NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

# Hearing the Moment: Measures and Models of the Perceptual Centre

By

Rudi C. Villing

A thesis presented to the
National University of Ireland
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Electronic Engineering
National University of Ireland Maynooth

September 2010

Research supervisors: Dr. Tomás Ward and Dr. Joseph Timoney
Head of department: Dr. Seán McLoone

# Abstract

The perceptual centre (P-centre) is the hypothetical specific moment at which a brief event is perceived to occur. Several P-centre models are described in the literature and the first collective implementation and rigorous evaluation of these models using a common corpus is described in this thesis, thus addressing a significant open question: which model should one use? The results indicate that none of the models reliably handles all sound types. Possibly this is because the data for model development are too sparse, because inconsistent measurement methods have been used, or because the assumptions underlying the measurement methods are untested. To address this, measurement methods are reviewed and two of them, rhythm adjustment and tap asynchrony, are evaluated alongside a new method based on the phase correction response (PCR) in a synchronized tapping task. Rhythm adjustment and the PCR method yielded consistent P-centre estimates and showed no evidence of P-centre context dependence. Moreover, the PCR method appears most time efficient for generating accurate P-centre estimates. Additionally, the magnitude of the PCR is shown to vary systematically with the onset complexity of speech sounds, which presumably reflects the perceived clarity of a sound's P-centre.

The ideal outcome of any P-centre measurement technique is to detect the true moment of perceived event occurrence. To this end a novel P-centre measurement method, based on auditory evoked potentials, is explored as a possible objective alternative to the conventional approaches examined earlier. The results are encouraging and suggest that a neuroelectric correlate of the P-centre does exist, thus opening up a new avenue of P-centre research.

Finally, an up to date and comprehensive review of the P-centre is included, integrating recent findings and reappraising previous research. The main open questions are identified, particularly those most relevant to P-centre modelling.

# Declaration

I hereby declare that this thesis is my own work and has not been submitted in any form for another award at any other university or institute of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

_____ _____

Signature Date

# Acknowledgements

Most of all I would like to thank my wife, Fiona, for being with me and being a constant source of support, belief, and encouragement throughout. Along the way, we were joined by Cian, who has provided us both with many hours of enjoyment and occasional responsibility. In his own way, he also helped me cross the final hurdles.

I would like to specially thank my two supervisors, Tomas Ward and Joe Timoney for providing guidance, focus, encouragement, and cool heads. I have learned a great deal from both of them. Special thanks must also go to Bruno Repp for being an excellent collaborator and for being stimulating, perceptive, encouraging, and patient throughout our work together.

I would like to thank current and former colleagues in the Department of Electronic Engineering who have all helped in one way or another: John M, Ronan, Sean McL, Bob, Frank, Seamus, John R, Sean D, Brian, Andrew, Denis, Yuriy, Martin, Damini, Alan, Eoin, Garret, Dave, Joanne, and Orla. I would also like to thank staff and students in the university who have aided me along the way, particularly Derek Walsh, Sean Commins, Richard Roche, Keith Finnerty, Ciara Whelan, Chris Soraghan, Brian Carty, and Tom Lysaght. For their recent patience I must also thank Alvaro Palomo and Tony Keenan.

I am thankful to my family (immediate and extended) and friends for being supportive and interested, but knowing when not to ask how the PhD was going. Finally, I am particularly grateful to both my parents who gave me the right start, believed in me, and showed that persistence pays off!

—

Wavelet software used in this thesis was provided by C. Torrence and G. Compo, and is available at http://atoc.colorado.edu/research/wavelets/

# Contents

# List of tables

# List of figures

# Nomenclature

| | |
|---|---|
| /abc/ | International Phonetic Alphabet transcription of phonemes |
| [abc] | International Phonetic Alphabet transcription of phones. (The degree of detail in a phonetic transcription can vary such that in some cases there is no difference from a phonemic transcription.) |
| "abc" | English orthographic transcription of speech tokens (having the normal relationship of spelling and sound). |
| $F^*_{SP}$ | Modified single point variance ratio (Stürzebecher, Cebulla & Wernecke 2001) |
| $F_{SP}$ | Single point variance ratio (Elberling & Don 1984) |
| $F$ | F-ratio or variance ratio |
| $M$ | Mean of a sample |
| $N$ | Sample size |
| $R^2$ | Coefficient of determination in a regression fit |
| $SD$ | Standard deviation of a sample |
| $SE$ | Standard error of the mean |
| $t$ | T-test statistic used in student's *t*-test |
| $\alpha$ | Strength of coupling in sensorimotor synchronisation |
| $\varepsilon$ | Greenhouse-Geisser correction for degrees of freedom in repeated measures with significant lack of sphericity |
| $\eta_G^2$ | Generalized eta squared, a measure of effect size that is more comparable across within-participant and between-participant experiment designs (Olejnik & Algina 2003) |

# List of abbreviations

| | |
|---|---|
| ACC | Acoustic change complex |
| AEP | Auditory evoked potential |
| BAEP | Brainstem auditory evoked potential |
| CAEP | Cortical auditory evoked potential |
| CV | Consonant-vowel (syllable) |
| CVC | Consonant-vowel-consonant (syllable) |
| EPC | Event-local P-centre |
| ERP | Event Related Potential |
| IOI | Inter-onset interval |
| IPA | International Phonetic Alphabet |
| IPI | Inter P-centre interval |
| ISI | Inter stimulus interval |
| jnd | Just noticeable difference |
| LLR | Long or late latency response |
| MLR | Middle latency response |
| MMN | Mismatch negativity |
| RMS | Root mean square |
| RMSE | Root mean square error |
| RPC | Relative P-centre |
| SNR | Signal to noise ratio |
| VC | Vowel-consonant (syllable) |
| VOT | Voice onset time |

# List of publications

## Conference and Workshop Publications

Villing, R., Ward, T. & Timoney, J. 2003, 'P-Centre Extraction from Speech: the need for a more reliable measure', paper presented to Irish Signals and Systems Conference, Limerick.

Villing, R., Timoney, J., Ward, T. & Costello, J. 2004, 'Automatic Blind Syllable Segmentation for Continuous Speech', paper presented to Irish Signals and Systems Conference 2004, Belfast, June 30 - July 2.

Soraghan, C., Ward, T., Villing, R. & Timoney, J. 2005, 'Perceptual Centre correlates in Evoked Potentials', paper presented to 3rd European Medical & Biological Engineering Conference (EMBEC '05), Prague, Czech Republic, 20-25 November.

Villing, R., Timoney, J. & Ward, T. 2006, 'Performance Limits for Envelope based Automatic Syllable Segmentation', paper presented to IET Irish Signals and Systems Conference, Dublin, Ireland, June 28-30.

Villing, R., Ward, T. & Timoney, J. 2007, *A review of P-centre models*, presented at the Rhythm Perception and and Production Workshop, Kippure Estate, County Wicklow, Ireland, July 1-5.

## Peer-reviewed Journals

Villing, R., Repp, B. H., Ward, T. & Timoney, J. 2010, 'Measuring Perceptual Centers using the Phase Correction Response', *[Submitted/in review]*.

# Chapter 1

# Introduction

The commonly held notion that there are just five senses derives primarily from the pioneering writings of Aristotle (350 BC/1993, Book II). These five senses, namely sight, hearing, smell, taste, and touch, are concerned with essentially external stimuli. There are, however, additional senses, including balance, proprioception[1], and pain, and these primarily communicate information about the state of the body rather than the external world. Falling easily into neither category is the "sense of time". Though time is certainly perceived it is not clear whether this perception can be considered to result from a primary and unitary sense or an abstraction inferred from more elementary percepts such as events (Grondin 2001). Despite this philosophical problem, it is still possible, useful, and necessary to investigate psychophysical properties of this sense-perception. This thesis, in particular, is concerned with measuring and modelling one specific aspect of time perception: the perception of event timing over relatively brief time scales.

The *perceptual centre* (P-centre) is the hypothetical[2] specific moment at which a brief event (generally shorter than about 1.5 seconds) is perceived

[1] Proprioception is the ability to sense the position, location, orientation, and movement of the body and its parts.

[2] The P-centre is hypothetical insofar as there is as yet no experiment design to prove that an individual event is perceived at a specific moment, a point which is taken up again in Chapter 4. Nevertheless, there is a substantial body of research, reviewed in this work, which supports the hypothesis.

to occur (Morton, Marcus & Frankish 1976). By the P-centre definition, when two brief events are synchronized, it is their P-centres that are (approximately) synchronous, and when a sequence of events occurs, it is the pattern of P-centres that determines whether the sequence is perceived as rhythmic (regular and predictable) or arrhythmic (unpredictable) and as expressively or mechanically timed. Figure 1.1 illustrates these relationships for isolated events.

The fundamental nature of the P-centre concept may be recognized by its relationship to the elementary temporal perceptions of simultaneity, successiveness, temporal order, and interval duration (Pöppel 1997), and the higher level perception of temporal patterns including rhythm. Although, the term P-centre has come to be associated with auditory and speech events only, Morton et al. (1976) explicitly specified the P-centre as a neutral concept applicable to events in any modality. It seems appropriate to return to this intended use.

Understanding the P-centre in detail depends on an understanding of events more generally. Segmenting continuous experience into discrete events appears to be a component of perception that takes place at multiple time scales concurrently (Kurby & Zacks 2008; Zacks et al. 2007). An event may be considered to be a segment of time that an observer conceives to have a beginning and an end (Zacks & Tversky 2001), though, in general, these boundaries may be imprecise and events may overlap. The description and identity of the event result from integration of the sensations and perceptions that occur during the event's span. Unlike objects, which persist and can be reexamined, individual events are ephemeral and can be experienced only once. For this reason, a collection of events is termed homogeneous if the events are identical except for a time shift, whereas heterogeneous (or mixed) events result when the underlying stimuli differ.

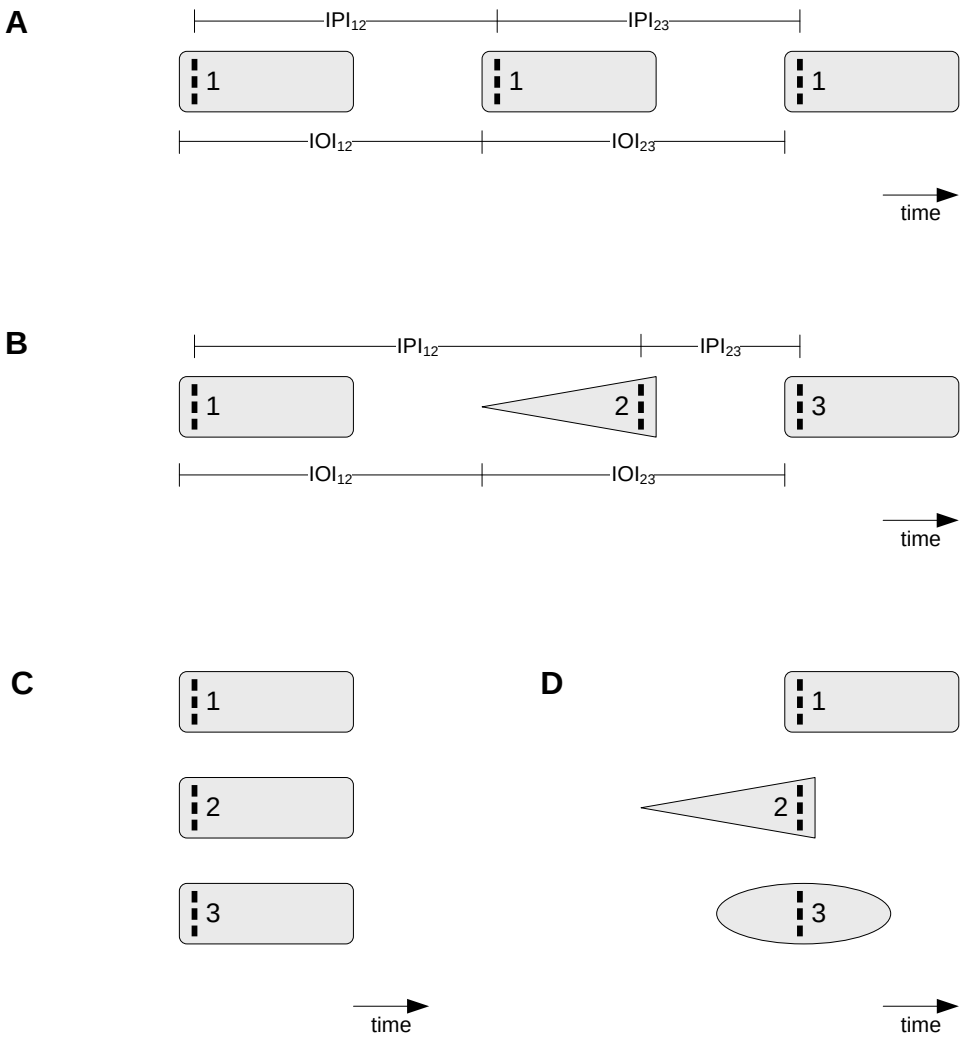


Figure 1.1 Schematic illustration of the relationship between onsets, P-centres, temporal patterns, and synchrony. Hypothetical P-centres are indicated by vertical heavy dashed lines. (A) Homogeneous events separated by intervals which are both objectively identical (inter-onset intervals $IOI_{12}$ and $IOI_{23}$) and perceptually identical (inter-P-centre intervals $IPI_{12}$ and $IPI_{23}$); (B) heterogeneous events separated by identical objective intervals but perceptually different intervals resulting from different onset to P-centre delays; (C) synchronous homogeneous events have synchronous P-centres and synchronous onsets; whereas (D) perceptually synchronous heterogeneous events have synchronous P-centres but asynchronous onsets.

This work considers only events and intervals that are directly sensed rather than remembered and take place within the timescale of the psychological present, or about 3 seconds (see for example Fraisse 1984; Pöppel 1997). Furthermore the P-centre is primarily concerned with events

that seem to occur at subjectively rather well defined times, for example, musical tones, speech syllables, visual flashes, and dance movements. When attending carefully, it may be possible to perceive both the event and its underlying percepts independently. For example the phonemes in a syllable such as "splash" may be perceived independently of the syllable as whole. Nevertheless, it appears that listeners can only reliably determine the timing of the syllable's P-centre (that is the perceived moment of occurrence of the syllable as a whole) with any precision (Whalen, Cooper & Fowler 1989). In particular, although it is possible for listeners to detect that the onset of the /s/ in "splash" occurs before the P-centre, it does not appear possible to accurately identify the timing of that initial onset or to use it for synchronisation or rhythmic timing of events. (This point is explored in more detail in Chapter 2.)

Ultimately, the goal of P-centre research is to accurately model human perception of event timing so that the perceived timing of a sequence of events can be predicted without constant recourse to empirical measurement. Although a homogeneous sequence of events can be easily timed using the intervals between any convenient corresponding time points, it is not possible to accurately measure or control the timing of heterogeneous events (either within or between sensory modalities) unless the corresponding P-centres are known (see Figure 1.1). Although this limitation is generally not mentioned, it has an effect on many research questions that concern timing. For example, research into sensorimotor synchronization (see Repp 2005 for a review) is generally constrained to use homogeneous (or nearly homogeneous) event sequences in order to avoid the potential effect of P-centre differences between events. Investigations of rhythmic timing and microtiming (the intentionally produced timing variations that give human performance its natural, expressive quality) cannot adequately measure the perceived timing of performances in which the P-centres of events in a sequence can vary substantially relative to each other. In particular, without knowledge of P-centres the rhythm of spoken language cannot be measured accurately and

thus questions about the perceived timing of individual languages can be answered only on the basis of flawed or indirect data at best. A researcher who needs to prepare event sequences with specific perceptual timing for use in an experiment cannot use heterogeneous events if the event P-centres are not known. Indeed, the P-centre term originated when Morton et al. (1976) discovered that they could not easily construct a perceptually regular sequence of recorded words for a memory experiment. In general, the P-centre is a necessary component of expressive performance in speech, music synthesis and other temporally sensitive activities, and it may well have a part to play in achieving natural interaction and gesture timing for anthropomorphic robot and virtual human models (for a suggestive example, see Murata et al. 2008).

## 1.1 Motivation and objectives

The work in this thesis grew out of a problem encountered while investigating expressive speech synthesis. Specifically, it is the prosodic aspects of speech, including pitch, stress, and rhythm, that most distinguish expressive speech from artificial synthetic speech. It seemed that there were well established methods for measuring and manipulating pitch and stress but not rhythm. Although the duration of speech units such as phonemes can be straightforwardly manipulated and measured, there was apparently no method to map from these durations to the perceived rhythm of speech. Subsequent research uncovered the P-centre concept and perhaps the potential solution to this problem.

Although the P-centre term is now more than thirty years old, at the outset of this work its state of development as a theoretical concept, a body of empirical findings, and a feature of events that could be manipulated or measured was unclear. That this was the case despite the P-centre's fundamental importance in event timing was surprising. Therefore, a critical review and integration of the published research on theoretical and empirical aspects of the P-centre phenomenon became the initial objective

of this work. This thesis focused on acoustic P-centres because of the motivating problem domain (expressive speech) and because essentially all published literature investigated acoustic P-centres only.

Due to its nature as an entirely perceptual construct, the P-centre is elusive and there is no truly objective means of measuring it. Though a variety of measurement methods have been described and used there had apparently been no analysis of their comparability, reliability, or efficiency. A second objective of the work, then, was to investigate empirical P-centre measurement in general and determine how best to measure P-centres.

Finally, the solution to the original expressive speech problem requires a P-centre model for two purposes: first, to analyse and extract rhythm from natural expressive speech; and second, to help synthesize a speech waveform with the appropriate perceptual rhythm. (Expressive music synthesis with heterogeneous sound events is a similar problem requiring a similar solution.) Because the literature described several models but provided no guidance regarding which one to use, evaluating the existing models became the last objective of this work.

## 1.2 Thesis contributions

The primary contribution of this work is a coherent integration of prior research providing a foundation upon which future P-centre developments can rely and lowering the barrier for entry into the field. This foundation comprises the following detailed contributions:

1. A detailed survey of published empirical findings and the theoretical arguments concerning the P-centre phenomenon. The findings were integrated to identify open empirical questions and to reassess the theoretical framework in which P-centres are analysed and modelled.

2. The introduction of a new behavioural method for P-centre measurement (the PCR method) and the experimental evaluation of this method with the two other principal methods leading to specific method recommendations. In addition, the concept of P-centre clarity was introduced and previously unknown effects on the strength of sensorimotor coupling (the coupling between sensory input and motor action in a synchronisation task) were discovered. (The work which produced this contribution was conducted in collaboration with Bruno Repp of Haskins Laboratories.)

3. The investigation of a novel neuroelectric method for measuring P-centres which found that there was correlation between neuroelectric and behavioural P-centre measures. Though further investigation and refinement is necessary, the technique has potential and may provide insight into the objective timing of the P-centre and its underlying physiology.

4. A detailed analysis of existing model specifications, integrated from several sources where necessary, leading to detailed operational descriptions and fully commented software implementations. With this contribution the cost of enhancing an existing model or developing a new model is greatly reduced.

5. A comprehensive evaluation of the existing models which indicated that the existing models make predictions which are both inconsistent with one another and fail to correctly predict the measured P-centres of at least some stimuli. Some of the models make sufficiently poor predictions (with at least some stimuli) that their future use is not recommended. The features of the remaining, partially successful models are assessed and future development directions proposed.

## 1.3 Thesis outline

The remainder of this thesis comprises five main chapters and appendices.

Chapter 2 presents a review of the empirical results from the three disparate approaches to prior P-centre research, namely speech oriented P-centres, general acoustic P-centres (including music), and articulatory P-centres. This chapter also includes a review of the principal theoretical arguments and concludes with a discussion which integrates the findings to date and identifies the research questions that remain open or require confirmation.

Chapter 3 begins with a review of P-centre measurement methodology, then describes a new measurement method (the PCR method), followed by a significant empirical study designed to determine which method allows P-centres to be measured most efficiently. The chapter concludes with a discussion that makes specific method recommendations and raises some new research questions.

Chapter 4 introduces neuroelectric measurement as a novel approach to measuring P-centres. The chapter begins by reviewing methodological issues related to EEG analysis and the relatively little available neuroelectric research in the P-centre related fields of rhythm and meter. This is followed by an exploratory empirical study comprising two separate experiments which provides evidence of a neuroelectric P-centre correlate.

Chapter 5 opens with a detailed operational description of each of the existing P-centre models based on the actual model implementations created specifically for this thesis. In all cases, this description incorporates necessary elements that were either omitted or ambiguous in the original model descriptions. At this early stage, common model strategies and potential problems are also identified. The remainder of the chapter is devoted to a detailed two part evaluation of the models tested against a specific corpus of sounds. At the chapter conclusion the most accurate

models are identified, their strengths are assessed, and suggestions for developing a more accurate model are made.

Finally chapter 6 concludes the main body of the thesis with a summary of the work performed and the main results. An extensive list of suggestions for future work is also included based on open questions identified in the literature and new questions resulting specifically from the work in this thesis.

# Chapter 2

# The P-centre phenomenon

Many aspects of the P-centre phenomenon have been explored empirically and at least two fundamentally different theoretical frameworks for the P-centre exist. Nevertheless, assembling the available information into a coherent representation of the P-centre phenomenon is challenging. A general problem is that P-centre research is a niche field with a relatively small set of available data and many results have been described in less readily available theses or conference presentations only. Though one may speculate on the reasons for this, it is nevertheless preferable to critically review all sources. At least two partial reviews of the field do exist, but both are now quite old (Scott 1993; Seton 1989) and an up to date inclusive review is certainly required.

The P-centre is, as noted in the Chapter 1, a general term intended to be applied to brief event timing in any modality. Researchers have used other terms to refer to the essentially the same concepts in more restricted domains, including the *syllable beat*[3] of (English) speech (Allen 1972a) and the *perceptual attack time* (PAT) of musical tones (Gordon 1987). A term which is quite distinct from the P-centre, though sometimes encountered in the same contexts, is the *perceptual onset* of an event, the moment at which

[3] Another term used in Allen's paper and favoured by some researchers is the *stress beat*. However this term implies that unstressed syllables have no beat and can elicit no perception of event timing, an interpretation which is not consistent with the P-centre as the perceived timing of any brief event whether it is stressed or unstressed.

the event is first detected. The distinction, discussed in more detail later, is that events with relatively long, gradual, or complex onsets (for example the syllable /sa/) generally appear to have rather late P-centres despite the initial onset of acoustic energy being perceived early.

In Chapter 1 the P-centre was related to temporal perceptions such as synchrony and rhythm. Before proceeding, it is necessary to introduce the set of related temporal percepts and terms with which the P-centre is often associated. *Rhythm* defines a temporal pattern and specifically defines a pattern with some element of regularity and predictability; an irregular pattern can be described as *arrhythmic*. The nature of the predictability in rhythm results from its relationship to a higher level canonical pattern of timing and stress termed *meter*.

Meter arises in both music (notated as the familiar time signature which specifies beats in a bar) and linguistics (as the patterns of stress in a sentence—particularly prominent in poetry and rhyme). Unlike rhythm which is fully determined by the timing of events, meter is an abstract percept which may be inferred from not only the times at which events occur, but also the intervals between them. Related to meter is the concept of *pulse*, the basic periodic beat in music. Like meter, pulse is an abstract percept which may be implied rather than present in a sequence of events. A *pulse group* is a group of pulses with a particular stress pattern (for example strong-weak-weak) and meter is ultimately defined by a repeating pattern of pulse groups. In spoken language, and poetry in particular, the basic element of meter is usually termed the *foot*. Each foot is composed of one or more syllables with a particular stress and duration pattern. Well known patterns include the stressed monosyllable (e.g. “cat”), the trochaic disyllable (long-short pattern, e.g. “peacock”), and the iambic disyllable (short-long pattern, e.g. “reprieve”).

The relationship of rhythm to meter is that meter forms a relatively stable temporal framework of pulse groups in reference to which rhythmic events may be predictably timed: some rhythmic events may occur on pulses

(stressed or unstressed) while others may occur between pulses; the occurrence of a pulse with no attendant rhythmic event is also possible. In musical contexts, *tempo* defines the rate at which pulses or pulse groups occur.

A particularly simple rhythm is *isochrony*, in which all intervals are identical. *Objective isochrony* (or *physical isochrony*) indicates that the objectively measurable (physical) onsets of events occur at identical intervals, whereas *perceptual isochrony* indicates, according to the P-centre definition, that it is the P-centres of those events which occur at identical intervals. In an isochronous rhythm, all rhythmic events occur on a pulse and all pulses within the meter are marked by rhythmic events. There is, however, a tendency for the meter induced by an isochronous rhythm to be perceived as two element pulse group consisting of a stressed and unstressed pulse (as in the familiar "tick-tock" of an objectively isochronous ticking clock). Just as asynchrony indicates a deviation from synchrony, *anisochrony* refers to a deviation from isochrony. Methods which distinguish between isochrony and anisochrony are the most common in P-centre research.

The remainder of this chapter comprises a review of the empirical findings, a review of the theoretical frameworks, and finally a discussion which reappraises both.

## 2.1 Empirical review

In collecting empirical data it is necessary to make P-centre measurements and a brief comment on the various approaches is warranted. Many researchers approach the P-centre problem from a speech and linguistics specific perspective. The majority of these researchers have investigated the relationship of the P-centre to various kinds of natural or edited speech stimuli (see for example Cooper, Whalen & Fowler 1986, 1988; Harsin 1997; Marcus 1981). The perceived benefit of edited natural stimuli is that

the edits can be designed to manipulate the stimuli along just one parameter dimension (for example vowel duration). It is worth viewing these manipulations cautiously however: auditory perception is complex and it is rarely possible to manipulate one parameter without simultaneously affecting several auditory percepts by side effect. For example, the very simple manipulation of steady state duration also affects perceived loudness, can induce timbre changes at short durations, and can induce loss of pitch strength in the case of periodic sounds[4]. Synthetic speech has also been used in certain cases (notably by Pompino-Marschall 1989), but while this undoubtedly has the benefit of greater stimulus control it remains to be seen whether the results obtained generalise to natural speech.

A smaller, but still significant, set of P-centre research has investigated the production of speech having a specific perceptual rhythm, both with and without an external pacing aid such as a metronome. Such investigations fall into two broad categories: those which focus on measuring properties of the produced acoustic waveform (e.g. Fowler 1979; Fox & Lehiste 1987a, 1987b; Rapp-Holmgren 1971), and those which instead measure the articulatory gestures required to produce the speech (e.g. de Jong 1994; Patel, Lofqvist & Naito 1999; Tuller & Fowler 1980).

The remaining P-centre research has been approached from a more general acoustic perspective. Some useful comparisons have been reported between P-centres in edited or natural speech and those in simpler synthetic sounds designed to have some features in common, such as, for example, a similar amplitude envelope. There have, however, been relatively few investigations which disregard speech entirely to focus on purely synthetic sounds (Schütte 1978; Vos, J. & Rasch 1981) or musical sounds (Gordon 1987; Wright 2008). There is apparently no research available on the

[4] As duration is varied from about 5 to 250 ms, the perception of a simple 1000 Hz tone changes from a click (with almost no sensation of pitch), to a pip with gradually increasing pitch strength, to a tone with clear pitch whose loudness gets louder.

possible P-centres of non-linguistic human vocalisations or non-human animal vocalisations.

In addition to these broadly different orientations (speech perception, speech production, and acoustic perception), researchers have used a variety of psychophysical methods and tasks to measure P-centres including adjustment (to isochrony or synchrony), synchronous tapping, and constant stimuli with a forced choice task. At this point the specific detailed operation of each method is unimportant (measurement methods are examined in detail in Chapter 3). It is, however, important to recognise that different methods and experimental configurations were used and that these differences may have had an effect on the results ultimately obtained.

The following subsections survey each of the main empirical P-centre research areas in turn.

### *2.1.1 P-centre precision and perceptibility of deviations*

Given that events span time and do not, as a whole, objectively occur at a specific moments, it is reasonable to question whether they occur at subjectively specific moments. Phrasing this question more explicitly: is the P-centre a specific moment? If the time of the P-centre was not specific but was instead distributed in time, then the perception of synchronization and rhythmic timing should be both highly imprecise and variable, particularly between heterogeneous events. But this is not what the evidence suggests.

Rasch (1979) reported that the absolute deviation from synchrony in natural music performance was typically about 30-50 ms depending on both the instrument timbres and the average inter-onset interval (IOI) or tempo. These deviations were not measured for P-centres or physical onsets but for onsets defined by a relative threshold 15–20 dB below maximum. These deviations from synchrony can be interpreted as having two contributing components: one due to perceptual tolerance for asynchrony and another due to P-centre differences between sounds.

The *just noticeable difference* (jnd) of events from isochrony may be studied by temporally displacing (shifting) events in an otherwise isochronous short sequence. Friberg and Sundberg (1995) list a variety of possible displacement patterns, including single event displacements (just one event in the sequence is shifted in time away from its perceptually isochronous point) and cyclic event displacements (the overall sequence is subdivided into repeating subsequences, termed cycles, and identical event displacements occur in each cycle). They found that the jnd depended on the type of deviation, the sequence length, and the IOI. Tempo changes were more detectable than cyclic displacements and single event displacements. For IOIs below 250 ms, the absolute jnd appeared approximately constant whereas above this value the relative jnd was approximately constant. Friberg and Sundberg's results—6 ms absolute jnd and 2.5% relative jnd—approximated the mid value of previous findings when doubled to correct for methodological differences[5].

Madison and Merker (2002) investigated the threshold of anisochrony in a nominally isochronous sequence, and the threshold of pulse attribution (the subjective experience of a periodic pulse) in an objectively anisochronous sequence. Using a short percussive stimulus, IOIs ranging from 570–630 ms, and sequences with essentially unpredictable deviations, they found that the detection threshold for anisochrony was 3.5% of IOI (20-22 ms), but the threshold for pulse attribution was 8.6% of IOI (49-54 ms). These disparate thresholds suggest a difference between the ability to detect anisochrony and the ability to tolerate it.

Repp (2002), in an investigation of sensorimotor synchronisation, demonstrated that participants can respond automatically to subliminal temporal deviations below the conventional threshold for consciously detectable anisochrony. Participants tapping in synchrony with a mostly

[5] Friberg and Sundberg used an adjustment method and estimated the jnd as the *SD* of adjustments whereas previous research typically estimated the jnd as the 50% detection level using, for example, a forced choice task.

isochronous sequence (IOI = 500 ms) containing occasional timing deviations as small as 10 ms (2% of IOI) exhibited a consistent compensatory correction in response to these deviations. Madison and Merker (2004) demonstrated even greater sensitivity. With a nominal IOI of 600 ms and a continuous unpredictable sequence of deviations, musicians and non-musicians responded to deviations as small as 1.5 and 3 ms respectively (less than 1% of IOI in both cases).

Taken together, these findings all suggest that the P-centre is a specific moment, that the jnd for deviations from a predictable rhythm (always simple isochrony in these studies) is no more than 5% of IOI, and that even substantially smaller subliminal P-centre deviations may be perceptually relevant.

Nonetheless, the subjective precision of P-centres associated with different sounds may differ (a point explored in more detail in Chapter 3). Relating measured anisochronies to the difference in rise time between sounds, Rasch (1979) made the assumption that "shorter and sharper rises of notes make better synchronization both necessary and possible" (p. 128). Allen (1972b), using a forced choice paradigm, found that listeners perceived the synchronisation between a click and one syllable in a continuous speech utterance as if the syllable beat were a "broad slur, approximately 200 msec. in duration" (p. 189). Gregory (1978), however, notes that there are various problems perceiving the synchronicity of clicks with music and speech—problems that seem to be at least partially caused by auditory streaming (Bregman 1990/1999). Thus, it would seem methodological problems may have produced Allen's broad slur.

Both Gordon (1987) and Wright (2008) found that the distribution of synchronisation responses to instrumental tones could in some cases be multi-modal, but again methodological issues may be to blame for this (see Chapter 3). Wright made the interesting proposal that the P-centre should be represented by a probability density function rather than a single moment. Nevertheless, unless there is more convincing evidence of a

multimodal P-centre, it would seem to be the central tendency (e.g. the mean or median) of the measured P-centre distribution that is most valuable in both a measurement and modelling context. P-centre variability almost certainly includes task specific components and can perhaps be represented more effectively by conventional measures such as the standard deviation or inter-quartile range.

### *2.1.2 The perceptual onset*

For separated, non-overlapping events, there is no a priori reason to assume that the P-centre is not located at the perceived event onset, that is, at the moment of event detection. In particular, musical notation encourages exactly this assumption: Rhythm is assumed to be specified by the timing of note onsets and not their durations or offsets (Rasch 1979). However, Morton, Marcus, and Frankish (1976) failed to construct perceptually regular sequences of recorded words for their memory experiment when they made the word onsets isochronous; clearly the P-centre is not coincident with the onset of a word or syllable. (Although they did not specify it, it must be assumed that Morton et al. used onset to mean the objective or physical onset rather than the perceptual onset. However, their Figure 1 does not demonstrate any alignment by a common threshold, a feature that would be expected if perceptual regularity resulted from perceptual onset isochrony.) Numerous subsequent studies support the idea that the P-centre in a speech syllable occurs somewhere in the vicinity of the vowel onset, substantially after the perceptually detectable onset of acoustic energy in the case of syllables with long initial consonants or consonant clusters.

Gordon (1987) similarly found that neither a simple absolute nor relative onset threshold could accurately predict the P-centre of all the musical tones he had empirically measured. In fact, the P-centre of acoustic and speech events does not appear to reliably correspond to any obvious acoustic or speech specific feature. Numerous candidate features have been

considered but shown to fail in at least some cases; these include local or global intensity peaks (Gordon 1987; Marcus 1981), the measured vowel onset (Marcus 1981), the number of initial consonants (Cooper, Whalen & Fowler 1986), and the vowel quality (Fox & Lehiste 1987b).

For continuous stimuli, which may result in imprecise and overlapping event boundaries, the interaction between events in the vicinity of their onsets and offsets would also seem to argue against the P-centre corresponding to a single simple onset-related feature and indeed, the concept of a perceptual onset is difficult to define in such a context.

### *2.1.3 General features of the P-centre*

Morton et al. (1976) showed that isolated digits had to be objectively anisochronous in order to sound perceptually isochronous. Fowler (1979) briefly investigated whether naturally produced anisochronies were perceived to be more "rhythmic" than sequences in which the silent periods were edited to create objectively isochronous sequences. In what seems subsequently to be a rather obvious outcome, the results showed that listeners chose natural sequences at far greater than chance frequency.

An interesting study conducted by Fowler, Smith and Tassinary (1986) investigated whether pre-babbling infants would show preference for similar objective anisochronies as adults. The results indicated that they did, from which it was inferred that infants, even before learning speech gestures themselves, perceive stress beat (P-centre) timing as adults do.

In research that pre-dates the P-centre term, Allen investigated the timing of syllable beats in English (Allen 1972a). His experiments indicated that when participants tapped in synchrony with syllables, the variability of those taps depended on the degree of syllable stress: taps were less variable with stressed syllables than unstressed syllables. He also reported that when participants adjusted clicks to synchrony with a target syllable, the resulting variability was less than that of their taps. Judging the timing of a

click relative to speech or music is, however, more complex than one might initially suspect: clicks tend to be attracted to phrase boundaries and are perceived early in speech and late in music (Gregory 1978).

One of the most commonly employed perceptual methods of estimating P-centres uses an adjustment paradigm, first described by Marcus (1981). Using this method with the digits "one" to "nine" Marcus found no evidence of participant or context[6] effects on the measured P-centres. In contrast Pompino-Marschall (1991) found that there was a significant effect of context when he measured P-centres for the syllables /pak, bak, fak, vak, mak/. Thus, the literature is inconsistent on this point.

Whalen, Cooper, and Fowler (1989) investigated whether participants could attend to temporal features of the stimulus other than the P-centre (the onset, vowel-onset, and offset) when making adjustments of the sort described by Marcus. Their data showed that participants were unable to perform the task of adjusting to offset and that for the other features their adjustments were either in the wrong direction or not significantly different from those made when attending to the P-centre. Seton (1989) also investigated whether participants attend to offset to maintain isochrony when the rise time of stimuli is varied, but his data did not support this hypothesis.

The effect of the presentation rate, or IOI, used when estimating the P-centre with the adjustment method has also been investigated (Eling, Marshall & van Galen 1980; Scott 1993). Eling et al. found that the P-centre estimate was independent of IOI for IOIs between 600 and 2500 ms. Scott found that a 600 ms IOI yielded more reliable estimates (having smaller standard errors) than a 400 ms IOI. Using a synchronous tapping paradigm Vos, Mates, and van Kruysbergen (1995) also found no significant effect on the P-centre estimates for IOIs between 500 and 900 ms. Curiously, these

[6] In this case, context refers primarily to the choice of "other" sound in an alternating sequence.

results appear to be at odds with the jnd for deviations from isochrony (Friberg & Sundberg 1995). Since the relative jnd is approximately constant over the range of IOIs tested, the accuracy and reliability should be better at shorter IOIs than longer ones.

Finally, in a finding which may be related to the time shrinking phenomenon (ten Hoopen et al. 1995), Lehiste (1973) found that in a set of equal intervals the last one was always perceived to be shorter than its objective duration. Nevertheless, listeners appear to perform better with non-speech which Lehiste interpreted as an indication that listeners tolerate larger timing deviations in speech. In a comment that is relevant to the methodology of much P-centre research she suggested that words produced in isolation may be produced as if they were in utterance final position and therefore may not be representative of sentence internal stress patterns and durations.

### *2.1.4 Syllable segments*

A large majority of the P-centre investigations undertaken to date have attempted to relate the P-centre to some feature of syllable segments, most commonly segment duration. As the definition of the syllable and related concepts is not universally standardised, the working definitions used in this thesis are introduced before proceeding further.

A syllable is an elementary constituent of spoken language and all languages have a syllabic structure (Holmes & Holmes 2001). The syllable is composed of a continuous sequence of one or more elementary sounds. The core of any syllable is a vowel or vowel-like sound and, subject to language specific constraints, this may be preceded or followed by one or more consonants. Denoting a consonant sound as C and a vowel (or vowel-like) sound as V, various syllable possibilities can easily be represented as shown in Table 2.1.

**Table 2.1 The relationship of consonants and vowels to syllable structure**

| Monosyllabic word | | | | Syllable Rhyme | |
|---|---|---|---|---|---|
| Orthography | Phonemes | C and V | Syllable Onset | Nucleus | Coda |
| "a" | /æ/ | V | — | V | — |
| "do" | /duː/ | CV | C | V | — |
| "at" | /æt/ | VC | — | V | C |
| "cat" | /kæt/ | CVC | C | V | C |
| "scratched" | /skrætʃd/ | CCCVCCC | CCC | V | CCC |

*Note*—Phonemic spelling derived from Cambridge Dictionaries Online (*Cambridge Dictionaries Online* 2009)

The syllable may be structurally decomposed into an onset (comprising the initial consonants, if any) and rhyme (comprising all subsequent sounds). The syllable rhyme may be further decomposed into the nucleus (the central vowel or vowel-like sound in the syllable) and the coda (the final consonants, if any). The onset, rhyme, nucleus, and coda may be generically referred to as syllable segments. Table 2.1 illustrates some of these possibilities.

Many P-centre studies have investigated how the duration of syllabic segments (such as the onset, or rhyme) or equivalently the boundaries of such segments (for example, the beginning of the syllable nuclear vowel) affect the P-centre. Whether the intent is to measure a duration or the timing of a boundary point, the requirement is the same: the boundary point (or points) must be unambiguously identified. In practice, acoustic signals rarely have unambiguous boundaries and researchers use a variety of different techniques and heuristics to identify them. For example, the time of vowel onset in a syllable was measured from spectrograms by Fowler and Tassinary (1981) using either the point where the "glottally excited, full formant pattern was first evident" (p. 526) or by matching the amplitude and frequency of the third formant between syllables. Rapp

(1971) measured vowel onset from printed oscillograms (waveforms), with a claimed accuracy of ±5ms, but did not specify the heuristic used to identify the vowel onset when it was embedded in the transition from a voiced consonant. Janker (1996a) specifically noted that the measured boundary for a segment can easily vary by one or two glottal pulses in either direction, depending on the segmentation heuristic used; for a speaker whose average fundamental frequency is 100Hz, this difference amounts to ±20ms. This source of variability must be considered when experimental results from different researchers are compared and can even be problematic within the analyses for a single experiment. A further complication highlighted by Tuller and Fowler (1980) is that the linguistic boundaries conventionally selected in acoustic waveforms (for example the time at which the features of the nuclear vowel dominate over the features of the initial consonant) may have no psychological significance.

#### *2.1.4.1 Syllable identity*

The simplest experimental manipulation used in P-centre investigations controls nothing other than the identity of syllables whose P-centre is to be measured.

Marcus (1981) used this approach to measure the P-centres of the digits "one" to "nine" and found that the interval between P-centre and vowel onset was linearly related to the initial consonant duration (slope = 0.75) for all tokens except "six" and "seven".

Janker (1996a) also used naturally produced stimuli, but specified syllables which varied in initial consonant only. Using a synchronous tapping paradigm, he found that the mean tapping position (corrected for individual differences and assumed to co-vary with the P-centre) varied from about 10 ms before to 30 ms after the vowel onset for the syllables [ʔast$^{h}$, past$^{h}$, fast$^{h}$, kast$^{h}$, hast$^{h}$, k$^{h}$ast$^{h}$, last$^{h}$, mast$^{h}$, p$^{h}$ast$^{h}$, ʁast$^{h}$, t$^{h}$ast$^{h}$]. Because individual productions of the syllable rhyme varied, it is difficult to

generalise from this data except to note that the mean tapping location did not depend only on the initial consonant duration. In what seems to be one of the only investigations into non-syllabic speech, his data also showed that two consonant-only interjections, [sːtʰ] and [pstʰ], elicited mean tapping positions that were 27 and 60 ms after onset respectively.

In related work (Janker 1996b) he showed that mean tapping position, (again corrected for individual differences) ranged approximately 20–40 ms before the nuclear vowel in a variety of monosyllables with short vowels (/ʃtɪl, ʃtɛl, ʃtal/), long vowels (/ʃtiːl, ʃteːl, ʃtɑːl/), and constant nucleus but varying onset and coda complexity (/ʃɑːl, ʃtɑːl, ʃtrɑːl, ʃɑːlt, ʃtrɑːlt, ʃɑːlst, ʃtɑːlst, ʃtrɑːlst/). There was no evidence that increased complexity in the rhyme affected the P-centre.

#### *2.1.4.2 Syllable onset*

One of the first effects noted by researchers was the apparent relationship of the P-centre in monosyllables and the syllable onset (initial consonant) duration. A different but closely related interpretation of the same effect is that the P-centre is located in the vicinity of the vowel onset. To investigate this effect, the initial consonant duration has been manipulated in a variety of experiments.

Marcus (1981) edited the token "seven" by deleting 0–150 ms from the initial frication (in 30 ms steps). His results indicated no effect with the first deletion and a linear shift (slope = 0.45) towards the onset for each subsequent deletion. This linear relationship was seemingly unaffected by either the categorical change in initial phoneme over the course of the deletions (from "seven" to "devon") or the abrupt onsets which resulted. The lack of effect from the first deletion suggests that the initial energy may have been below a perceptual threshold.

Cooper, Whalen, and Fowler (1986) edited a /ʃa/ syllable by deleting 0–135 ms of the initial frication (in 15 ms steps) and correspondingly applying a

linear ramp to the first 150–15 ms of the onset. They found that the manipulation altered the P-centre by almost precisely the same amount (slope = 0.95). A second experiment, using a /sa/ syllable edited by inserting 0–100 ms of silence (in 10 ms steps) between the initial consonant and the vowel, found a 1:1 effect of the manipulation on the P-centre. Pompino-Marschall (1987) replicated this experiment with essentially identical results.

Harsin (1997) edited naturally produced CV syllables, [ʃa, na, ra], to manipulate the initial consonant duration (120, 160, and 200 ms) while holding the vowel duration constant (280 ms). He found that longer onsets resulted in later P-centres (approximately 1:1 for [na], but slightly less for the other syllables). He also examined the stop-consonant CV syllables [ta, da, ka, ga] edited to have constant consonant duration (80 ms) and vowel duration (320 ms). In this case the results showed that the voiced stops had earlier P-centres than the unvoiced stops (the mean difference was 27 ms).

In summary, manipulating the initial consonant duration (or the temporal onset of the vowel relative to the syllable onset) appears to have a strong effect on the P-centre. There is, however, some disagreement among the results regarding precisely how strong this effect is.

#### *2.1.4.3 Syllable rhyme*

Research generally indicates that syllable rhyme duration has an effect on the P-centre, though the effect seems to be weaker than that of the syllable onset.

Marcus (1981) measured the P-centres of natural /bæ, dæ, gæ, pæ, tæ, kæ/ syllables and lengthened /bæ, dæ, gæ/ syllables whose vowel duration was extended (~60 ms) by duplicating pitch periods. The results showed that lengthening the vowel duration shifted the P-centre later (by ~20 ms).

Marcus also found that altering the duration of the rhyme in the syllable "eight" by changing the stop closure duration had a small effect on the P-centre (duration changes of -30 and 30 ms shifted the P-centre by -9 and 13 ms respectively). In contrast, changes in the level of the final t-burst, described by Marcus as much more perceptible than the duration changes, had almost no effect on the P-centre. Thus Marcus concluded that it is the temporal makeup and not the amplitude or energy which most affects the P-centre.

Cooper, Whalen, and Fowler (1988) manipulated the rhyme duration in two experiments. In the first of these the vowel duration was edited by deleting pitch periods to create matched /a/ and /sa/ syllable continua. The latter syllable was formed by adding frication (202 ms) to the vowel (424–526 ms). The effect of vowel duration on the P-centre was significant but unfortunately subject to a significant participant effect which appears to prevent generalization. The second experiment created two /at/ continua by deleting 8–99 ms from the vowel both with and without compensatory change in the silent stop closure duration. The P-centre of the first continuum, whose vowel duration, rhyme duration, and syllable duration changed simultaneously, shifted earlier as the durations reduced. In contrast, the P-centre of the second continuum, whose total duration remained constant (549 ms), showed no effect for two out of three participants. Again the effect of participant was significant. Cooper et al. concluded that the effect of vowel duration is present but weaker and less reliable than effect of vowel onset time on the P-centre.

Harsin also examined final consonant duration and quality (1997). The consonant duration (120, 160, and 200 ms) of naturally produced VC syllables, [aʃ, an, ar], was manipulated while the vowel duration was held constant (280 ms). Unlike previous researchers, Harsin found no reliable effect of final consonant duration or class.

Therefore it seems that the duration of the nuclear vowel has a weak effect on the P-centre but the potential effect of final consonant duration is less certain.

#### *2.1.4.4 Combined effects*

Several investigations have examined combined effects of the syllable onset and syllable rhyme, or individual constituents of the syllable rhyme, namely the nucleus and coda.

Cooper, Whalen, and Fowler (1986) examined the effect of compensatory segment duration changes on the P-centre in a /sa/ syllable such that the total syllable duration (566 ms) remained constant. In the first experiment 0–100 ms silence was added between the consonant and the vowel and a corresponding amount of frication was deleted from within the consonant. They found that this manipulation, which did not affect the timing of the vowel onset, did not alter the P-centre. A second experiment inserted 0–93 ms silence between consonant and vowel but compensated by deleting an equivalent duration (in whole pitch periods) from the vowel. In this case, the manipulation, which shifted the vowel onset as silence was inserted, did alter the P-centre, but the effect was smaller than when the vowel duration was not edited (the slopes relating the manipulation to the P-centre were 0.83 and 1.00 respectively). Thus reducing the vowel duration appeared to weaken the effect of its onset on the P-centre. When Pompino-Marschall replicated this experiment (1987), he found a smaller effect on the P-centre (slope = 0.53) than Cooper et al.

Pompino-Marschall (1989) investigated whether the effects of the initial consonant duration and vowel duration were linearly independent in CV syllables. Using synthetic /ma/ syllables whose consonant duration (40–200 ms) and vowel duration (100–260 ms) were independently manipulated his results showed an approximately 1:1 effect of consonant duration and weaker effect of vowel duration (slope ≈ 0.25) on the P-centre.

There was significant interaction between these effects (they were not independent) and they exhibited some non-linearities. Specifically the effect of consonant duration was weaker for longer durations. Replicating the experiment with a square wave (100 Hz) whose envelope was matched to the syllable found similar general effects but the P-centres were earlier on average than those of syllables. A confounding factor in these results, however, is that for most combinations of consonant and vowel duration the overall duration must have changed also.

Pompino-Marschall applied the same consonant and vowel duration manipulations to a synthetic /ʃi/ syllable whose envelope was identical to the /ma/ syllable. The results were similar in trend to those for /ma/ but the small differences were nevertheless significant. The effect of vowel duration on the P-centre in particular was weaker in this case (slope ≈ 0.16).

In a second experiment Pompino-Marschall examined whether there was a single effect of the syllable rhyme duration on the P-centre or if instead there were independent effects of the nuclear vowel and final consonant duration. Using synthetic /am/ syllables whose vowel and consonant durations were independently manipulated (100–260 ms and 40–200 ms respectively) and square wave tones with identical envelopes his results showed a weak effect of vowel duration (slope ≈ 0.2) and final consonant duration (slope ≈ 0.14) on the P-centre of syllables. Furthermore he found a significant interaction between these two manipulations. In this case the P-centre of square wave tones tended to be later than corresponding syllables and the strength of both vowel and consonant duration effects was slightly larger.

### *2.1.5 Syllable segment envelope*

Existing data made it clear that the duration of all syllable segments appeared to have an effect on the P-centre of monosyllables, although the

duration of the syllable onset was the factor with the strongest effect. Does the amplitude envelope of a syllable also have an effect on the P-centre? Several studies have addressed this question.

Marcus (1981) modified the amplitude (by 4.5 and 9 dB) of the final t-burst in the token "eight". Though the manipulation was clearly perceptible to Marcus it had almost no effect on the P-centre. Scott (1993) replicated the experiment (with a 6 dB amplitude modification) and found that there was a weak effect: the P-centre of the token with the modified burst was 5 ms earlier than that of the unmodified token. This result is strange, however, and seems contrary to most other data on the P-centre which would either indicate no effect or possibly a shift later.

Howell (1984) modified the envelope of a naturally produced /ʃa/ syllable by ramping up a portion of the initial frication (the first 40 or 120 ms of 148.8 ms) and ramping down the vowel over its entire duration (312 ms) producing stimuli that were perceived as /tʃa/ and /ʃa/ for short and long ramps respectively. An effect of the envelope was found: the P-centre shifted later as the onset time increased.

Building upon Howell's work, Scott investigated the effect of rise time resulting from linear ramps applied to the onset of naturally produced /wa/ and /æ/ syllables and a /tʃa/–/ʃa/ continuum (comprising a natural vowel prefixed by synthetic fricative). The segment durations and rise times varied for each stimulus (the consonant and vowel durations of /tʃa/ were 210 and 494 ms respectively and the ramp durations used were 10, 60, and 120 ms; the duration of /wa/ was 433 ms and the ramp durations were 0, 120, and 240 ms; finally the /æ/ duration was 213 ms and ramps of 10, 50, and 90 ms were used). The results showed almost no effect of rise time on the P-centre for the /tʃa/ sound whereas, for the other sounds, the P-centre shifted later as rise time increased (slope ≈ 0.3). It is worth noting that these onset ramps were applied by multiplying the existing onset envelope, so while it appears to be the case that rise time does affect the P-centre (in

some cases) it is not possible to directly generalise the effect size from these results.

Prior to Howell's first investigation, Tuller and Fowler (1981) attempted to examine the effect of amplitude on the P-centre by amplifying the sound waveforms to the point of clipping. Though they referred to the technique as infinite peak clipping, Howell (1988) noted several problems with their execution. Problems notwithstanding, their principal finding was that participants found sequences with the original, naturally produced timing more regular than those with altered, objectively isochronous timing whether the sounds were peak clipped or not. They concluded that neither the peak increment of spectral energy nor the amplitude characteristics in general play an important role in the perception of isochrony (and hence the P-centre). On the first point, however, they provided no evidence that they actually examined the peak increment in spectral energy as defined by Marcus (1976), a sub-band of about 1000 Hz that may well show increments even though the overall signal does not. On the second point, their experiment could only reveal an effect of the peak clipping manipulation if the effect was sufficiently large to make the objectively isochronous peak clipped waveforms sound more regular than the naturally timed versions. It was therefore premature to conclude that there was no effect.

Fowler, Whalen, and Cooper (1988) responded to Howell's critique of the original peak clipping manipulations and methodology by creating new stimuli that were more evenly clipped. The waveforms shown in their paper still do not appear to be infinitely peak clipped, though the overall signal envelope is closer to rectangular. Their results show that peak clipping a /ba/ syllable had negligible effect on the P-centre, whereas the effect of peak clipping a /sa/ syllable was generally to shift the P-centre earlier. The shift was largest when only the consonant was clipped, smallest (or even reversed) when only the vowel was clipped, and between the two when the entire syllable was clipped.

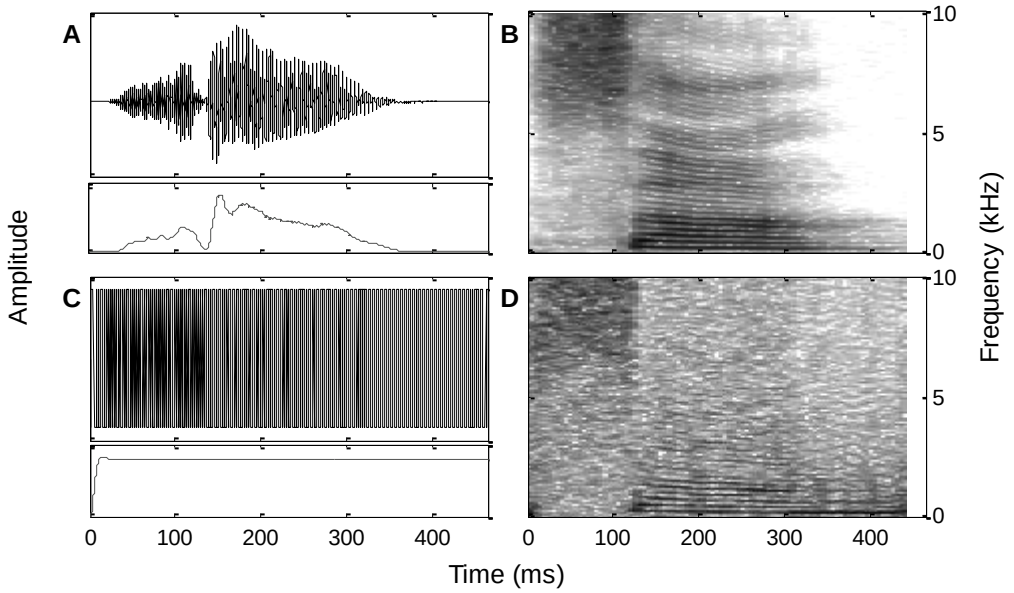


Figure 2.1 The technique of infinite peak clipping applied to the syllable /sa/. (A) The original sound waveform and envelope (full wave rectified and low pass filtered at 50 Hz); (B) spectrogram of the original signal (90 dB dynamic range shown); (C) the infinitely peak clipped waveform and its envelope; (D) the spectrogram of the peak clipped sound (dynamic range reduced to 50 dB to increase contrast for presentation purposes).

Scott (1993), taking care to execute the infinite peak clipping manipulation correctly, showed that P-centres of peak clipped "la", "ya", "ra", and "wa" syllables were earlier than their natural equivalents. P-centres of peak clipped stimuli, nevertheless, were still later than the P-centre of the reference sound (a 50 ms noise burst), despite their rectangular instantaneous envelope. Figure 2.1 illustrates the effect of infinite peak clipping on a /sa/ syllable and it is clear that while there is a gross distortion of the signal amplitude, the spectral structure remains at least partially intact. In particular it is possible see that even in the peak clipped spectrogram there is an offset of high frequency energy and an onset of lower frequency energy at about 130 ms.

Pompino-Marschall (1989) investigated an alternative to the peak clipping approach which he expected would have much the same effect (i.e. that it would result in a rectangular envelope). Specifically he modified a synthetic

/ʃi/ syllable so that the amplitude of the consonant and vowel were identical. (In practice, this manipulation is much less damaging to the original sound than infinite peak clipping because the signal is never in fact clipped.) His manipulations of both the initial consonant duration and vowel duration showed effects on the P-centre that were broadly similar to the equivalent /ʃi/ syllable with more natural envelope. Nevertheless, the small differences were significant.

### *2.1.6 Language and phonetic effects*

Almost all speech specific P-centre investigations used English only. Can the P-centre be produced (and perceived) equally by non-English speakers? Are there P-centre effects which seem to depend specifically on phonetic categorisation? The following investigations addressed these questions.

Hoequist (1983) examined the ability of speakers of English, Spanish, and Japanese (languages which are nominally stress timed, syllable timed, and mora timed respectively) to produce isochronous sequences of heterogeneous syllables, "a, ma, ba, pa" and "sa". He found that all were able to produce isochronous sequences. He further found that the P-centres must have occurred after consonant onsets but before the vowel onset (where an initial consonant was present), though he ignored the relative nature of his paradigm in reaching this conclusion. Most importantly, however, he concluded that the onset anisochronies produced by all speakers were consistent with a P-centre explanation and that this concept was therefore not specific to language rhythm categories.

When Marcus truncated the initial consonant duration for "seven", by truncating the initial frication, he reported that the token was categorically perceived as "seven"-"devon" (1981). Although his data showed there was no effect of this categorical change on the P-centre, Cooper et al. criticised his conclusions because he had not formally tested the categorical perception (Cooper, Whalen & Fowler 1986). In response, Cooper, Whalen,

and Fowler (1986) manipulated the onset duration of a /ʃa/ and a /sa/ syllable (by deleting frication and inserting silence respectively) and formally determined that the initial consonants were categorically perceived as /ʃa/-/tʃa/-/ta/ and /sa/-/sta/. Nevertheless, their results confirmed those of Marcus: the P-centre varied smoothly with onset duration but showed no effect of phonetic categorisation, even at the transitions between categories. This result was replicated once more by Pompino-Marschall (1987).

Fox and Lehiste (1987b) investigated the effect of varying the vowel quality on the P-centre in syllables with identical initial and final consonants: /siːt, sɪt, seɪt, sɛt, sæt, sat, sʌt, sɔt, suːt, soʊt, saɪt, saʊt, soɪt, sɚt/[7]. Using a forced choice method they found that the relationship of vowel duration to the P-centre was significant only when /soʊt, saɪt, saʊt, soɪt, sɚt/ were excluded. A subsequent experiment in which the vowel durations of /siːt, sɪt, seɪt, sɛt, sæt, sat, sʌt, sɔt, suːt saɪt/ were edited to be constant exhibited no P-centre effect. Thus Fox and Lehiste concluded that, at least for monophthongs, the vowel quality per se has no effect on the P-centre and only its duration is important.

Though it has not really been addressed rigorously in perception experiments, there is some evidence that the segment duration effects on the P-centre already reported may depend somewhat on the phonetic class of the consonants involved (see for example Harsin 1997; Janker 1996a; Pompino-Marschall 1989).

### *2.1.7 Affixes, Disyllables, and longer sequences*

Most P-centre research with speech uses (isolated) stressed monosyllables which all theories agree should have a single P-centre. There are significantly fewer investigations which examine longer sequences. Such

[7] IPA transcription of Fox and Lehiste's orthography was obtained from Perez (1997)

longer sequences are particularly valuable, however, as they provide a glimpse of how P-centres may be perceived in continuous speech.

Fox and Lehiste (1987a) investigated the perceptual effect of unstressed prefixes and suffixes on the P-centre. The words "peal, pealer, pealing, appeal, appealer" and "appealing" were edited so that the same initial [ə] sound was used where needed and the intervocalic [$p^h$] was the same duration in all cases. The results of a forced choice task indicated that the addition of an unstressed suffix appeared to shift the P-centre later but the effect was non-significant. The unstressed prefix, however, had a significant effect on the P-centre shifting it approximately 250 ms earlier than the words without prefix. Unfortunately, Fox and Lehiste did not specify the duration of the initial [ə] and so it is not possible determine whether the shift is closely related to its duration.

Bell and Morishima (1994) reported several results from manipulations on Japanese disyllables, whose accent patterns depend on pitch and not duration. First, the P-centre shifted with first syllable onset duration but was unaffected by placement of the accent on the first or second syllable—the slope of the relationship (0.62–0.76) was smaller than typically reported for monosyllables however. (A subsequent experiment suggested that accent placement did have a small effect, but this result may have been confounded by other factors.) Second, the P-centre shifted somewhat later as "tail" duration (incorporating the first syllable vowel and entire second syllable) increased—the effect was broadly similar to that of the rhyme duration in monosyllables. Third, the P-centre was unaffected by compensatory duration changes made to the first syllable vowel and second syllable consonant. They concluded that the two main effects (syllable onset and rhyme duration) found in monosyllables can be extended to unstressed disyllabic words and that P-centres are not literally equivalent to stress beats because the P-centre was unaffected by accent placement when it did not simultaneously alter the duration or amplitude.

In related work, Bell and Biasca (1994) examined the effect of English disyllable manipulations on the P-centre. They found an effect of onset duration similar to that of monosyllables but somewhat larger than usual (slope = 1.26). The P-centre of initially stressed and finally stressed disyllables shifted by approximately equal amounts (similar to shifts in monosyllables) when the first syllable onset duration was manipulated but the remaining durations held constant. The P-centres of finally stressed syllables occurred later than those of initially stressed syllables but by an amount which was only about half the interval between the first and second vowel onset. Together, these interesting results appear to show that the P-centres[8] of finally stressed disyllables depend more on the timing of the initial vowel onset than the second stressed vowel onset.

Only Allen's early investigations into the rhythm of English used continuous speech stimuli in perception experiments (1972a; 1972b). Although this method was flawed (as already noted), it nevertheless provides some insight into how P-centres in continuous speech may be evaluated.

### *2.1.8 Speech production versus perception*

P-centre perception studies generally have the benefit of control but are time-consuming to execute. As a result the data from perception experiments are relatively sparse. In contrast, production paradigms allow large amounts of data to be generated quickly. Of course individual productions are quite variable and so considerable data is still required to measure parameters accurately.

In one of the earliest relevant studies, Rapp-Holmgren (1971) investigated the stress beat (P-centre) of Swedish syllables by asking participants to produce nonsense speech tokens (/aˈsɑːd, aˈtɑːd, aˈdɑːd, aˈlɑːd, aˈnɑːd,

[8] It is probably not correct to refer to "the P-centre" of a disyllable as Bell and Biasca do. Whether a syllable is stressed or not it is still a rhythmic event with an associated P-centre. Thus, a better term would be the *stressed P-centre*.

aˈstɑːd, aˈstrɑːd/) in synchrony with a pacing sequence of clicks (IOI = 500 ms). She measured and averaged syllable segment durations across 40 productions of each token and then compared the average segment boundaries with the distribution of pacing clicks. Her results revealed a linear relationship between the stressed syllable initial consonant duration and the position of the mean pacing click relative to the vowel: as consonant duration increased approximately[9] 100–220 ms, the pacing click position shifted approximately 0–85 ms. It is worth noting that this meant the pacing click occurred well before the vowel onset with longer initial consonants, though results may have been confounded by slightly shorter stressed syllable rhyme durations for these same tokens.

Fowler (1979) conducted a number of experiments in which she investigated isochronously produced speech. In the first of these, a single participant produced homogeneous or alternating sequences using the syllables /ad, bad, mad, nad, tad, sad/. The results indicated that onsets were nearly isochronously produced for homogeneous sequences and for the same order within alternating sequences. In contrast IOIs for alternating sequences exhibited systematic differences of isochrony which were closely related to the prevocalic (consonant) duration.

In a second experiment, Fowler found that when participants were forced to choose the more "rhythmic" sequence between naturally produced onset anisochrony and edited onset isochrony, they chose the naturally produced timing at with significantly greater than chance frequency. This result is exactly what would be expected based on the majority of perceptual P-centre experiments. Scott (1993) conducted a more sophisticated version of this experiment. Using naturally produced "one" and "two" tokens from seven speakers her results showed that produced tokens were objectively anisochronous (although the amount depended on speaker). Unlike Fowler, who only forced discrimination between two coarse timing categories

[9] The accuracy of these values is limited the lack of tabulated values; these values were determined from graphs.

(naturally anisochronous or objectively isochronous), Scott asked participants to adjust the timing of exemplar token productions until they were perceptually isochronous. She found that the adjusted tokens exhibited anisochronies that were similar to those produced. Thus the notion that speakers produce exactly the objective anisochrony required to be perceptually isochronous was supported.

Fowler's third experiment required speakers to produce the words "acts, bats, mats, gnats, tacks" and "sacks" in the framing sentence "Jack likes black —". Like in her first experiment, Fowler's results showed that long intervals preceded words with short initial consonants while short intervals preceded words with long initial consonants. After discussing the articulation and quality of various consonants, Fowler hypothesised that to produce a (perceptually regular) stress timed utterance speakers initiate the production of stressed syllables at regular intervals and that moreover listeners judge rhythmicity by inferring the articulatory timing from the acoustic signal.

A fourth experiment examining vocal reaction time to speak a syllable in response to a visual prompt found minor differences between reaction times to syllables with different initial consonant classes (affricates, were later than stops, which were later than the remaining classes). Fowler took these results as evidence that speakers start producing responses at approximately the same time and that the resulting anisochronies (whether in a reaction time task or in ordinary speech) are a natural consequence of the articulation and not an attempt to achieve a specific perceptual timing.

In Fowler's final experiment she investigated the produced timing of /bad, dad, gad/ both with and without pre-voicing of the initial consonants. Pre-voicing changes the acoustic realisation of the consonants but has almost no effect on the articulation. Therefore the prediction was that if the P-centre was primarily an articulatory phenomenon there would be no anisochrony between the initial stop releases of voiced and pre-voiced consonants,

whereas if the P-centre was primarily acoustic in nature, then anisochrony would be observed. The results supported the articulatory hypothesis.

Fowler and Tassinary (1981) investigated the produced timing of homogeneous and alternating syllable sequences both with a metronome, replicating Rapp-Holmgren's method (1971), and without. The syllables used were /ad, bad, dad, fad, mad, nad, pad, sad, tad, stad, trad, strad/ which varied in both the phonetic class and complexity of the syllable onset. Their results closely resembled the earlier results of Rapp-Holmgren. The metronome pulse generally fell within the syllable onset but the specific location appeared to depend on both phonetic class and onset duration. They suggest that it may be the articulatory onset (rather than the realised acoustic onset) of the nuclear vowel that is regularly timed.

Perez (1997) examined segment duration effects on naturally produced monosyllables in a series of experiments, the first two of which replicated Fowler's first and third production experiment and obtained similar results. Using a framing sentence in which, to avoid certain confounding factors, the test word was no longer sentence final ("they like — mats"), her next two experiments found effects of both initial consonant duration and vowel duration on the produced timing of monosyllables that were broadly in line with previous research. Her data showed the initial consonant duration effect was quite large and linear (slope ≈ 0.75) whereas the weaker vowel duration effect (slope ≈ 0.25) was quite a bit more variable ($R^2$ = .18 to .32). A following experiment showed that the final consonant duration also had an approximately linear effect (slope ≈ 0.36) on the produced timing. In all these experiments the tokens were naturally produced and so variations in one segment duration were not completely independent of variations in others though the degree of interaction may have been small.

Perez also investigated the effect of segment duration on the naturally produced timing of disyllables, complementing previous perception experiments (Bell & Biasca 1994; Bell & Morishima 1994). Over the series of experiments both initially stressed and finally stressed disyllables were

tested. The effect of initial consonant duration on disyllables was very similar to that of monosyllables. The effect of medial consonant duration was present but somewhat smaller (slope ≈ 0.3) and more variable. The effect of final consonant duration was not significantly different from that of medial consonant duration. Her data showed that finally stressed disyllables were produced significantly earlier than initially stressed disyllables, a result that is compatible with the interpretation that finally stressed disyllables have later P-centres than initially stressed disyllables. Furthermore, there was no significant difference in word durations to confound this effect.

Fox and Lehiste reported two production experiments which were matched to equivalent perception experiments (1987a; 1987b). In the first of these, using syllables whose nuclear vowel quality was manipulated (varying the duration by side effect), there was a tendency for the vowel onset to be produced earlier as the vowel duration increased, though this tendency was only reliable for monophthongs. This matched their perception findings. In the second experiment the production effect of unstressed prefixes ("a-, de-, con-") and suffixes ("-er, -ing, -able") on stressed monosyllables having a variety of initial and final consonantal classes was examined. Results were again similar to those in perception. Addition of a suffix tended to shift the measurement point (e.g. the onset of vocalic energy) somewhat later relative to the base form but this effect did not reach significance. Addition of a prefix shifted the measurement point earlier (27–86 ms depending on prefix) and this effect was significant.

Investigating acoustic and kinematic candidates for the P-Centre Patel, Lofqvist, and Naito (1999) asked subjects to produce sequences consisting of eight pairs of alternating syllables. The first syllable was always /ba/ while the second syllable was taken from the set /tʃa, ha, sa, ja, la, ma, pa, ta, lad, spa, deˈla, li/. Subjects produced the sequences without a rhythmic aid after a brief practice trial with a metronome having a tick interval of 500ms. Patel found that the onset anisochrony between syllables in these

sequences was systematic and stable "suggesting that speakers have a clearly defined focus in their timing strategy" (p. 3). His results showed that /pa, ta/ exhibit the least anisochrony, followed by /ja, la, li/, then /tʃa, ma, lad/ and finally /ha/. The syllables /sa, deˈla/ exhibit substantially greater anisochrony. All of the previous syllables exhibit negative anisochrony; their onsets occur earlier than physical isochrony would require. While the /spa/ syllable also exhibits negative anisochrony, the /pa/ part of this syllable exhibits slight positive anisochrony which is quite different from the simple /pa/ syllable. These results are broadly compatible with previous production and perception experiments.

### *2.1.9 Articulatory correlates*

Perception studies of the P-centre implicitly assume that the P-centre is based on the acoustic waveform only. Fowler in particular has argued that the P-centre may in fact be an articulatory feature of speech. Several experiments have been conducted to examine this hypothesis.

Tuller and Fowler (1980) investigated possible articulatory correlates of the P-centre. Participants were asked to produce the monosyllables /bak-fak/, /duk-suk/, and /dup-sup/, either in alternation (as shown) or by repeating just one of the syllables. Electromyography (EMG) measurements of the Orbicularis Oris-Inferior (used for lip rounding) showed smaller departures from isochrony than acoustic onset measures. Their results did not, however, identify which, if any, of the articulatory gestures actually corresponded to the P-centre.

Fowler (1983) subsequently reported a set of three experiments investigating the hypothesis that vowels are produced cyclically in sentences composed of monosyllabic stress feet. The first experiment made use of a categorical perception illusion: the final consonant in /ad/ may be perceived as /d/ or /t/ depending on the perceived duration of the preceding vowel. The perceived duration of the vowel can in turn be

affected by a preceding consonant. The results indicated that a variety of initial consonants induced minor perceptual duration changes (approximately 10 ms) in the vowel. A second experiment using reaction time indicated that vowel identity is signalled by co-articulation occurring before the conventional acoustic vowel onset time, though there was an interaction between initial consonant type and vowel type. The third and final experiment revealed that the mean tap asynchrony was closely correlated with the vowel identity reaction time which Fowler interpreted as evidence that the perceived timing of syllables is in fact the perceived timing of vowels.

De Jong (1994) examined the relationship of articulatory gestures to the P-centre over the course of two experiments. In the first experiment, stimuli were 12 productions each of "toast" and "totes" which varied in their degree of naturally produced accent (and as a consequence in all of their segmental boundaries). Listeners adjusted sequences of alternating stimuli to perceptual isochrony so that P-centres could be estimated and compared with articulatory information that had been recorded with the original token productions. The results indicated that timing of the tongue tip minimum predicted as well as voice onset timing whereas other articulatory events under or over-predicted measured P-centres. The second experiment attempted to distinguish between the acoustic and articulatory predictor using tokens that differed in initial aspiration ("gap/cap, gob/cob, gab/cab, dot/tot" and "dab/tab"). In this case the acoustic and articulatory measures which performed best were different than the first experiment. Though various articulatory features predicted the P-centre as well as acoustic features there was no single articulatory feature which predicted all P-centres well.

Complementing their production experiment, Patel et al. (1999) measured the primary articulator velocity and jaw velocity for each of the syllables /tʃa, ha, sa, ja, la, ma, pa, ta, lad, spa, deˈla, li/. The choice of primary articulator depended on the syllable. Although their results indicated that

the interval between primary articulator velocity maxima was more nearly isochronous than the acoustic onsets (which had been produced with systematic anisochronies as expected), these articulator intervals were still significantly different from isochrony for many of the syllables.

Therefore, no definitive articulatory correlate of the P-centre has been found.

### *2.1.10 Acoustic envelope and duration*

Several psychoacoustic effects of envelope and duration are known. For example Efron (1970a; 1970b; 1970c) found that the minimum perceived duration of an acoustic stimulus appeared to be about 130 ms. Envelope also appears to affect perceived duration in non-symmetric ways: damped sounds with gradual offset are perceived to be shorter than ramped sounds with gradual onset (Grassi & Darwin 2001; Schlauch, Ries & DiGiovanni 2001). Furthermore, perceived loudness is affected by duration (e.g. Buus, Florentine & Poulsen 1997; Epstein, Florentine & Buus 2001; Florentine, Epstein & Buus 2001; Glasberg & Moore 2002; Heil & Neubauer 2001; Zimmer, Luce & Ellermeier 2001). The specific confounding factors that these psychoacoustic phenomena introduce have not been specifically investigated and existing results must be viewed cautiously as a consequence.

Although music is perhaps the most obviously rhythmic activity, initial P-centre investigations were focused on speech and monosyllables specifically. Do effects equivalent to those observed for syllable segment durations and envelope arise with non-speech stimuli? Several researchers have examined these questions.

Vos and Rasch (1981) examined the effect of rise time on the P-centre[10] of sawtooth tones (400 Hz). They used an adjustment paradigm that differed in one important respect from those typically employed: adjustments to the timing of the test sound were achieved by altering its onset time while keeping its offset time fixed; thus the duration of the test sound changed with each adjustment and this is a confounding factor on their results. Nevertheless, they found that increasing the rise time shifted the P-centre later. Vos and Rasch interpreted their results as being compatible with a simple threshold based explanation of the P-centre.

Gordon (1987) measured the perceptual attack time (P-centre) of 16 re-synthesised instrumental tones with varying timbres (including differences in onset time and shape). He found that the difference between the earliest and latest P-centres was 49 ms and that the P-centres of sounds with impulsive onsets were very close to the perceptual onsets of those sounds. For sounds with more gradual onsets, the P-centre appeared to depend somewhat on the timbre of synchronous sounds; an impulsive sound could possibly mask part of the onset of a gradual onset sound. He found that the P-centres in his data were best explained by a combination of an onset threshold delayed by a fraction of the rise time. Even with musical instruments, however, he was forced to introduce heuristics to handle non-monotonically increasing onsets and it seems likely that with more complex onsets the P-centres would not be well explained by this rather simple approach.

Using a synchronous tapping paradigm, Vos, Mates, and van Kruysbergen (1995) investigated the effect of duration (1, 2, 50, 300 ms) on the mean tap asynchrony (assumed to co-vary with the P-centre) of a square wave tone (440 Hz) having a rectangular envelope. The results showed that the P-centre shifted later as the duration increased. The effect may have been

[10] They used the term *perceptual onset* but their paradigm was essentially identical to that of researchers investigating P-centres so it would seem that the percept they actually measured was the P-centre.

linear up to 100 ms (slope = 0.2) but was weaker between 100 and 300 ms (slope = 0.06). It is worth noting that without correction the perceived loudness of the tones would vary with duration, an effect that may also become less pronounced at 300 ms.

In a second experiment, Vos et al. varied the rise time (0, 40%, and 80% of duration) of 500 Hz square wave tones having a number of durations (2, 50, 100, and 300 ms). The mean tap asynchrony (and P-centre) shifted later as rise time increased but the size of the effect depended somewhat on stimulus duration. A third experiment revealed no significant effect of tempo (IOI = 500, 700 or 900 ms) on the mean tap asynchrony. A possible confounding factor in the results is that subjects were instructed to keep tapping speed and duration as constant as possible; previous experiments in the set had indicated that the duration of a tap increased as the duration of the stimulus increased.

Howell (1984), as part of his investigation into the effect of envelope on speech applied the envelope of a modified /ʃa/ syllable to a synthesised sound comprised of white noise and a sawtooth tone whose durations matched those of the fricative and vowel respectively. The P-centre of the sound with short (40 ms) onset time was earlier than that of the long (120 ms) onset time, but the effect was weaker than for modified speech.

Scott (1998) examined the effect of onset time (5–75 ms) and offset time (5–75 ms) on the P-centre of a constant duration (200 ms) synthetic /a/ vowel. Onset and offset ramps were both linear. There was a significant effect of onset time (slope = 0.235) and a small non-significant effect of offset time (slope = -0.05).

In a somewhat similar experiment, Seton (1989) investigated the effect of onset time (40–160 ms) and level (65, 75 dB SPL) on the P-centre of a synthetic /a/ vowel. In this case, the offset was cosine shaped whereas the onset provided linear power increase (decelerating amplitude). The results revealed very little effect of rise time, but this may have been a consequence

of the initially rapid amplitude increase. Seton argued that the result is most compatible with a threshold interpretation of the P-centre.

Scott also examined the effect of duration (76–280 ms) on the P-centre of a synthetic /a/ vowel (1998). She found a weak non-significant relationship (slope = 0.04) between duration and P-Centre. Because of the small effect size, it would be premature to conclude that there is no effect of duration. Furthermore, there is some suggestion from the results that the duration effect may not be linear and may get weaker at longer durations. A confounding factor in the results is that the stimulus rise time (5ms) is both more abrupt than encountered in natural speech and shorter than the glottal period of the vowel (8.8ms). It is possible that stimuli with less abrupt onsets would yield different results.

Seton also investigated the effect of duration (50–250 ms) on the P-centre. Using a sawtooth tone (400 Hz) with cosine shaped onset and offset (10 ms each) he found a weak effect of duration (slope = 0.1) on the P-centre and there was a tendency for this effect to become non linear (weaker still) at long durations. Vos and Rasch's interpretation that the P-centre could be represented by a simple threshold already appeared to be discounted by syllable rhyme duration results, but perhaps the effects were different for speech. Seton's result showed that the duration effect, although weaker for non-speech, was present and this could not be handled by a simple threshold explanation.

Seton (1989) also investigated the ability of a participant (himself) to make reliable adjustments to synthetic tones that were not isolated from one another as is typical but instead were synthesised as amplitude "bumps" over a constant pedestal level sound. Though the task became subjectively difficult at low signal to pedestal levels, the results suggested that the adjustments were reliable. This may prove to be a useful experimental method to bridge the gap between general acoustic investigations and continuous speech.

In general, then, the results appear to show that, similar to speech stimuli, general acoustic stimuli exhibit a (possibly weak) effect of both onset time (rise time) and overall duration.

### *2.1.11 Level and loudness*

Most P-centre investigations do not specifically investigate the effect of presentation level or perceived loudness. The most typical configuration is for sound presentation to be at a "comfortable level". Nevertheless investigations into presentation level are particularly relevant to any explanation of the P-centre in terms of a threshold effect. Even if the P-centre is not primarily a threshold effect it seems reasonable that it would be affected by threshold effects due to gain control adaptation and perceptually relevant dynamic range constraints in hearing.

Vos and Rasch (1981) manipulated sawtooth tones so that not only their onset time varied but also their sensation level (level relative to silence and a masker level). Over the course of three experiments, they found that the threshold (relative to peak level) which best explained the P-centre shifts decreased (from -7 to -15 dB) as the sensation level increased (from 20 to 70 dB).

Seton (1989) questioned whether the P-centre could be shown to be distinct from the perceptual onset, i.e. the moment of detection, of a sound. Using a sawtooth tone (400 Hz) with a fixed duration (250 ms) he examined the effect of onset duration (cosine shaped, 5–200 ms) and level (60, 70, and 80 dB SPL) on the reaction time. His results showed two effects: the mean reaction time was later when the level was lower and mean reaction time shifted later as the rise time increased (the shift depended on the level; the maximum shift was 22, 28, and 40 ms for 80, 70, and 60 dB respectively). This first experiment had grouped all identical levels together for presentation. A subsequent experiment which grouped mixed levels together found similar results. Seton compared these results to those of Vos

and Rasch (1981) and showed that, if interpreted identically, the reaction time results would correspond to lower thresholds (shifting from -20 to -30 dB compared to Vos and Rasch's -10 to -15 dB over the same range of levels). Another interpretation is that the point to which the participant reacts (e.g. the perceptual onset) is not the same as the point used to adjust sounds into perceptual isochrony (the P-centre).

In a following experiment, Seton measured P-centres for the reaction time stimuli using the adjustment paradigm. In this case the results for blocked levels (i.e. mixed levels did not occur in a single trial) showed no effect of level and a linear effect of rise time (slope ≈ 0.22). This appears to support to idea that the P-centre and perceptual onset are different points. Results for mixed levels exhibited the same general trends but a dependency on level which resulted in smaller shifts for higher levels. Seton interprets the result as evidence that P-centres may exhibit context dependence on the level of preceding and succeeding events in a sequence. If the hearing system adapts continuously to the short term average sound level, then it is easy to imagine that the onset of quiet sound following a loud sound may be more difficult to detect (or alternatively that it will be detected only at a higher level relative to the sound's peak).

### *2.1.12 Frequency, streaming, and compound events*

The research reviewed to this point makes it obvious that the vast majority of investigations have focused on relatively simple manipulations of duration and amplitude envelope. There appear to have been just two investigations of frequency specific effects despite natural speech, in particular, incorporating continuous pitch and spectral peak (formant) modulations. Similarly, although several investigators have commented on problems that appear to be attributable to auditory streaming effects (Bregman 1990/1999) such effects have been directly investigated just once.

Janker and Pompino-Marschall (1991) investigated the effect of pitch ($F_0$) manipulations on the P-centre of an edited /ka/ syllable whose duration and amplitude was matched to a single template production. They based their pitch alterations on the five Thai tones: low, mid, high, fall, and rise. Only the fall and rise tones exhibited substantial pitch changes (54 and 77 Hz respectively) and in both cases most of the change occurred in the latter half of the vowel duration. The results (with just two participants) indicated that the P-centre of the rising tone was delayed by 17 ms relative to mid (almost constant) tone but that no significant P-centre differences were found between the remaining tones. This is an important finding that deserves further investigation. Existing P-centre explanations have almost nothing to say about the effect of pitch. If the effect of pitch was confirmed, then existing explanations would require modification.

Using stimuli based on those of van Noorden (1975), Seton investigated the effect of auditory streaming on P-centre perception (1989). Using low and high frequency tones (1000 and 4000 Hz) with fixed duration and envelope (30 ms steady state with 5 ms cosine shaped onset and offset) he measured the P-centre in low, high, and mixed frequency conditions. The results showed that the P-centre of the high frequency tone occurred 9 ms later on average than that of the low frequency tone. This intriguing result may indicate an absolute frequency dependent effect on the P-centre or it may be a psychoacoustic artefact: equal loudness curves (ISO/TC43 2003) predict that the 4000 Hz tone would have been perceived almost 10 phon louder than the 1000 Hz tone.

In a related experiment, Seton wished to determine whether listeners could attend selectively to noise and periodic components (which may stream apart under repetition) in a single compound sound (1989). Stimuli were composed of noise bursts (65 dB SPL, 250 ms duration, 50 ms linear onset and offset) and a pure tone (1000 Hz, 75 dB SPL, 50 ms duration, 5 ms linear onset and offset) added at one of several delays (0, 50, 100, or 150 ms). His results showed that participants could indeed attend selectively to

either the noise or tone component of the sound. When asked to attend to the noise, there was no effect of tone delay on the P-centre whereas when asked to attend to the tone, the P-centre shifted later as the tone was delayed although the size of the shift was not 1:1 (150 ms delay resulted in 108 ms shift). Because the shift was not 1:1, streaming alone cannot explain the results—if it did then the noise should have no effect. An alternative interpretation is that the amplitude "bump" (or perhaps some combination of the amplitude and spectral change) was the main contributor to the P-centre shift, but this would need to be investigated by replicating the experiment with sounds that do not have such quality differences. The most similar experiments in the literature are those in which Pompino-Marschall (1989) matched the envelope of square wave tones to synthetic syllables and those stimuli did elicit P-centre changes which may be compatible with Seton's results.

Finally, in recent work, Hove, Keller, and Krumhansl (2007) investigated the effect of small asynchronies (25–50 ms) between the constituent tones of chords. Such asynchronies could potentially be expected in natural performance on the basis of previous research (Rasch 1979). Their results indicated that the P-centres of chords with asynchronies were later than those of synchronous chords.

## 2.2 Theoretical review

Two theoretical frameworks have principally been used to analyse the P-centre phenomenon and frame hypotheses upon which to base investigations. As is apparent from the previous survey of empirical data, the majority of existing P-centre research implicitly assumes that the P-centre of an acoustic stimulus is based on the acoustic constitution of that stimulus only (see for example Gordon 1987; Howell 1984, 1988; Janker & Pompino-Marschall 1991; Marcus 1981; Pompino-Marschall 1989; Scott 1993; Vos, J. & Rasch 1981). In contrast, the competing theory hypothesises that the P-centre is determined by articulatory gestures in speech

production (see for example de Jong 1994; Fowler 1979, 1983, 1996; Patel, Lofqvist & Naito 1999) or, for more general events, perhaps by the sound producing mechanisms and actions underlying those events (Fowler 1996). In this work, these competing theories are termed the *acoustic theory* and the *articulation-production theory*.

### *2.2.1 The acoustic theory*

The principal prediction of the acoustic theory is simply that the P-centre should arise from one or more features of the acoustic stimulus only. Perhaps because it is best to discount simple explanations before invoking more complex accounts, much of the early acoustic P-centre research examined just one or two such acoustic features. The evidence to date, however, does not support such simple accounts of the P-centre phenomenon. In particular, results from multiple experiments manipulating the duration of the rhyme of monosyllables or the tail of disyllables indicate that is not sufficient to relate the P-centre to a single threshold (Vos, J. & Rasch 1981) or features within the acoustic onset alone (Gordon 1987; Rapp-Holmgren 1971; Scott 1993). Marcus's modelling of the P-centre as a function of syllable segmental durations alone is undoubtedly too simplistic also. Notwithstanding the notion of segment duration for non-speech sounds being problematic, Tuller and Fowler (1980) rightly criticise the questionable psychological significance of the segment boundaries typically chosen. For example, Marcus defined the vowel onset as the peak increment in mid band energy, a choice based on signal processing tractability rather than psychological significance; researchers working with oscillograms (waveforms) often choose the vowel onset as the point where the vowel periodicity becomes evident, but the vowel is often signalled much earlier by co-articulatory transitions within the consonant.

If the P-centre is not determined by the timing or magnitude of just one or two acoustic features, on what does it depend? Howell (1984; 1988) proposed that acoustic energy might be integrated in such a way that its

*centre of gravity* might represent (or at least co-vary with) the P-centre. Although this theory makes predictions that are often qualitatively correct it has been subjected to close and repeated examination which has eventually concluded that it is not viable in its basic form (see for example Fowler, Whalen & Cooper 1988; Scott 1993; Seton 1989). Howell went on to suggest that the acoustic energy may be pre-weighted in some manner, prior to integration in the centre of gravity calculation, but what form such pre-weighting might have has never been made clear.

Although Howell's centre of gravity P-centre theory is not directly supported by the empirical evidence, it has influenced researchers who have used a centre of gravity calculation to integrate the effect of acoustic features temporally distributed throughout a sound in order to model P-centre perception (Harsin 1997; Pompino-Marschall 1989). Even if centre of gravity style integration does not prove to be correct, it now seems certain that P-centres modelled according to the acoustic theory will have to rely on some, possibly complex, integration of acoustic features.

### 2.2.2 *The articulation-production theory*

In contrast to the acoustic theory, the articulation-production theory of P-centres predicts that it is the produced articulatory gestures of speech that define the P-centre rather than their acoustic side effects. Fowler (1979) argues that acoustic anisochronies "do not arise because the talker *intentionally* causes the onsets of acoustic energy [...] to occur when they do. Instead the anisochronies are a by-product of the talker making articulatory gestures at a stress-timed rate." (p. 382). Therefore, isochronous production of speech should be relatively straightforward: a speaker should simply time their articulatory gestures isochronously.

The prediction that articulatory gestures define P-centres has been tested a number of times (de Jong 1994; Patel, Lofqvist & Naito 1999; Tuller & Fowler 1980) but as yet no evidence for a universally reliable articulatory

predictor has been reported; on the contrary recent research concludes that none of the articulatory candidates examined to date appears to be the cue underlying the P-centre (Patel, Lofqvist & Naito 1999).

Speech production is a complex motor task requiring the co-ordination of several articulators (including the lips, tongue tip, and jaw). The movements of these articulators overlap in time and there simply may not be a single articulatory feature which determines the P-centre. If the articulation-production theory is to remain viable in any form then it seems there are only two possible solutions: either the specific articulatory gesture which determines the P-centre depends on the sound being produced (e.g. for one syllable it may be the jaw and for another it may be the lips), or else the P-centre is defined by some integration of the overlapping gestures. The evidence to date does not appear to support the first alternative and the second has not yet been investigated in detail.

It is also important to observe that the articulation-production theory (as typically framed) is limited to speech stimuli (but for a more general formulation see Fowler 1996). As a consequence, the typical interpretation of the theory is that P-centre mechanisms for speech and non-speech[11] stimuli must be separate and distinct. This would imply that non-speech musical sounds may be timed differently than speech sounds and raises a number of questions. For example, why should humans have evolved multiple mechanisms, and how should speech and non-speech be synchronised (in the case of vocals and accompanying instrumentals in a song, for example).

Fowler et al. (1988) predicted that listeners use the acoustic consequences of natural sound-producing events as information about the events themselves. This "direct realist" explanation of P-centres (Fowler 1996) is a

[11] Fowler argues that non-speech is not a class of sounds in the same way that speech is—it may well correspond to several distinct classes of sounds. A more accurate distinction for the articulation-production theory may be between sounds produced by the human vocal apparatus and those produced by other means.

more general expression of the articulation-production theory since the class of sound producing events extends beyond speech to non-speech vocalisations, animal and natural sounds, and possibly acoustic instruments. It is more difficult to imagine how it might apply to sounds produced by unnatural means, sounds subjected to extensive signal processing, or wholly synthetic sounds. In particular Fowler et al. (1988) state that they can make no predictions for how the P-centre of synthetic sounds without an identifiable distal source might be determined. This is problematic as modern cinema and popular music are filled with such sounds, often presented with and perceived as having very specific timing.

### *2.2.3 Discussion*

It is surprisingly difficult to choose between the acoustic and articulation-production theories on the basis of empirical tests alone. The P-centre does not appear to be correlated with a single simple feature of either the acoustic waveform or the articulatory gestures involved in speech production (and its relationship to the mechanics of production in non-speech sounds has not been investigated at all). Without a single simple feature, some as yet unknown integration function must be invoked to combine multiple features and the approximation of this integration function is a P-centre model. Although one might imagine that a model based on articulatory features which successfully predicts measured P-centres would be a strong argument in favour of the articulation-production theory and against the acoustic theory, such a conclusion would need to be reached with great care. Unless a model has been validated against a very large corpus of stimuli, the possibility of other, perhaps more general, models will exist and evidence in favour of one theory will not automatically discount the other.

Another strategy that has been applied to distinguishing between the two theories is to make a modification to the acoustic realisation of a (speech) sound without significantly altering its articulation (Fowler 1979). The

hypothesis is that this will affect the P-centre if it is defined only in acoustic terms, but have no effect if it is determined only by the articulatory gesture. It is not, however, quite enough to alter the acoustic realisation; the acoustic signal must also be altered in a way that is perceptually salient. In particular changes which do not affect the mid frequencies at which hearing is most sensitive would seem likely to have a small effect or perhaps even no effect and this may well have been the case in Fowler's experiment. Nevertheless, the general approach has merit. If it could be shown that a set of sounds with similar acoustic properties but different production mechanisms had similar P-centres, this would support the acoustic theory. If, instead, the same set of sounds exhibited significant P-centre differences, then the articulation-production theory would receive more support.

A particularly important consequence of the articulation-production theory is that it appears to require multiple P-centre prediction mechanisms, because a listener must recover the timing as produced rather than as acoustically realised. Recovering a speech gesture, for example, would seem to be quite different than recovering the beat gesture of a drum or the bowing gesture of a stringed instrument. In contrast, the acoustic theory requires just one P-centre mechanism (although of course it does not preclude there being more). Perhaps, therefore, the discovery of a single P-centre model which could reliably predict P-centres in speech and non-speech sounds would provide the most compelling evidence in support of the acoustic theory.

The final question that might be considered is how the two theories help or hinder the practical development of a P-centre prediction model. Like a listener, a P-centre model has access only to the acoustic stimulus and not the production gestures directly. To derive gestures, then, would require interpreting the stimulus in the context of some internal model of gesture production and the auditory world. Therefore, a P-centre model based on the articulation-production theory would seem to require two complex processing stages: first, the production gestures as produced would need to

be approximately recovered from features of the acoustic stimulus; and second, relevant features from the recovered gestures would need to be integrated to obtain the P-centre. In contrast, an acoustic P-centre model seems to require just one complex processing stage, namely, the direct mapping (again via some complex integration function) from acoustic features to the P-centre. While the potential model complexity is not in itself a sufficiently strong argument in favour of one theory or the other, it certainly would seem to be more productive to continue modelling based on the acoustic theory in the short to medium term.

## 2.3 Questions and conclusions

### *2.3.1 Theoretical questions*

Perhaps one of the first and most fundamental theoretical questions that could be asked is why is there a P-centre in the first place? What is its purpose? Merker and colleagues suggest possible evolutionary benefits that entrainment to an isochronous pulse would have given human ancestors, including the ability to synchronise group activities such as chorusing to attract mates from greater distances (Merker 2000; Merker, Madison & Eckerdal 2009). It is the P-centre that permits individuals to synchronise their activities to external events (such as the activities of others) and it is the apparently universal nature of the P-centre that permits groups to have a common mutual understanding of what synchronisation actually means. Was this its original purpose?

The acoustic P-centre theory has an important implication for sound producers including speakers. They must produce their sounds, not by conveniently timing production gestures according to the desired timing, but by anticipating (or actively measuring in feedback) the acoustic consequences of their actions. Is there evidence that sound producers can do this? Certainly the literature on sensorimotor synchronisation (Repp 2005) would seem to suggest that humans can initiate relatively simple

motor actions early in anticipation of the desired effect, namely synchronising with an external stimulus. P-centre production experiments in which speakers synchronise with a metronome pacing sequence seem suggestive of a similar capability in the complex motor task of speech production.

As previously discussed, the most compelling evidence in favour of the acoustic theory would be the development of a single P-centre model which reliably predicts the measured P-centres of a wide variety of sounds, including speech, instrumental, and synthetic sounds. Can such a model be developed? A prerequisite step is certainly to measure the P-centres for a large corpus of sounds. Only then might a sufficiently general model be developed.

When considering the nature of the auditory P-centre a question that arises is whether auditory P-centre perception should be considered a dedicated and distinct perceptual process, or a side effect—an emergent property of the known (and unknown) psychoacoustic operation and constraints of the hearing system. Between stimulation and perception various transformations of the sound signal are known to occur; these include frequency dependent amplitude sensitivity, temporal integration effects, asymmetric onset/offset sensitivity, frequency masking, and temporal masking. Furthermore, research on the neurophysiology of hearing indicates low level (individual neuron) and high level (auditory cortex) sensitivity to signal changes which may go some of the way towards explaining the apparent importance of co-articulatory transitions in speech. In consideration of these factors, it seems like a useful approach to advancing the field of P-centre research may reformulate the acoustic theory in terms of psychoacoustic plausibility. That is, assume the P-centre depends on acoustic features of the sound only (not the production gestures) but select and evaluate candidate features on the basis of psychoacoustic plausibility rather than convenience of analysis.

Of course, this notion of the P-centre as an emergent percept does raise a number of related theoretical questions. For example, how is the P-centre in other modalities perceived? Could it be that the P-centre in each sensory modality is simply (or primarily) a side effect of the psychophysics of that modality? Furthermore, how might the P-centres from multiple sensory modalities be compared or integrated?

Finally, how is the P-centre related to event perception and segmentation? Are P-centres and events two co-dependent features of the same phenomenon? In particular, is it the case that if an event is perceived it must have a P-centre and if a P-centre is perceived, by definition, a new event has occurred? Considered from this perspective, the process of event segmentation may in fact correspond to P-centre detection. This is an intriguing question deserving further consideration. In particular it may shed light on the reliability with which speech can be parsed into syllables (events) despite the apparent ambiguity over where the syllable boundaries should be located. Nonetheless, significant progress towards a reliable P-centre model for continuous event sequences will have to be made before it can be examined empirically.

### *2.3.2 Empirical questions and replication*

From the review of empirical P-centre research it is clear that there are a large number of open research questions. Moreover, a number of findings deserve replication, either to confirm the original results, or to resolve inconsistencies between existing investigations.

Some of the principal open questions are as follows:

1. A variety of measurement methods have been used to estimate P-centres and in some cases particular results have been reported only for one method. Does the choice of method have a significant effect on the P-centre estimates? Can results obtained with different methods be integrated?

2. Does the P-centre depend on frequency? If so, is the dependency related in any way to the absolute frequencies or only to relative frequency difference? Finally, does static frequency have any role or is it just frequency modulations, and if the latter is it rate or amount of change that is most important?

3. On a related note, how does the P-centre depend on phonetic qualities (rather than categorical identities per se)? Fox and Lehiste (1987b) found that diphthongs (which feature a degree of spectral change over their time course) had a different effect than monophthongs. Other research has indicated some differences in the P-centre associated with consonant phonetic classes, but the investigations have been far from exhaustive. Though less easy to generalise, it would appear that investigations with synthetic speech could be productive in this regard.

4. Why are almost all duration or boundary effects weaker for synthetic than speech stimuli? Is it because more spectrally complex speech stimuli typically undergo change to several properties at segment boundaries, while it is typically only the envelope of simpler synthetic stimuli that is manipulated?

5. Would the examination of a synthetic continuum that varies from a spectrally simple constitution to one of synthetic speech while varying parameters such as segment durations and envelope independently provide insight into why the P-centre typically shifts differently for speech and non-speech?

6. Significantly more data is required for compound and complex events. This includes not only chords with asynchronies, but also synchronous speech and singing for between two and many performers (e.g. a crowd), synthetic compounds of the type investigated by Seton (1989) and sounds with multiple rapid articulations (e.g. drum flams).

7. How does loudness affect the P-centre? Only Vos and Rasch (1981) and Seton (1989) appear to have conducted any empirical study relating the P-centre to the stimulus loudness. Natural, expressive performance of speech and music exhibits continuously changing loudness and it would seem that much more empirical data is required before a reasonable attempt at modelling any loudness dependence could be attempted.

8. How should investigation of the P-centre phenomenon be extended to continuous event streams (e.g. ensemble music and continuous speech) which would seem to be more typical of everyday experience? Early work on disyllables in speech may be a useful start point and more formal replication of these results, particularly in perception, is certainly warranted.

9. Related to question 8 (and perhaps question 6), over what period or duration do sound changes or manipulations affect the P-centre? Currently, researchers seem to be divided between those favour local onset-only effects and those who integrate features of the entire (isolated) sound signal. While the former may be compatible with psychoacoustics the latter is not. It seems then that an intermediate duration, a P-centre integration window (possibly the same as existing temporal integration windows in hearing) may exist, but the duration and weighting of this window remains to be investigated.

10. Seton's reaction time experiment (1989) seemed to show that participants reacted to the perceptual onset which occurred before the P-centre for simple ramped tones. Fowler (1983), using a different approach found that the reaction time for identifying a vowel was correlated with the synchronous tap times of participants. More investigation of the reaction time seems warranted. In particular, if no identification or recognition task is

invoked, does reaction time naturally relate to the (perceptual) onset of sounds or to their (possibly different) P-centres?

11. Finally, is there anything more that we can discover about the nature of P-centre perception. There is essentially no work examining the neurophysiology of the P-centre, though recent research has started looking at the neurophysiology of rhythm and meter (e.g. Snyder & Large 2005; Zanto, Snyder & Large 2006).

In the list above, questions 1–7 focus on progressing research into phenomena which may be useful for modelling the P-centre of isolated events. Other than the time investment required, there should not be any significant obstacles to answering these questions. Questions 8 and 9 are focused on the extension of P-centre research to continuous event streams. This extension is extremely important (since many natural event streams are continuous) but represents a significant departure from P-centre research to date and is likely to be difficult. Finally questions 10 and 11 relate to the nature of the P-centre and represent more exploratory research whose value remains to be seen.

### *2.3.3 Conclusions*

Although the size of P-centre effects observed for speech and non-speech undoubtedly seem different, it is not clear that these differences are due to anything other than complexity or amount of change typically resulting from speech manipulations compared to those of non-speech. Therefore a unified approach to investigating speech and non-speech stimuli seems to be warranted.

Furthermore a general review of the P-centre research indicates that while there is broad agreement about a strong initial consonant (or onset) effect and a weaker rhyme (or duration and offset) effect on the P-centre there still remain many details to be resolved—details which may well determine whether a P-centre model works reliably for all sounds, or is limited to a set

of sounds whose properties are rather like those already investigated (e.g. isolated stressed English CV and CVC monosyllables).

# Chapter 3

# Measuring P-centres

In all existing P-centre measurement methods participants must either consciously classify the temporal pattern of a set of events (as synchronous/asynchronous or isochronous/anisochronous for example) or synchronize actions with those events. From these responses, the intervals between the P-centres of successive (or even simultaneous) events can be inferred. If, for example, a participant perceives a perfectly isochronous rhythm, then the intervals between consecutive P-centres must be equal (except for some perceptual tolerance of deviations). Similarly, if a participant perceives two events as synchronous, then the interval between their P-centres must be close to zero, that is, their P-centres must be synchronous. There is one limitation common to all methods, however. Without knowing the absolute location of at least one P-centre in the pattern beforehand, the relative locations implied by the intervals between P-centres cannot be used to derive absolute P-centre locations.

P-centres mark specific moments in time and must be defined with respect to a time origin for their values to have meaning. Several P-centre variants can be distinguished based on the time origin used: the *absolute P-centre* (or simply P-centre), the *event-local P-centre* (EPC) and the *relative P-centre* (RPC). Figure 3.1 illustrates these variants.

The absolute P-centre is defined relative to a time origin that is common to the set of events under consideration, such as the objective beginning of a continuous acoustic stimulus. This is the form required to describe a

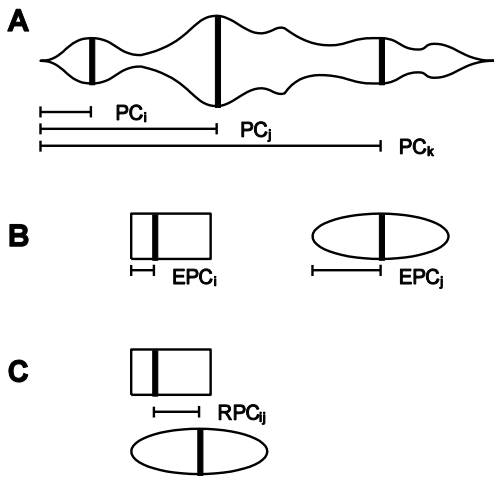


Figure 3.1 Three different P-centre measurements. (A) A continuous stimulus with absolute P-centres measured relative to the stimulus origin; (B) discrete events with event-local P-centres (EPCs) measured relative to each event's onset; and (C) the relative P-centre (RPC) of two discrete events. Hypothetical P-centre locations are marked by vertical lines.

pattern of P-centres when it is either impossible or inconvenient to explicitly segment the continuous stimulus into individual events with well defined boundaries, or if there was reason to suspect that the P-centres of individual events were highly context dependent. The absolute P-centre is the most general and useful form, allowing the temporal pattern of an arbitrary set of events to be measured or controlled. Unfortunately, there is no known method of directly detecting the perception of the P-centre at the moment it occurs, and consequently, no way to directly measure absolute P-centres.

Morton et al. (1976) hypothesized that the P-centre of a sound was independent of context, such as temporally nearby sounds, and that its temporal location relative to the sound thus remained constant. This is termed the *context independence hypothesis*.

Based on the assumption of context independence, it is useful to define the event-local P-centre: the P-centre of an event relative to an event-local origin, which is normally the physical onset or start of the event. The EPC can be related to the absolute P-centre by the difference

$$EPC = PC - EO \tag{3.1}$$

where *PC* is the absolute P-centre and *EO* is the event local origin. Finally, the relative P-centre expresses the relationship between the P-centres of two events *i* and *j*. It is defined by the difference

$$RPC_{ij} = EPC_i - EPC_j \tag{3.2}$$

where $RPC_{ij}$ should be read as the relative P-centre of event *i* with respect to event *j*. If $RPC_{ij}$ is positive, then the P-centre of event *i* occurs further from its onset than the P-centre of event *j*, i.e. the P-centre of event *i* occurs relatively later than the P-centre of event *j* when the two event onsets are synchronous. Conversely, if $RPC_{ij}$ is negative, then the P-centre of event *i* occurs relatively earlier than that of event *j*. An RPC of zero implies that the P-centres of both events are equally far from their onsets.

The true EPC cannot be measured if the absolute P-centre timing is unknown (see equation 3.1). Conversely, if it was possible to measure the true EPC, then the absolute P-centre could be derived from it (at least for isolated events). Nevertheless, both the RPC and the biased EPC[12] can be measured. However, it is worth noting that the definitions of EPC and RPC do not work for continuous event streams without well defined event boundaries.

The context independence hypothesis predicts two properties of the RPC that the various measurement methods use. First, if the roles of the two sounds in the RPC are swapped, then the RPC will simply change sign, that is, $RPC_{ij} = -RPC_{ji}$, since $EPC_i$ and $EPC_j$ should be invariant under the change of role. Second, the RPC of any two sounds may be calculated by simple addition if the RPC of each of those sounds relative to a common third

[12] Existing measurement methods incorporate an unknown delay, assumed to affect all events equally, when estimating the EPC; this is the source of bias. The tap asynchrony discussion explores this in more detail.

sound is known. Specifically, the *indirect* RPC of sound *i* relative to sound *k* is the sum of the *direct* RPCs of sounds *i* relative to *j* and *j* relative to *k*:

$$RPC_{ik} = RPC_{ij} + RPC_{jk} \quad (3.3)$$

Assuming there is no consistent bias, only the variance of the indirect RPC will be affected by the sum. A useful consequence of this additive property is that the RPCs of stimulus sets used in different experiments require just one sound in common to be directly comparable if context independence holds.

Biased EPCs for individual sounds can be inferred if a common reference sound is included in the stimulus set. There are several properties the common reference sound should have: It should be of short duration so that it will not overlap the previous or following event in a perceptually isochronous sequence; it should have a subjectively clear P-centre (as sounds with relatively abrupt onsets tend to have); it should not easily induce auditory streaming effects when alternating with other stimuli; and it should minimize RPC estimate variability. Auditory streaming (Bregman 1990/1999), where the single acoustic sequence is perceived as multiple perceptual streams whose temporal coordination is unclear, makes it difficult to tell whether the sounds in the sequence have the required rhythm. Marcus (1981) reported two conditions affected by streaming: when a single syllabic sound is repeated different components of the syllable may stream apart; and when one of the sounds in an alternating pair is a click, the sequence may be perceived as two streams. Streaming will tend to increase variability of the RPC estimates as will a reference sound that has an ambiguous P-centre. Using very short click-like reference sounds, Wright (2008) showed that estimate variability was reduced when the reference sound spectrum was modified to approximate the average spectrum of the test sound, but this approach, which entails synthesizing a custom click sound for each test sound, may not be practical in general. Nevertheless a good reference sound is likely to be short, with a relatively

abrupt onset, and a spectrum that is at least somewhat similar to the sounds under test.

## 3.1 Existing measurement methods

### *3.1.1 Rhythm adjustment method*

First described in detail by Marcus (1981), rhythm adjustment is by far the most commonly used method for measuring P-centres (see for example Cooper, Whalen & Fowler 1986; Harsin 1997; Pompino-Marschall 1989; Scott 1998). In this method, sequences are constructed by cyclic repetition of a short rhythm using just two sounds, the base sound and test sound.

Figure 3.2 shows key features of the experimental procedure. Initially, the repeating pattern is not isochronous. The participant's task is to adjust the timing of the test sound within the cycle until the *point of subjective isochrony* (where consecutive P-centre to P-centre intervals are equal) is reached. Each final adjustment yields one estimate of the RPC of the test sound with respect to the base sound.

Typically a duple rhythm (base-test-base-test…) is used and the base-base interval is fixed whereas the base-test interval is adjustable. Until perceptual isochrony is reached, the perceptual base-test interval is not equal to the perceptual test-base interval and neither interval is equal to the target isochronous interval (the base-base interval divided by 2). Harsin (1997) used a different rhythmic grouping, a triple rhythm (base-base-test-base-base-test…) where the first and second instances of the base sound were fixed, at the start and 1/3 of the cycle duration respectively, while the base-test interval was adjustable as before. Using this scheme, the target isochronous interval is presented once each cycle (between the first and second base sounds) and may potentially be used as a reference interval by participants. There is a possibility that rhythmic grouping (subjective or objective) may bias P-centre measurements if, for example, it leads to

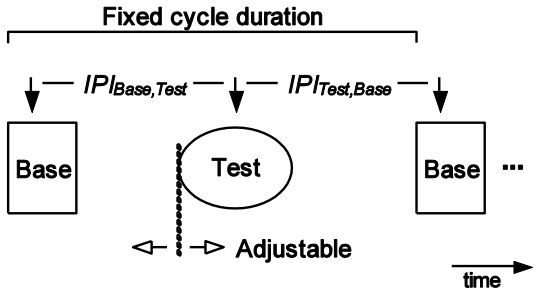


Figure 3.2 A schematic illustration of the rhythm adjustment method. The sequence consists of cyclic repetition of two sounds, the base and test. A participant adjusts the onset timing of the test sound within the cycle until the point of subjective isochrony is reached. At that point the inter-P-centre interval between the base and test sounds ($IPI_{Base,Test}$) will approximate that between the test and subsequent base sound ($IPI_{Test,Base}$). Downward pointing arrows indicate hypothetical P-centre locations.

subjective expectation of intervals that deviate systematically from isochrony. This question has not yet been investigated, however.

It is customary to measure the RPC for both possible assignments of sounds to roles, that is, $sound_i$ and $sound_j$ assigned first to *base* and *test* respectively and subsequently to *test* and *base* respectively. The resulting measures are averaged (assuming context independence) so that $\overline{RPC}_{ij} = (RPC_{ij} - RPC_{ji})/2$.

For a set of $N$ sounds, measuring only linearly independent RPCs (those which cannot be derived from any combination of the others) requires the fewest experimental conditions. For example, designating one sound as a common reference, and the other $N$ - 1 sounds as test sounds the RPC for each test sound can be directly measured relative to the reference. The indirect RPC for any pair not directly compared can be calculated as described previously. This approach is sensitive to the choice of reference sound since all other sounds are compared directly with the reference sound only. A poor choice of reference sound may result in larger estimate variance overall.

A more complex approach is to test all possible $N \times (N - 1)$ pairs of different sounds. The resulting RPC measurements are not all linearly independent;

at least some of the measured RPCs can be derived from a combination of others. Therefore, multiple linear regression is used to solve for EPCs with the exception of an unknown constant. RPCs may then be calculated between any two sounds in the set using the estimated regression parameters. Although this approach tends to balance the errors across all sounds in the set, larger overall variability can be expected if a number of sounds in the set have relatively unclear[13] P-centres (under the assumption that a participant will find it more difficult to detect anisochrony when both P-centres are unclear than when just one of them is unclear, hence making more variable adjustments in the former case.)

The main benefits of the rhythm adjustment method are that it is straightforward for participants to understand and can be implemented without special apparatus. For example, the method is not particularly sensitive to input delays when processing a participant's responses (in contrast to the synchronized tapping methods described later). Unfortunately, participants can find the task rather difficult and fatiguing to perform reliably since they must continuously judge whether or not the rhythm is isochronous. Judgment of isochrony seems to be even more difficult when one or both P-centres are unclear.

A variant of the rhythm adjustment method involves adjusting the test sound to the *point of subjective synchrony* (cf. Figure 1.1 [D]) rather than the point of subjective isochrony with the base sound (Gordon 1987; Wright 2008). The difference between the base and test sound onset times after adjustment is an estimate of the RPC. Multimodal distributions of RPC observations (perhaps implying competing candidate P-centres) have been found using this method, but it is possible that these distributions are simply artefacts of the method itself. Potential problems with the method include auditory masking (the onset of one sound may mask portions of the

[13] Although this thesis is focused on acoustic events with relatively well defined P-centres, it is still the case that the P-centres of some sounds are subjectively clearer and more precise than others. This point is explored in more detail in the discussion section of this chapter.

onset of the other), stimulus fusion (the two sounds may fuse into a single composite sound), and timbre changes at short onset delays (interference patterns occur if, as a control condition, the base and test sounds are identical). Although the synchrony adjustment task would superficially seem to be closely related to ensemble music performance, there may be other mechanisms involved in achieving synchronous musical performance (see for example Goebl & Palmer 2009).

### *3.1.2 Tap asynchrony method*

Tapping in synchrony with a regular rhythmic sequence is a simple task that many people perform naturally when listening to music. The motor actions associated with a tap take a certain amount of time to execute, so the person tapping must predict when the stimulus will next occur (based on an established rhythm) and begin the movement early so that the tap and the stimulus are perceived to be synchronous. However, when presented with a pacing sequence of short, abrupt sounds (such as the clicks of a metronome) it is commonly found that a participant's taps precede the sounds by some tens of milliseconds on average, a phenomenon referred to as *negative mean asynchrony* (Aschersleben 2002; Repp 2005). Furthermore, the negative mean asynchrony has been shown to depend on tapping force (Aschersleben, Gehrke & Prinz 2004), with more forceful taps exhibiting less negative asynchrony. Participants are generally unaware of any asynchrony; the sounds and taps appear subjectively synchronous. Although several different explanations have been proposed for negative mean asynchrony, these explanations are not of concern here. In the context of this work it is only necessary to consider that negative mean asynchrony adds an unknown constant (bias) to the EPC. Furthermore, the negative mean asynchrony is quite variable both within and between individuals.

Vos, Mates, and Van Kruysbergen (1995) showed that the asynchrony varied systematically when either the duration or rise time of acoustic stimuli were varied, and concluded that participants synchronize with the

P-centre rather than the (perceived) onset of sounds. Synchronization with the P-centre was predicted by Morton et al. (1976). Specifically, they predicted that it is the P-centre of a tap that is synchronized with the P-centre of a sound.

The *tap asynchrony* method for P-centre measurement begins with a pacing sequence, consisting of repeated presentations of the test sound at fixed isochronous intervals. The participant's task is simply to tap synchronously with each presentation of the test sound. The P-centre of a tap (the moment at which the participant perceives that the tap occurs) is assumed to occur at some unknown, but constant, offset from the moment of initial physical contact. (Although this offset may be influenced by factors such as the tapping force and tap duration, it seems reasonable to assume that the net effect is a constant offset when averaged across many taps.) Therefore, the mean tap asynchrony (relative to the sound onset) is taken to be an estimate of the EPC except for some unknown bias, that is, $\overline{A}_i = EPC_i - b$, where $\overline{A}_i$ is the mean tap asynchrony to sound *i* and *b* is the anticipation bias (relative to the sound's P-centre). The average anticipation bias is assumed to be invariant within an individual participant (at least in the context of an experiment). This assumption allows the RPC to be easily calculated from the difference in mean tap asynchronies for any pair of sounds, that is $RPC_{ij} = \overline{A}_i - \overline{A}_j = EPC_i - EPC_j$, because the bias is cancelled out. If the assumption of anticipation bias invariance within a participant is violated, the resulting RPC estimates will not be reliable.

Only Janker (1996a) appears to have used the *tap asynchrony* method as described for general P-centre measurement. Allen (1972a) also made use of a synchronized tapping task (to identify syllable beats in his experiments) but the details of his procedure differ from those described and thus the procedures are not likely to be comparable.

The synchronized tapping task used in the tap asynchrony method is performed automatically by participants and does not require them to make

conscious decisions. For this reason participants generally seem to find the task easier than rhythm adjustment. Furthermore, this method allows initial RPC estimates to be measured quickly, though the variability of asynchrony may require more observations to obtain sufficiently reliable estimates. An important difference between the tap asynchrony and rhythm adjustment methods is that the former must derive RPCs from the estimated EPCs for individual sounds assuming constant bias between trials and P-centre context independence, whereas the latter explicitly measures RPCs for pairs of sounds heard together in a sequence.

### *3.1.3 Other methods*

There are two remaining previously used methods that are worth discussing briefly. The first of these is a two alternative forced choice task using the method of constant stimuli (Fox & Lehiste 1987b). In this method, a sequence of 4 sounds (base-base-base-test) is presented. The inter-onset interval is fixed between the base sounds and manipulated between the last base sound and test sound. Participants are forced to choose whether the test sound is presented *too early* or *too late*. With a sufficient number of results, psychometric functions can be constructed and the RPC of the test sound with respect to the base estimated from the point of subjective equality on the psychometric function. The method is easy to implement and can be readily executed with multiple simultaneous participants (resulting in a useful time efficiency). Nevertheless, the method, though it seems to be little more than a straightforward constant stimulus variation of rhythm adjustment, suffers from some problems. Fox and Lehiste noted that listeners tend to underestimate the duration of the last interval in the sequence, a behavior which may distort RPC measurements (see also Benguerel & D'Arcy 1986; Repp 1995). Additionally, the task depends on judging the temporal order of a perceived event and an internally timed moment of isochrony. Participants in a rhythm adjustment experiment frequently find it easier to detect anisochrony than to choose the direction of adjustment required to reduce it, suggesting that temporal order

judgments are more difficult than anisochrony judgments. Until these problems can be resolved, this method does not seem suitable for P-centre measurement.

In the final method, speech production, participants are required to produce specific speech tokens, usually monosyllables, in either a rhythmic framing sentence or a simple repeating sequence paced with or without the aid of a metronome (for examples see Fowler 1979; Fox & Lehiste 1987b; Perez 1997; Rapp-Holmgren 1971; Tuller & Fowler 1980). This method is generally used to discover relationships between acoustic or articulatory features and the P-centre of a syllabic sound and not to estimate RPCs. Due to the complex nature of the motor task involved in speech production, the variability between repeated productions of the same token, and the limitation to speech sounds only, the speech production method is not a suitable candidate for a general P-centre measurement method.

## 3.2 The PCR Method

Research on sensorimotor synchronization, in particular finger tapping in synchrony with an auditory sequence, has investigated the phase correction process that enables a person to stay in synchrony with a pacing sequence that may incorporate phase perturbations. A key feature of this process is the *phase correction response* (PCR), which denotes the phase shift of a tap in response to a phase-shifted event in an otherwise isochronous pacing sequence (Repp 2002, 2005). The PCR occurs involuntarily and generally without a participant's awareness. Two kinds of phase perturbation are commonly employed: a phase shift, which affects the test event and all subsequent events, and an *event onset shift* (EOS), which affects only the test event[14]. The PCRs elicited are equivalent because a phase shift, by definition, begins with an EOS. A schematic illustration of an EOS and the subsequent PCR is provided in Figure 3.3.

[14] The EOS and phase shift have also been respectively described as single event displacement and single interval lengthening/shortening (Friberg & Sundberg 1995).

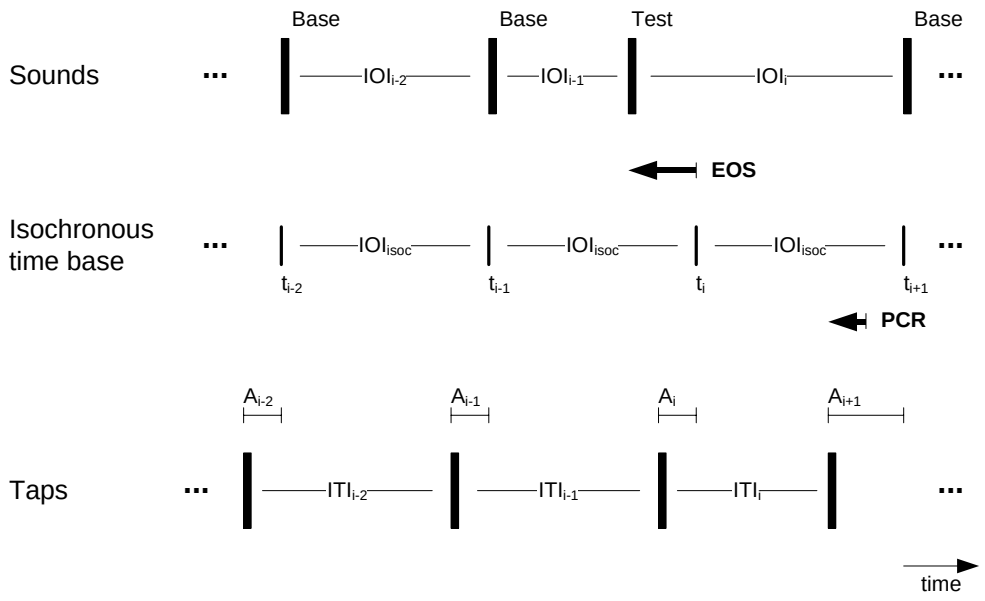


Figure 3.3 Schematic illustration of an event onset shift (EOS) and the phase correction response (PCR). In a pacing sequence of sounds, base events are timed according to the isochronous time base whereas the test sound represents a displacement from that timing. Taps generally anticipate sounds and the asynchrony between the tap and sound onset is denoted *A*. The EOS displaces just one event and the PCR appears on the subsequent tap. The PCR may be measured by subtracting the unperturbed inter-onset interval (IOI) from the current inter-tap interval (ITI), or equivalently as the difference between the tap asynchronies at and immediately after the EOS, i.e. $PCR = ITI_i - IOI_{i-2} = A_{i+1} - A_i$ .

The PCR may be calculated in two equivalent ways for a sequence of the form shown in Figure 3.3. In general, the current inter-tap interval (ITI) depends on the previous inter-onset interval[15] (IOI). For a test event onset shifted at time index *i* the immediately following ITI is affected. Therefore the first method of estimating the PCR uses the difference between the current ITI, $ITI_i$, and the pre-perturbation IOI, $IOI_{i-2}$ :

$$PCR = ITI_i + IOI_{i-2} \tag{3.4}$$

Referring again to Figure 3.3 it can be seen that the ITI can be related to the isochronous time base instants, *t*, and the tap asynchronies, *A*, as

[15] In fact this is simplification which assumes that the P-centre and onset of the sounds approximately coincide. More correctly, the inter-tap interval depends on the previous inter-P-centre interval.

$ITI_i = (t_{i+1} + A_{i+1}) - (t_i + A_i)$. Furthermore, the difference between consecutive time base instants is simply the isochronous IOI, i.e. $t_{i+1} = t_i + IOI_{isoc}$. Substituting these two expressions into equation 3.4 yields the alternative PCR calculation in terms of the difference between tap asynchrony at and immediately after the EOS perturbation:

$$PCR = A_{i+1} - A_i \tag{3.5}$$

As long as phase perturbations are within about ±15% of the sequence baseline IOI, the PCR can be well described by a linear model, termed the *PCR function*, (Repp 2002). In the linear range, each tap corrects for some fraction, $\alpha$, of the preceding tap-sound asynchrony. This parameter, $\alpha$, can be estimated mathematically from the complete time series of tap asynchronies with an isochronous sequence (Schulze & Vorberg 2002) or, alternatively, from the PCRs which immediately follow phase perturbations introduced into an otherwise isochronous sequence (Repp 2002). To apply this latter technique to estimate $\alpha$, the perturbation magnitude is varied within the range that elicits a linear PCR, and the resulting PCRs are regressed onto perturbation magnitude. The sloped of the regression line is the desired estimate of $\alpha$, as illustrated in Figure 3.4.

The discussion of PCR measurement and $\alpha$ estimation to this point was concerned primarily with their established application to the study of sensorimotor synchronisation. The PCR phenomenon and PCR measurement techniques have not previously been applied to P-centre estimation, however, and it is this novel application that is hereinafter termed *the PCR method*.

To apply the PCR to P-centre measurement, participants are asked to tap in synchrony with a pacing sequence in which the onset-shifted events are termed the test events and the other events are termed the base events. As illustrated in Figure 3.3 a base sound is presented repeatedly at

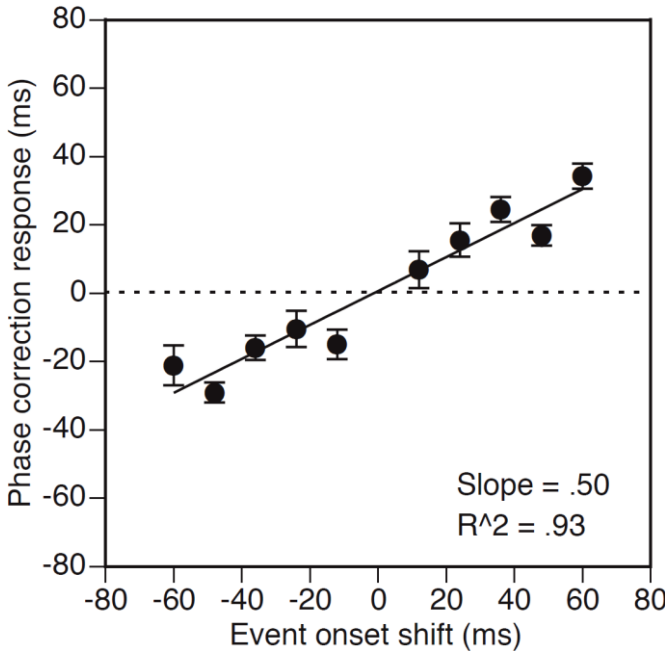


Figure 3.4 Illustration of the calculation of the phase correction coefficient α as the slope of a regression line relating the PCR to EOS magnitude. Each data point is the mean of a number of observations, with standard error bars. The value of $R^2$ (R^2) indicates the goodness of the linear fit (very good in this example). The baseline IOI was 600 ms in this example.

isochronous intervals while a test sound is inserted occasionally with various EOS values and PCRs are measured in response to each test event.

When a sequence of events is isochronous there cannot be an expected (mean) PCR since there is no phase perturbation requiring a correction. Therefore, when the PCR function is estimated from a range of EOS perturbations, the point at which the estimated function is zero, the EOS axis intercept, indicates the point of subjective isochrony relative to the unperturbed base sounds in the pacing sequence. It is this observation which enables the P-centre to be estimated.

If the sequence events are all homogeneous, as they typically are in research on the PCR, their P-centres will be identical. For symmetrically distributed EOS values, the PCR function should pass through the origin (see Figure 3.4) although random variability and systematic phase drift can cause small deviations of the regression line's EOS axis intercept from zero.

(It should be clear that because the EOS axis intercept will always be approximately zero for homogeneous base and test events such sequences cannot be used to estimate P-centres.) If, instead, the onset-shifted test event is different from the preceding events and has a different EPC, its point of subjective isochrony (eliciting a zero PCR) will occur at some EOS value other than zero. If, for example, the P-centre of the test event is 20 ms later than that of the preceding sounds then the expected PCR would be positive at the point of onset isochrony ($EOS = 0$). Correspondingly, the EOS axis intercept ($PCR = 0$) would occur at an EOS of -20 ms. In other words, the test event has to occur 20 ms earlier than the point of onset isochrony to be perceptually isochronous, in which case no PCR is elicited.

Since each PCR function is a line, $PCR = b_0 + b_1 x$, defined by the regression constant, $b_0$, and slope, $b_1$, the x-axis intercept ($PCR = 0$) may be calculated as $x_{\text{Intercept}} = -b_0 / b_1$. This intercept value defines the onset anisochrony required to place the test event at the point of subjective isochrony relative to the base events. To estimate the RPC of the test event relative to the base event, the intercept value is simply negated. Like rhythm adjustment, the PCR method is used to measure the RPC values of mixed sound pairs and RPC estimates can be obtained for both possible role-to-sound assignments. (Just like rhythm adjustment, the RPC estimates for these role-to-sound assignments should differ in sign but be approximately equal in absolute value.)

If, in addition to measuring the minimum asynchronies or inter-tap intervals necessary for PCR estimation, the tap asynchrony to all base events is measured, then a second method of estimating P-centres may also be used. This second estimation method is in fact almost identical to the tap asynchrony method except that instead of averaging the asynchrony for all events, only a subset of the base events are included. By its nature, the PCR disturbs the mean tap asynchrony for the taps immediately following a perturbation, but the effect subsides over subsequent taps. Although including asynchronies in the vicinity of a perturbation tends to increase

the variance of the asynchrony estimate, even when the perturbations are balanced around zero, the effect of the PCR is minimal after 3 or 4 taps (see Repp 2005). If too many taps are excluded, the benefits of reduced variance may be outweighed by variability due to the smaller sample size from which to estimate the mean asynchrony. Thus a reasonable compromise is to exclude the first few base event asynchronies after each phase perturbation (at least 3) from the averaging and P-centre estimation.

The PCR method has several beneficial properties while suffering from just one notable drawback. First, like tap asynchrony and in contrast to rhythm adjustment, it enables P-centre measurement without requiring explicit perceptual judgements from participants—P-centre estimates are by-products of an automatically performed synchronisation task. Second, unlike tap asynchrony, difference measures (the PCR) rather than absolute measures are used to estimate the P-centre which might make the method more accurate or less susceptible to bias. Third, the method actually yields two somewhat independent measures (the PCR, and tap asynchrony) that can be used to estimate P-centres. The main drawback of the PCR method arises from the limited range of EOS values over which the PCR is approximately linear. EOS values must be constrained and centred approximately on the point of subjective isochrony to remain within the linear range. This means that a prior estimate of the P-centre difference between two sounds must be obtained with some other method to guide the relative timing of the sounds in the pacing sequence. Therefore, the PCR method thus is not very useful for initial P-centre measurements. It is, instead, more appropriate for confirming and perhaps fine-tuning existing P-centre estimates. The precision and reliability of the PCR method (and both its P-centre estimation methods) is an empirical question the present study was intended to address.

## 3.3 The present study

There were several objectives to be addressed by the present study which was conducted as part of an international collaboration with Bruno Repp of Haskins Laboratories[16]. The primary objective was to determine which methods might be best for assessing P-centres consistently and efficiently so that those methods could be recommended for future investigations. Two methods without problems that would prevent general P-centre measurement (as noted in the review) were selected: rhythm adjustment and simple tap asynchrony. Rhythm adjustment has been the most commonly used method and would serve as the control against which other methods would be compared. The new PCR method would also be evaluated.

The accuracy and reliability of each method was assessed using the variability of RPC estimates it produced, both within and between participants. The most fundamental question to be addressed was: Do these methods all measure the same percept, the P-centre? The agreement of RPC estimates between the methods was assessed by using each method to obtain estimates for the same set of stimuli. These stimuli were seven speech syllables that pilot experiments suggested had a wide range of P-centres, and a non-speech reference sound comprising a harmonic complex and noise mixture. (Except insofar as they allowed comparison of the measurement methods, the specific P-centre values were of no particular concern in this study and there was no independent variable whose level was controlled between sounds other than sound identity.) The RPC of each syllable with respect to the reference sound provided a minimum set of estimates that allowed all syllable P-centres to be compared within and between methods. Various syllable-syllable pairings were also examined, though not all combinations.

[16] Based on a collaborative design, Experiment 3 was conducted by Bruno Repp at Haskins Laboratories and this author was not present while it was running. All analyses presented in the thesis were conducted by the author.

All the measurement methods being tested rely on the assumption of P-centre context independence to produce reliable RPC estimates. Although Marcus (1981) tested this hypothesis for rhythm adjustment, it has been reexamined just once and then with just one participant (Eling, Marshall & van Galen 1980). The context independence hypothesis was tested in two ways. First, direct RPC estimates were obtained for various syllable-syllable pairs and compared to indirect RPCs calculated by addition of the results for appropriate noise-syllable pairs. Second, RPC estimates were obtained for pairs of sounds in both orders (i.e., with their roles interchanged), because independence predicts a negative relationship between the RPCs. Based on the findings of Marcus and Eling et al., it was expected that context independence would be supported by the rhythm adjustment method, but whether or not it was supported by the PCR method was an empirical question to be answered. If there was any context dependence due to order its effect should be greater with the PCR method, since there is a greater difference between the presented event sequences in the two orders using this method (due to repetition of the base sound). Unfortunately, context independence must be assumed and cannot be evaluated for the tap asynchrony method, since this method can only estimate EPCs (and thereafter derive RPCs assuming independence) rather than measuring RPCs directly.

The PCR method was also applied to homogeneous sound sequences typical of general PCR investigation. Pilot observations had suggested that the slope of the PCR function might be steeper in homogeneous than in heterogeneous sequences. If confirmed, this novel finding would suggest that phase correction is less effective in the presence of sound change. Furthermore, homogeneous sequences were expected to yield a better estimate of mean asynchrony for each sound, as well as additional information about the accuracy and reliability of the PCR method because the EOS axis intercept was expected to be at zero. Although there were no specific predictions regarding differences in slope among heterogeneous sequences, the experiments examined this issue as well.

The study's final aim was to assess the efficiency of each method in terms of its accuracy (standard error of RPC estimates) relative to its execution time (for each participant and number of participants required). P-centre measurement methods are often rather time consuming to execute and the objective was to discover which method provided the optimum return on time invested.

## 3.4 Experiment I

The aim of Experiment I was to measure RPCs by rhythm adjustment, the most commonly used method. RPCs measured using this method would serve as a baseline or control against which measures from other methods could be compared. Sound pairs that could be used to directly estimate RPCs were augmented by additional pairs that could be used to derive equivalent indirect RPC estimates. The P-centre independence hypothesis predicts that direct and indirect RPC estimates should not differ significantly. If confirmed, this would support the findings of Marcus (1981) and justify the continued use of the rhythm adjustment method, which fundamentally depends on the assumption of context independence to generate sensible RPC estimates. Finally, pilot experiments suggested that some sound pairs were harder to align than others. It was predicted that trial duration, which is a coarse indicator of difficulty in an adjustment task, would show an effect of sound pair if there were any pairs that were systematically more difficult than others.

### *3.4.1 Method*

#### *3.4.1.1 Participants*

The participants were 2 females and 6 males (21–45 years old) comprised of 7 unpaid volunteers at the National University of Ireland Maynooth and the author. Three participants had previously performed the rhythm

adjustment tasks, but only the author was practiced. None of the participants had any known hearing deficiencies. All were native speakers of English and had a range of music training (0–17 years).

#### *3.4.1.2 Stimuli*

The stimuli were seven naturally produced monosyllables and a synthetic reference sound. The relationship between specific acoustic features of these sounds and their P-centres was not the concern here. The syllables /ba/, /la/, /pa/, /pla/, /sa/, /spa/, and /spla/ were produced by a female native speaker of English and digitally recorded. After trimming leading and trailing silence, the recordings ranged in duration from 420–560 ms. Individual phoneme productions were not edited, so the recordings exhibited some natural variation in those productions. For example, the /l/ in /la/ differed acoustically from that in /pla/.

The reference sound was designed not only for the present study but for anticipated use as a generally applicable reference sound that could be used in a variety of P-centre experiments. For this reason, the reference sound was a synthetic, 200 ms, 1:1 mixture of noise and a harmonic complex. The harmonic complex had a 100 Hz fundamental frequency and phases designed to reduce the crest factor (Schroeder 1970). Both the harmonic complex and the noise had a pink (1/f) spectrum which was intended to be relatively similar to the long term spectral average of speech (and many natural sounds). The amplitude envelope (a cosine shaped 20 ms onset and 180 ms offset) was designed to elicit a relatively early P-centre so that test sounds would be likely to have relatively later EPCs and, hence, RPCs using the noise as a reference would tend to be positive. Together, the combination of harmonic and noise components, spectral profile, and envelope were expected to mitigate the effects of streaming, and pilot experiments suggested this was the case. Most participants described the timbre of this reference sound as noise-like and thus it was referred to simply as noise. For convenience, the 7 syllables and reference sound are

hereinafter referred to as: BA, LA, PA, PLA, SA, SPA, SPLA and N. (0 shows the waveform and spectrogram of all these sounds.)

Sounds were paired for measurement and formed two main groups: noise-syllable pairs and syllable-syllable pairs. Noise-syllable pairs consisted of each of the 7 syllables paired with the reference sound (N) in both orders (with N as the base sound and the syllable as test sound and vice versa). There were thus 14 unique permutations from which RPCs could be estimated. Syllable-syllable pairs consisted of two sub-groups in which all combinations of 3 syllables each were tested. These were LA-PLA, PLA-SPLA, LA-SPLA, and PA-SA, SA-SPA, and PA-SPA. Once again both orders of each pair were tested so that there were 12 permutations in all. Syllable-syllable pairs provided independent RPC estimates that could be compared to those measured for noise-syllable pairs to test the context independence hypothesis. Moreover, the RPC estimates for each triplet of syllable-syllable pairs should be internally consistent if RPCs are context independent.

#### *3.4.1.3 Apparatus*

Custom software, running under Windows XP on a personal computer, controlled the adjustment procedure (see Appendix B). Participants could adjust asynchrony over a ±400 ms range (permitting the sounds to overlap if so chosen) using the keyboard, mouse pointer, or mouse scroll wheel. There was no visible indication of the absolute adjusted asynchrony, and participants could make adjustments as small as 1 ms.

The timing of the output audio events was sample accurate. The digital audio for each sequence was mixed in real time at a sampling rate of 48 kHz, converted to analogue by an M-Audio USB Duo 2 audio interface, and presented diotically using Sennheiser HD280 Pro closed-back circumaural headphones in a quiet room.

#### *3.4.1.4 Procedure*

In each trial, a pair of sounds was used to construct a cyclic sequence having a mean inter-onset interval (IOI) of 650 ms and cycle duration of 1300 ms. The base sound was fixed to the start of each cycle, while the asynchrony of the test sound relative to the cycle mid-point was adjustable by the participant. At the start of each trial the initial asynchrony of the test sound was randomly selected from the discontinuous range -200 to -100 ms and 100 to 200 ms. (This choice of values had three desirable properties: The initial rhythm was generally not isochronous and thus required adjustment; participants were exposed to trials where the test sound initially occurred both too early and too late; and finally, the asynchrony was not so large that parts of the base and test sounds would overlap.) The trial began when the participant clicked an onscreen button. Their task was to adjust the asynchrony of test sound until the rhythm of the cyclic sequence was perceptually isochronous. Participants could stop and restart the sequence with a button press as necessary if, for example, they became confused about which sound was taking the base or test role. The most recent adjustment of the asynchrony was always used when the sequence was restarted. The participant clicked an onscreen button to end the trial. The software saved the initial asynchrony, time-stamped sequence of adjustments, and final adjusted asynchrony for each trial.

Trials were blocked, and each block consisted of trials for all 13 sound pairs in both orders (that is 26 trials in all). The order of trials was randomized in every block. Six blocks were presented in the course of 2 sessions taking approximately 45 minutes each. Sessions were typically a week apart.

### *3.4.2 Results*

Data for repetitions of each condition were first aggregated within participants. One participant appeared unable to perform the task adequately. This participant's adjustments exhibited much larger than

**Table 3.1 Direct RPC estimates obtained using the rhythm adjustment method.**

| | SD RPC | | RPC | | Pooled RPC | |
|---|---|---|---|---|---|---|
| Pair | Fwd | Rev | Fwd | Rev | *M* | *SE* |
| N-BA | 19.30 | 31.45 | 11.90 | -0.19 | 5.86 | 4.92 |
| N-LA | 24.50 | 34.41 | 39.26 | 38.64 | 38.95 | 6.35 |
| N-PA | 21.04 | 26.21 | 51.02 | 43.60 | 47.31 | 3.74 |
| N-PLA | 27.26 | 31.27 | 54.17 | 57.21 | 55.69 | 5.86 |
| N-SA | 22.03 | 24.46 | 109.21 | 110.71 | 109.96 | 3.68 |
| N-SPA | 31.40 | 36.29 | 181.98 | 189.79 | 185.88 | 8.51 |
| N-SPLA | 29.65 | 30.98 | 176.45 | 173.83 | 175.14 | 5.50 |
| LA-PLA | 13.97 | 13.43 | 16.33 | 9.95 | 13.14 | 0.99 |
| LA-SPLA | 17.39 | 19.02 | 138.57 | 133.81 | 136.19 | 3.02 |
| PLA-SPLA | 22.58 | 19.51 | 117.86 | 119.50 | 118.68 | 2.62 |
| PA-SA | 14.67 | 18.08 | 56.14 | 48.64 | 52.39 | 2.06 |
| PA-SPA | 22.24 | 16.88 | 128.05 | 128.38 | 128.21 | 2.48 |
| SA-SPA | 14.81 | 17.90 | 67.79 | 67.76 | 67.77 | 2.11 |

*Note*— Each pair ($sound_j$-$sound_i$) acted in the roles base-test in the forward order (Fwd) and test-base in the reverse order (Rev). All RPC values shown are for $sound_i$ relative to $sound_j$, thus reverse order ΔPCs, measured for $sound_j$ relative to $sound_i$, were negated. The SD RPC measure is the average within-participant standard deviation of the RPC. The Pooled RPC columns give the between-participant mean (*M*) and standard error (*SE*) of the RPC estimates pooled between orders. All values are in milliseconds.

average variability between replications of each condition. As other researchers have excluded participants judged unable to perform the task adequately on the basis of screening trials (Harsin 1997), this participant's data were excluded from the analysis. The main results, averaged across the remaining participants, are shown in Table 3.1.

The mean trial duration was 48.2 s (*SD* = 14.4 s). Trial duration can be interpreted as an indicator of task difficulty (though subject to confounding effects such as participant attention) and was subjected to a two way

repeated-measures ANOVA[17] with the independent variables of Pair (13 levels) and Order (2 levels). Neither the main effects nor their interaction were significant; therefore, it seems there were no individual conditions in which participants consistently experienced greater or lesser difficulty than average. Furthermore, though some participants reported having more difficulty with noise-syllable pairs than syllable-syllable pairs, the noise-syllable trial durations ($M$ = 49.8, $SD$ = 17.2) were not significantly longer than the syllable-syllable trial durations ($M$ = 46.3, $SD$ = 13.8), $t(6) = 0.73$, $p = 0.49$.

The within-participant standard deviation of the RPC estimate is expected to indicate both how reliably a participant can reproduce his or her own adjustments and how clear or ambiguous the RPC is for a particular sound pair. A two-way repeated-measures ANOVA indicated that the effect of Pair on the standard deviation of RPC was of medium size and significant, $F(12, 72) = 4.11$, $\varepsilon = .17$, $p = .04$, $\eta_G^2 = .19$. Neither the Order effect nor the Pair × Order interaction was significant, $F(1, 6) = 1.04$, $p = .35$, $\eta_G^2 = .01$, and $F(12, 72) = 1.03$, $\varepsilon = .18$, $p = .39$, $\eta_G^2 = .03$, respectively. Closer inspection of the differences among pairs revealed that standard deviation of RPC was higher for noise-syllable pairs ($M$ = 27.9, $SD$ = 13.2) than for syllable-syllable pairs ($M$ = 17.5, $SD$ = 4.2), and this effect was both large and significant, $t(6) = 2.61$, $p = .04$, $r = .73$.

From pilot experiments, it was expected that the RPC would differ significantly between pairs. However, it is a fundamental prediction of the P-centre context independence hypothesis that the sign-corrected within-order RPCs for a pair of sounds should not differ significantly. Table 3.1 shows that these matching RPC values differed by less than 10 ms in all

[17] The Greenhouse-Geisser correction was applied to all repeated-measures factors with more than two levels unless two conditions were met: Mauchly's test for sphericity was not significant and $\varepsilon > .8$. Where used, the correction factor $\varepsilon$ is reported so that departures of sphericity are clear. The effect size statistic generalized eta squared, $\eta_G^2$, is used to facilitate comparability across between-participant and within-participant designs (Bakeman 2005; Olejnik & Algina 2003).

**Table 3.2 Direct and indirect RPC estimates for syllable-syllable pairs obtained with the rhythm adjustment method.**

| | RPC | | RPC via N | | RPC via syllable | | |
|---|---|---|---|---|---|---|---|
| Pair | *M* | *SE* | *M* | *SE* | Syl. | *M* | *SE* |
| LA-PLA | 13.14 | 2.4 | 16.74 | 3.78 | SPLA | 17.51 | 2.88 |
| LA-SPLA | 136.19 | 2.48 | 136.19 | 3.96 | PLA | 131.82 | 3.17 |
| PLA-SPLA | 118.68 | 3.22 | 119.45 | 6.12 | LA | 123.05 | 3.28 |
| PA-SA | 52.39 | 1.17 | 62.65 | 4.17 | SPA | 60.44 | 3.51 |
| PA-SPA | 128.21 | 3.52 | 138.57 | 7.41 | SA | 120.17 | 3.19 |
| SA-SPA | 67.77 | 3.35 | 75.92 | 4.7 | PA | 75.82 | 4.06 |

*Note*—For each pair ($sound_j$-$sound_i$), the direct RPC of $sound_i$ relative to $sound_j$, $RPC_{ij}$, is reproduced from Table 3.1 for comparison. Indirect RPCs were calculated via a third sound, $k$, such that $RPC_{ij} = RPC_{ik} + RPC_{jk}$. The identity of sound $k$ was either the reference noise (RPC via N) or a syllable (Syl.). All values are in milliseconds.

cases (except for N-BA) which is quite a bit less than the typically reported jnd for the presentation rate used (Friberg & Sundberg 1995). A two-way repeated-measures ANOVA showed the expected large and significant Pair effect, $F(12, 72) = 282.77$, $\varepsilon = .29$, $p < .01$, $\eta_G^2 = .96$. The effect of Order and the Pair × Order interaction were both small and nonsignificant, $F(1, 6) = 0.51$, $p = .50$, $\eta_G^2 = .01$, and $F(12, 72) = 1.15$, $\varepsilon = .33$, $p = .36$, $\eta_G^2 = .05$, respectively.

Under the context independence hypothesis, RPCs may be measured directly between a pair of sounds, or calculated indirectly by simple addition of RPCs between each sound in the pair and a common third sound. All direct and indirect RPC estimates of syllable-syllable pairs resulting from the data are shown in Table 3.2. Pairwise comparisons of direct and indirect RPCs for each sound pair yielded just one comparison that approached significance: PA-SA direct compared to PA-SA via N, $t(6) = -2.40$, $p = .05$. With Bonferroni correction, none of the differences reached significance, so there was no evidence of P-centre context dependence.

The RPC estimates and the efficiency of the rhythm adjustment method are compared with the other methods in this study after the method-specific experiment sections.

## 3.5 Experiment II

The purpose of Experiment II was to estimate RPCs using the simple tap asynchrony method with homogeneous sound sequences constructed from the same set of sounds used in Experiment I. In this method it is (biased) EPCs that are measured and RPCs are subsequently derived using the assumption of P-centre context independence. Specifically, the assumption is that the EPC does not change whether the sound is presented among homogeneous or heterogeneous sounds. As tap asynchronies suffer from a number of potential sources of variability (individual anticipation bias, phase drift, and motor variability) there were two key questions to be addressed: Would asynchronies prove to be stable within participants and would RPC estimates agree with those of adjustment?

### *3.5.1 Method*

#### *3.5.1.1 Participants*

All the participants from Experiment I participated again in Experiment II. All but two were right handed.

#### *3.5.1.2 Stimuli*

The 8 sounds used in Experiment I were used again here. In this experiment sounds were not tested in pairs; instead each sound was tested individually.

#### *3.5.1.3 Apparatus*

The experimental procedure was controlled by custom software (see Appendix B), running under Windows XP on a personal computer. The audio presentation apparatus was identical to that of Experiment 1. Taps were registered on a custom touch sensor with 4 × 4 cm sensing area: A short strip of conductive tape affixed to the participant's index finger enabled the moment of tap contact and release to be detected with precision. The presented audio signal was routed through a simple circuit to generate a synchronized 2 channel signal containing the finger tap signal in one channel and the presented audio in the other. This 2 channel signal was routed to the line input of a Griffin Technology iMic audio interface, digitized at a sample rate of 11,025 Hz per channel, and recorded. Additional custom software processed the digital recording after the experiment to identify the timing of finger tap events in relation to the presented audio onsets (exceeding a threshold just above the signal noise floor) with an accuracy of better than 1 millisecond.

#### *3.5.1.4 Procedure*

Each trial consisted of a sequence constructed from a single sound repeated 40 times at a constant IOI of 700 ms[18]. Participants sat in front of the computer with the index finger of their dominant hand over the tapping device. They started the trial by pressing a key on the computer keyboard. Thereafter a short warning tone was played, followed by a brief pause and then the trial sequence. Participants were instructed to start tapping with the third sound in the sequence and to stay synchronized throughout (giving 39 expected taps per sequence, the last of which did not accompany a sound). They were further instructed not to count the sounds or try to

[18] This IOI, which was slightly larger than that of Experiment 1, reduced the occurrence of streaming effects with the more complex syllables. Prior research had indicated that presentation rate, at least within the narrow range of values used here, should not play a significant role in P-centre measurement (Eling, Marshall & van Galen 1980).

form rhythmic groups. Participants were free to cancel and restart the trial at any time if they noticed that they had skipped a tap, double tapped, or lost synchrony with the sequence. Only data from completed trials were saved.

For each participant, the order of trials was randomized within each block of 8 trials. Four blocks were presented, typically in a single session, taking approximately 20 minutes.

### *3.5.2 Results*

Participant taps were successfully matched to pacing sounds in all but 3 cases where no tap was present (indicating a skip), resulting in 9,725 usable taps registered. The mean and standard deviation of tap asynchrony, measured relative to the pacing sound onsets, were calculated separately for each trial. These statistics were then aggregated across sequence repetitions within participants and all subsequent hypothesis tests used the within-participant summary data only. The within-participant RPC estimate for each sound relative to the noise reference sound was calculated as the difference between their mean tap asynchronies. The main results averaged across participants are shown in Table 3.3.

There was a tendency for the within-participant standard deviation of asynchrony to be larger for more complex syllabic sounds with late RPCs, while the noise sound had the smallest standard deviation. A one way repeated measures ANOVA found that the pacing sound had a small but significant effect on the standard deviation of asynchrony, $F(7,49) = 4.43$, $\varepsilon = .44$, $p < .05$, $\eta_G^2 = .06$.

As expected, the asynchrony itself shows a large systematic variation with the pacing sound. The rather large standard errors reflect individual differences in the magnitude of anticipation bias.

**Table 3.3 Asynchronies and RPC estimates from the tap asynchrony method.**

| | SD Async. | | Asynchrony | | RPC | |
|---|---|---|---|---|---|---|
| Sound | *M* | *SE* | *M* | *SE* | *M* | *SE* |
| N | 23.97 | 2.17 | -32.37 | 10.94 | — | — |
| BA | 26.18 | 2.05 | -28.44 | 10.49 | 3.92 | 3.77 |
| LA | 28.91 | 2.77 | 12.64 | 13.13 | 45.01 | 6.61 |
| PA | 28.10 | 3.06 | 13.66 | 11.88 | 46.02 | 3.24 |
| PLA | 27.34 | 3.18 | 11.50 | 13.70 | 43.86 | 3.37 |
| SA | 27.11 | 2.96 | 69.04 | 12.22 | 101.41 | 3.66 |
| SPA | 30.20 | 3.19 | 125.65 | 17.24 | 158.02 | 6.81 |
| SPLA | 30.47 | 3.64 | 121.96 | 17.40 | 154.33 | 8.27 |

*Note*—SD Async. = average within-trial standard deviation of asynchrony. Each RPC is for the specified sound relative to the common reference sound N and is calculated from the difference between their asynchronies, that is, $RPC_{sound,N} = Async_{sound} - Async_{N}$. All values are in milliseconds.

Once again, the RPC estimates and the efficiency of the tap asynchrony method are compared with the other methods in this study after the method specific experiment sections.

## 3.6 Experiment III

The PCR method for measuring P-centres was assessed in Experiment III. Like tap asynchrony (Experiment II), the PCR method employs a synchronized tapping task, but uses phase perturbed rather than strictly isochronous sequences. Like rhythm adjustment (Experiment I), the sequences usually use two different sounds so that RPC estimates can be obtained directly. Two measures were used to estimate RPCs: the PCR in response to an EOS perturbation and the mean asynchrony of taps sufficiently far from perturbations to be largely unaffected by them.

There were several questions to be addressed by this experiment. Would PCR based RPC estimates agree with those of the adjustment method? Would mean asynchronies yield estimates that agreed with those derived from the PCR functions and those obtained with the simpler tap asynchrony method? Finally, what effect, if any, do the P-centre and type of sequence have on the slope of the PCR function?

### *3.6.1 Method*

#### *3.6.1.1 Participants*

There were 9 participants. Bruno Repp (who also ran the experiment) was 63 years old at the time, has been an active amateur pianist all his life, and is highly experienced in synchronization tasks. The remaining 8 participants were paid volunteers (3 men, 5 women). The volunteers were all highly trained musicians (graduate students at the Yale School of Music, 22–28 years old) who had agreed to serve in a series of sensorimotor and perceptual experiments at Haskins Laboratories. Although music training was not required for the task, advantage was taken of ready availability of this rhythmically skilled and highly motivated group of participants.

#### *3.6.1.2 Stimuli*

Once again, the 8 sounds of Experiment I were used. The sounds were tested in three main groups. As in Experiment I, there were 7 noise-syllable pairs consisting of each syllable paired with the reference sound and 6 syllable-syllable pairs (LA-PLA, PLA-SPLA, LA-SPLA, and PA-SA, SA-SPA, and PA-SPA). These two groups were used to form mixed sequences (in which the base sound and test sound differed) and each pair was tested in both orders (with each sound serving once as the base sound and once as the test sound). In addition all 8 sounds were tested singly in homogeneous sequences (in which the same sound served as base and test sound) and

this formed the last group. Taken together, there were 34 distinct sequences to be tested. These were divided into 3 sets: Sets 1 and 2 both contained various mixed pair sequences and shared the N-BA sequences in common (for consistency checking); set 3 also contained some mixed pair sequences but primarily consisted of homogeneous sequences.

The PCR method requires initial RPC estimates for all sounds to be used in mixed sequences so that EOS perturbations of the test sound can be approximately centred about the point of subjective isochrony. For this purpose, a pilot adjustment experiment was run testing all noise-syllable pairs 4 times in both orders. (Only the author participated and the mean IOI and adjustment range were 600 ms and ±250 ms respectively.) Estimated RPCs relative to N (analyzed as in Experiment I) were 7, 42, 53, 55, 106, 184 and 183 ms for BA, LA, PA, PLA, SA, SPA and SPLA respectively. To simplify the experimental software, silence was prepended to each sound according to its estimated RPC so that when the onsets of the modified sound files were isochronous the corresponding sounds would be approximately perceptually isochronous. The prepended silence ranged from 200 ms for N to 17 ms (= 200 - 183 ms) for SPLA. These silent delays were subtracted again in the data analysis.

Each trial consisted of a nearly isochronous sequence of varying length (generated on-line by the software) in which a base sound occurred repeatedly and a test sound was inserted from time to time. Each sequence contained 11 test sounds, with the number of intervening base sounds varying randomly from 4 to 6. The first test sound occurred in the 8$^{th}$ sequence position at the earliest. The IOI between base sounds was 700 ms, which prevented any overlap of base and test sounds. The 11 test sounds occurred at temporal offsets (EOS values) ranging from -50 to 50 ms, in increments of 10 ms, relative to the point of sound file onset isochrony. The order of EOS values within a sequence was random.

#### *3.6.1.3 Apparatus*

The experimental procedure was controlled by customized MAX/MSP 4.6.3 software (designed for MIDI applications) running on an Intel iMac computer (OS 10.4.10). The timing accuracy of the sequential audio output, which was controlled by the MSP (signal processing) component of the software, was verified by acoustic measurements to be within 1 ms. Measurements were also conducted to determine the electronic processing delay between the impact sound of a tap and a sound triggered by the tap via the MAX (MIDI) component of MAX/MSP. This revealed a mean delay of 26 ms; this constant was subtracted from the nominal tap-tone asynchronies (time of MIDI input minus theoretical time of sound output) registered by the MAX component of the MAX/MSP program, which also triggered the beginning of a sequence. Taps were registered by a Roland SPD-6 electronic percussion pad connected to the computer via a MOTU Fastlane MIDI interface. Sound sequences were presented diotically over Sennheiser HD540 Reference II headphones.

#### *3.6.1.4 Procedure*

Each stimulus set, repeated 5 times in different random orders (blocks), required a separate session of about 1 hour. The order of Sets 1 and 2 was varied between participants; the two sessions were typically one week apart. Set 3 was presented at a later time.

Participants sat in front of the computer and tapped manually on the percussion pad, which they held on their lap. Participants were free to tap in any style they preferred. They started each sequence by pressing the space bar on the computer keyboard and started tapping with the third sound they heard. They were instructed to stay in synchrony throughout and to ignore any small deviations from temporal regularity in the sequence. After each presentation of the block of trials, they saved their data in a file.

### *3.6.2 Results*

A total of 97,111 taps was recorded; a small additional number of expected taps (338) were not registered for various reasons. The PCR to each test sound EOS was calculated by subtracting the baseline IOI (700 ms) from the interval between the taps coinciding with the test sound and the following base sound (cf. Figure 3.3). Occasionally, a PCR could not be calculated because one or both of the critical taps had failed to be registered or were anomalous (double taps or unusually large asynchronies[19]). A total of 0.3% of the PCR data was excluded due to these causes. Simple linear regression of the PCRs on EOS magnitude was used to estimate the parameters of the PCR function (EOS axis intercept, slope, standard error of the estimate) separately for each participant, sound pair, and order.

Mean asynchronies were also calculated for all base event taps except the first three immediately following each EOS. Here again a small number of taps (0.1%) were excluded due to unusually large asynchronies. The mean and standard deviation of tap asynchrony were calculated separately for each sequence presentation (mean $N$ = 26.0), then aggregated across sequence repetitions within each participant's data. Once again hypothesis tests used the within-participant summary data only.

The main results, averaged across participants, are shown in Table 3.4 (mixed sequences) and Table 3.5 (homogeneous sequences). Sounds within each pair are ordered so that the less complex sound, which is also the sound with the earlier EPC, comes first. Within Table 3.4, noise-syllable sequences are followed by syllable-syllable sequences. All results for the pair N-BA were averaged – this pair had been presented in two separate sessions as a consistency check (with highly consistent results).

[19] Asynchronies with z-scores > 3.29 were excluded from the analysis. These generally occurred in the vicinity of skipped taps probably indicating that the participant had temporarily lost synchronisation with the pacing sequence.

**Table 3.4 PCR slope, EOS axis intercept, and tap asynchrony from mixed EOS sequences.**

| | | Slope | | Intercept | | Asynchrony | |
|---|---|---|---|---|---|---|---|
| Pair | Order | *M* | *SE* | *M* | *SE* | *M* | *SE* |
| N-BA | Fwd | 0.56 | 0.05 | -3.82 | 4.86 | 3.63 | 3.37 |
| | Rev | 0.65 | 0.06 | -9.54 | 2.78 | 7.18 | 3.33 |
| N-LA | Fwd | 0.55 | 0.06 | -39.46 | 5.44 | 1.72 | 2.87 |
| | Rev | 0.68 | 0.06 | -46.45 | 3.57 | 40.11 | 3.98 |
| N-PA | Fwd | 0.54 | 0.08 | -62.77 | 8.53 | 5.27 | 4.20 |
| | Rev | 0.70 | 0.06 | -59.68 | 2.22 | 56.30 | 2.98 |
| N-PLA | Fwd | 0.50 | 0.06 | -56.36 | 7.17 | 1.66 | 2.71 |
| | Rev | 0.69 | 0.04 | -58.29 | 2.99 | 52.78 | 5.43 |
| N-SA | Fwd | 0.54 | 0.08 | -125.05 | 7.89 | 4.79 | 3.11 |
| | Rev | 0.66 | 0.06 | -113.38 | 2.16 | 115.35 | 4.00 |
| N-SPA | Fwd | 0.52 | 0.08 | -181.07 | 4.68 | -0.55 | 3.05 |
| | Rev | 0.63 | 0.05 | -180.99 | 4.62 | 179.74 | 5.71 |
| N-SPLA | Fwd | 0.54 | 0.06 | -183.73 | 6.65 | 2.16 | 3.34 |
| | Rev | 0.68 | 0.08 | -182.76 | 4.35 | 178.58 | 6.27 |
| LA-PLA | Fwd | 0.54 | 0.06 | -17.08 | 6.53 | 49.15 | 3.39 |
| | Rev | 0.61 | 0.06 | -10.92 | 3.50 | 60.98 | 4.91 |
| LA-SPLA | Fwd | 0.50 | 0.05 | -137.64 | 8.50 | 40.21 | 4.23 |
| | Rev | 0.61 | 0.04 | -137.02 | 7.81 | 177.76 | 7.97 |
| PLA-SPLA | Fwd | 0.56 | 0.04 | -133.78 | 5.55 | 57.96 | 4.97 |
| | Rev | 0.53 | 0.06 | -113.37 | 7.35 | 178.71 | 7.14 |
| PA-SA | Fwd | 0.66 | 0.06 | -55.00 | 3.41 | 54.46 | 3.13 |
| | Rev | 0.62 | 0.06 | -47.51 | 4.04 | 112.71 | 4.48 |
| PA-SPA | Fwd | 0.54 | 0.05 | -117.89 | 6.17 | 56.77 | 6.35 |
| | Rev | 0.58 | 0.07 | -115.38 | 3.88 | 181.79 | 6.84 |
| SA-SPA | Fwd | 0.53 | 0.08 | -70.72 | 5.66 | 115.22 | 4.03 |
| | Rev | 0.60 | 0.05 | -66.69 | 4.16 | 181.36 | 5.51 |

*Note*— Each pair ($sound_j$-$sound_i$) acted in the roles base-test in the forward order (Fwd) and test-base in the reverse order (Rev). All intercept values shown are for $sound_i$ relative to $sound_j$, thus reverse order intercept values, measured for $sound_j$ relative to $sound_i$, were negated. (Negative intercept values imply positive RPCs). The asynchrony measure applies only to the base sound in each sequence. All values are in milliseconds.

**Table 3.5 PCR slope, EOS axis intercept, and tap asynchrony from homogenous EOS sequences.**

| | Slope | | Intercept | | Asynchrony | |
|---|---|---|---|---|---|---|
| Sound | *M* | *SE* | *M* | *SE* | *M* | *SE* |
| N | 0.84 | 0.07 | -0.66 | 1.11 | 9.15 | 2.38 |
| BA | 0.74 | 0.05 | -0.49 | 1.45 | 14.27 | 4.12 |
| LA | 0.70 | 0.05 | -4.43 | 1.66 | 52.54 | 4.54 |
| PA | 0.70 | 0.06 | -4.12 | 1.61 | 61.49 | 4.65 |
| PLA | 0.62 | 0.05 | -3.35 | 2.84 | 62.76 | 4.86 |
| SA | 0.61 | 0.06 | 0.14 | 2.88 | 120.95 | 4.51 |
| SPA | 0.53 | 0.05 | -2.42 | 2.40 | 181.98 | 5.80 |
| SPLA | 0.45 | 0.04 | -2.13 | 3.42 | 184.75 | 6.87 |

*Note*— All values are in milliseconds.

#### *3.6.2.1 PCR function standard error and slope*

Within-participant PCR variability was summarized by the PCR function standard error of the estimate (*SEE*). This statistic did not exhibit any consistent pattern (grand $M = 24.6$, $SD = 7.1$) and is not shown in Table 3.4 for that reason. A one way repeated-measures ANOVA showed no significant effect of sound on the *SEE* for homogeneous sequences, $F(7,56) = 1.35$, $\varepsilon = .52$, $p = .28$, $\eta_G^2 = .03$. For mixed sequences, a two way repeated-measures ANOVA revealed that the Order effect was nearly significant but small, $F(1,8) = 5.25$, $p = .051$, $\eta_G^2 = .01$. Neither the Pair effect nor the Pair × Order interaction was significant, $F(12,96) = 1.09$, $\varepsilon = .33$, $p = .38$, $\eta_G^2 = .02$, and $F(12,96) = 1.11$, $\varepsilon = .29$, $p = .37$, $\eta_G^2 = .02$, respectively.

The slope of the PCR function affects the confidence interval of within-participant RPC estimates, with shallower slopes resulting in larger confidence intervals and less certain estimates. In general, slopes were not excessively shallow, though they were rather variable (grand $M = 0.60$, $SD =$

0.19). Slope also showed a clear participant effect: Some participants exhibited consistently larger or smaller slopes than others[20]. The PCR function slope is also an estimate of $\alpha$, the phase correction parameter, and inspection of the data reveals some systematic variation. There was a wide range of mean slopes obtained from homogeneous sequences, with the steepest slope for N and the shallowest slopes for the syllables starting with consonant clusters. A one-way repeated-measures ANOVA on these data showed that the differences were substantial and highly significant, $F(7, 56) = 12.59$, $\varepsilon = .55$, $p < .01$, $\eta_G^2 = .37$.

The experiment design did not include all possible combinations of base sound and test sound. However, two subsets of sounds did include all combinations: N, LA, PLA, and SPLA; and N, PA, SA, and SPA. Figure 3.5 shows the PCR slope for each combination of base sound and test sound measured. Several effects are apparent. First, the range of slopes for mixed sequences tends to be smaller than the range for homogeneous sequences. Second, slopes show systematic variation by test sound for each base sound. This variation seems to follow the same trend as the corresponding homogeneous sequence slopes, except when N is the base sound. Finally, slopes for each test sound were generally (but not always) larger when the sequence was homogeneous rather than mixed.

A two-way repeated measures ANOVA on the subset of sounds N, LA, PLA, and SPLA found no significant effect of the base sound, $F(3, 24) = 1.00$, $\varepsilon = .75$, $p = .40$, $\eta_G^2 = .02$, a highly significant, moderate size test sound effect, $F(3, 24) = 25.52$, $p < .01$, $\eta_G^2 = .21$, and a small to medium interaction effect that approached significance, $F(9, 72) = 2.62$, $\varepsilon = .40$, $p = .06$, $\eta_G^2 = .09$. Planned contrasts indicated that homogeneous and mixed sequence slopes were not significantly different, $F(1, 8) = 3.17$, $\varepsilon = .28$, $p = .11$. A similar two-way repeated measures ANOVA on the subset defined by N, PA, SA, and SPA once again found no significant effect of the base sound, $F(3, 24) = 1.26$, $\varepsilon =$

[20] This variability in mean slopes seems to reflect individual differences in sensitivity to phase perturbations and the speed of response to such perturbations.

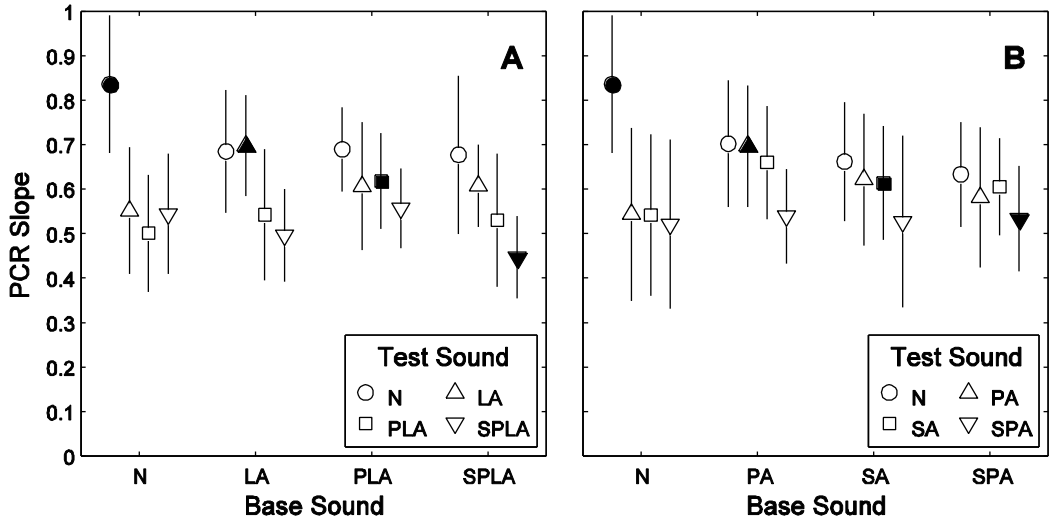


Figure 3.5 Mean slope of the PCR function for all combinations of the sounds N, LA, PLA, SPLA (A) and N, PA, SA, SPA (B). The slopes for all test sounds are clustered for each base sound. Mixed sequence slopes are shown with empty symbols whereas homogeneous sequences (with identical base and test sounds) are shown with filled symbols. Error bars show 95% confidence intervals for the mean slope.

.65, $p$ = .31, $\eta_G^2$ = .02, a moderate, highly significant effect of the test sound, $F(3, 24)$ = 14.34, $\varepsilon$ = .80, $p$ < .01, $\eta_G^2$ = .11, and a small non-significant interaction effect, $F(9, 72)$ = 1.72, $\varepsilon$ = .46, $p$ = .17, $\eta_G^2$ = .06. Again planned contrasts indicated that differences between homogeneous and mixed sequence slopes were not significant, $F(1, 8)$ = 2.73, $\varepsilon$ = .25, $p$ = .14.

### *3.6.2.2 PCR function EOS axis intercepts*

When the test sound is located at the point of subjective isochrony relative to the base sounds, there should be no perceived phase error on average and thus the expected PCR is zero. This point is located at the intercept of the PCR function and the EOS axis. For homogeneous sequences the intercept should occur when the EOS is zero. It is clear from Table 3.5 that the EOS intercepts for homogeneous sequences deviate very little from zero, as expected. Each deviation was subjected to a *t*-test, and though the LA and PA deviations were individually significant, with Bonferroni correction none of the deviations reached significance.

Intercepts for mixed sequences were sign-corrected according to the order of sounds within each pair because the null hypothesis was that the intercepts for both orders would be symmetric around zero. A two way repeated-measures ANOVA conducted on the EOS intercepts for heterogeneous sequences showed the expected large and significant effect of sound pair, $F(12,96) = 325.88$, $\varepsilon = .21$, $p < .01$, $\eta_G^2 = .93$. The main effect of Order was small and far from significance, $F(1,8) = 0.29$, $p = .61$, $\eta_G^2 = .01$, but the Pair × Order interaction approached significance, though its effect was small, $F(12,96) = 2.41$, $\varepsilon = .39$, $p = .058$, $\eta_G^2 = .05$. Bonferroni post hoc tests revealed a significant difference between orders only for PLA-SPLA, $CI_{.95}$ = -40.6 (lower) -0.2 (upper), $p < .05$; no other comparisons were significant.

Each PCR intercept directly estimates the RPC for a specific (mixed) sound pair and order. (The PCR intercept of homogeneous sequences cannot be used to estimate RPCs.) As there was no reliable effect of order, the intercepts from both orders for each sound pair were averaged (after sign correction) to form a single direct RPC estimate. For each direct estimate, up to two further indirect RPC estimates were calculated where the data permitted. All these RPC estimates are shown in Table 3.6. As usual, the hypothesis of P-centre context independence predicts that indirect and direct estimates would not differ significantly. This hypothesis was tested by pairwise comparisons of direct and indirect estimates for each syllable-syllable pair, thus there were 10 comparisons. Only the comparison of the direct estimate to the indirect estimate (via N) for the pair PA-SPA reached individual significance, $t(8) = -2.97$, $p = .02$. With Bonferroni correction, none of the comparisons were significant and so there was no evidence of context dependendence.

#### *3.6.2.3 Tap asynchrony in EOS perturbed sequences*

One participant showed extreme differences among asynchronies, tending to tap very early when the background syllable started with /s/. He also

showed large variability of asynchronies in some conditions, and his data were therefore omitted from analysis.

Despite the music training of the participants, there were considerable individual differences in variability, with some individuals being twice as variable as others. Nevertheless, the standard deviation of tap asynchrony did not exhibit any obvious systematic variation and is not included in Table 3.4 or Table 3.5. A one-way repeated-measures ANOVA confirmed that the base sound effect on the standard deviation of tap asynchrony was small and not significant, $F(7,49) = 2.29$, $\varepsilon = .43$, $p = .11$, $\eta_G^2 = .03$. A second one-way repeated-measures ANOVA showed that the effect of the inserted sound for sequences with N as the base was also not significant, $F(7,49) = 0.58$, $\varepsilon = .46$, $p = .64$, $\eta_G^2 = .02$.

The mean tap asynchrony in Table 3.4 and Table 3.5 shows the expected systematic effect of the base sound in each sequence. Since the tap asynchrony was expected to depend on the base sound only, no particular asynchrony relationship was expected between (mixed) sequences in which the sound roles had been reversed. To test the hypothesis that tap asynchrony depends on the base sound and not on the test sound in each sequence, a two-way repeated-measures ANOVA, with the base sound and test sound identity as factors, was conducted on each of the sequence subsets balanced for these factors. For the subset N, LA, PLA, and SPLA, the analysis showed the expected large and significant effect of base sound, $F(3, 21) = 853.70$, $\varepsilon = .49$, $p < .01$, $\eta_G^2 = .96$. The test sound effect was small and not significant, $F(3, 21) = 1.77$, $\varepsilon = .73$, $p = .20$, $\eta_G^2 = .01$, but there was a small significant interaction effect, $F(9,63) = 5.03$, $\varepsilon = .28$, $p < .05$, $\eta_G^2 = .07$. Pairwise tests indicated significant differences between each of the following: N and N-PLA, LA-N and LA, LA-N and LA-PLA, and finally LA-PLA and LA-SPLA. Analysis of the subset N, PA, SA, and SPA, once again showed the expected large, significant effect of the base sound, $F(3, 21) = 1005.03$, $\varepsilon = .52$, $p < .01$, $\eta_G^2 = .97$. The effect of the test sound was small and not significant, $F(3, 21) = 0.92$, $\varepsilon = .67$, $p = .42$, $\eta_G^2 = .01$, and, in this case, the

small interaction effect was also not significant, $F(9,63) = 1.95$, $\varepsilon = .41$, $p = .14$, $\eta_G^2 = .04$.

Unlike the PCR, the tap asynchrony measure allows RPC estimates to be made from both mixed and homogeneous sequences. Strictly, tap asynchrony provides a biased estimate of each base sound's EPC, but from these, unbiased RPC estimates can be derived (assuming that bias is constant between trials and that P-centres are context independent). Although there was generally no reliable effect of the test sound identity on the mean asynchrony, RPCs were derived slightly differently for mixed and homogeneous sequences. For homogeneous sequences, which are most similar to sequences used in Experiment II, the RPC for any pair of sounds was calculated as the asynchrony difference between their respective homogeneous sequences. So for example, the (homogeneous sequence) RPC of LA relative to N was obtained by subtracting the mean asynchrony of the N sequence from that of the LA. In mixed sequences, RPC estimates for each pair only used asynchronies from sequences containing that pair of sounds. Asynchronies from sequences which shared a base sound but had different test sounds were not combined. By way of example, the (mixed sequence) RPC of LA relative to N was obtained by subtracting the mean asynchrony of N-LA (the forward order where N acts as the base sound) from the complementary sequence, LA-N (the reverse order of N-LA where LA acts as the base sound). Direct RPC estimates were calculated for all noise-syllable pairs using both homogeneous sequence and mixed sequence data. Direct and indirect RPC estimates were calculated for all syllable-syllable pairs. All these estimates are shown in Table 3.6.

The RPC estimates (both PCR and asynchrony based) and the PCR method efficiency are compared with those of rhythm adjustment and simple tap asynchrony in the next section.

**Table 3.6 Mean direct and indirect RPC estimates from the PCR method.**

| | PCR RPC | | MTA RPC | | HTA RPC | |
|---|---|---|---|---|---|---|
| Pair | *M* | *SE* | *M* | *SE* | *M* | *SE* |
| N-BA | 6.68 | 3.34 | 3.56 | 2.71 | 5.12 | 2.39 |
| N-LA | 42.96 | 2.89 | 38.40 | 1.98 | 43.39 | 2.97 |
| N-PA | 61.22 | 4.88 | 51.03 | 2.53 | 52.34 | 3.04 |
| N-PLA | 57.33 | 3.78 | 51.12 | 5.17 | 53.61 | 3.44 |
| N-SA | 119.22 | 4.41 | 110.56 | 4.17 | 111.80 | 3.21 |
| N-SPA | 181.03 | 3.55 | 180.28 | 4.70 | 172.83 | 5.37 |
| N-SPLA | 183.24 | 3.64 | 176.42 | 5.50 | 175.60 | 5.67 |
| LA-PLA | 14.00 | 3.02 | 11.84 | 2.80 | 10.22 | 3.72 |
| via N | 14.37 | 2.92 | 12.72 | 4.06 | — | — |
| LA-SPLA | 137.33 | 7.12 | 137.55 | 4.77 | 132.21 | 3.63 |
| via N | 140.28 | 3.03 | 138.02 | 4.05 | — | — |
| PLA-SPLA | 123.57 | 4.81 | 120.76 | 4.00 | 121.99 | 4.64 |
| via N | 125.91 | 1.52 | 125.30 | 4.95 | — | — |
| via LA | 123.33 | 6.00 | 125.71 | 3.03 | — | — |
| PA-SA | 51.26 | 2.07 | 58.25 | 2.03 | 59.46 | 2.51 |
| via N | 57.99 | 2.11 | 59.53 | 3.89 | — | — |
| PA-SPA | 116.64 | 3.71 | 125.02 | 5.44 | 120.49 | 5.80 |
| via N | 119.81 | 4.80 | 129.25 | 3.93 | — | — |
| SA-SPA | 68.71 | 2.21 | 66.14 | 4.07 | 61.03 | 4.65 |
| via N | 61.81 | 4.30 | 69.72 | 3.18 | — | — |
| via PA | 65.38 | 3.35 | 66.77 | 5.91 | — | — |

*Note*—For each pair ($sound_j$-$sound_i$), the RPC values shown are for $sound_i$ relative to $sound_j$, i.e., $RPC_{ij}$. PCR RPC = RPC estimates from PCR function EOS axis intercepts; MTA RPC = RPC estimates from mixed sequence tap asynchrony; HTA RPC = RPC estimates from homogeneous sequence tap asynchrony. PCR and MTA RPC estimates are averaged from both possible role orders for each pair. Indirect RPCs via a third sound, *k*, were calculated as usual, so that $RPC_{ij} = RPC_{ik} + RPC_{jk}$. A dash indicates that the indirect RPC was mathematically identical to the direct RPC by definition and therefore redundant. All values are in milliseconds.

# 3.7 Method comparison

## *3.7.1 RPC estimate consistency*

An important motivation for this work was to investigate whether the methods in this study all measure the same percept and give consistent estimates. For comparison, the syllable-noise RPC estimates from each method are shown in Figure 3.6.

It is apparent that the methods generally yield very similar estimates despite having been measured in very different ways, and with different sets of participants. Despite this general similarity, the tap asynchrony method exhibits some of the largest confidence intervals, yields the smallest RPC estimates for most sounds, and, in particular, may differ significantly in its estimates for SPA and SPLA. As RPCs for the 3 methods were obtained from two independent groups of participants—non-expert musicians, for the rhythm adjustment (Experiment I) and tap asynchrony (Experiment II) methods, and expert musicians, for the PCR method (Experiment III)—an omnibus comparison of all methods was not performed and a subset of

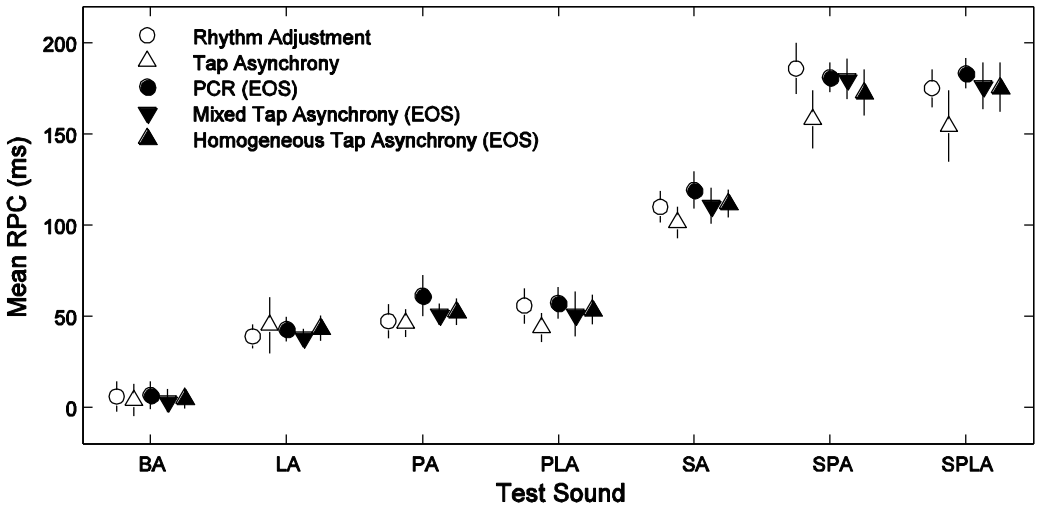


Figure 3.6 Between participant RPC estimates from each method compared. Symbols indicate the mean RPC relative to the reference noise, N, in ms. Error bars indicate the 95% confidence interval of the mean.

pairwise method comparisons was instead evaluated.

In the following analyses, the main effect of Pair (which was always highly significant by design) is not relevant to the hypothesis under and was not reported for that reason. RPCs from the rhythm adjustment and the tap asynchrony methods were compared in a two-way repeated-measures ANOVA which revealed a small to medium, significant effect of Method, $F(1, 6) = 7.42$, $p < .05$, $\eta_G^2 = .12$, while the Method × Pair interaction was not significant, $F(12, 72) = 1.80$, $\varepsilon = .24$, $p = .19$, $\eta_G^2 = .01$. A two-way mixed ANOVA, conducted on RPCs from the rhythm adjustment and PCR methods, revealed no significant effect of either Method or the Method × Pair interaction, $F(1, 14) = 0.50$, $p = .49$, $\eta_G^2 = .01$, and $F(12, 168) = 1.78$, $\varepsilon = .37$, $p = .14$, $\eta_G^2 = .08$, respectively. A further two-way mixed ANOVA, conducted on RPCs from the tap asynchrony and PCR methods, showed a medium size significant effect of Method, $F(1, 15) = 9.55$, $p < .01$, $\eta_G^2 = .17$, while the Method × Pair interaction approached significance, $F(12, 180) = 2.79$, $\varepsilon = .21$, $p = .06$, $\eta_G^2 = .11$.

The most methodologically similar RPC estimates are generated by the simple tap asynchrony method and homogeneous EOS sequence tap asynchrony. Unlike the other methods, these methods eliminate the possibility of within-sequence pair interactions from the execution of the experiment. Once again, a two-way mixed ANOVA was conducted. The effect of Method on RPC was of small to medium size, but did not reach significance, $F(1, 14) = 4.21$, $p = .06$, $\eta_G^2 = .11$. The Method × Pair interaction was also non-significant, $F(6, 84) = 2.00$, $\varepsilon = .40$, $p = .14$, $\eta_G^2 = .08$.

Figure 3.6 reveals a tendency for RPCs obtained from the tap asynchrony method to be smaller than those of the other methods for PLA, SA, SPA, and SPLA. The significance of this tendency could not be tested satisfactorily with the available statistical power. Furthermore, the tests that were performed on the existing data are inconclusive on this point: In the comparison of tap asynchrony with rhythm adjustment and tap asynchrony with the PCR method the method effect was significant, whereas in the

comparison of tap asynchrony with homogeneous EOS sequence tap asynchrony the method effect failed to reach significance.

In summary, there was no evidence of significant RPC differences between the rhythm adjustment and PCR methods. Therefore it appears that both these methods do indeed measure the same percept. The position regarding the tap asynchrony method is less certain.

### *3.7.2 Accuracy and efficiency*

The accuracy of each method was evaluated by comparing the within-participant standard error of the RPC estimate for equal numbers of trials and the between-participant standard error of the RPC estimate for equal number of participants. Since the method experiments had in fact used different numbers of trials and participants, these standard errors were estimated from the corresponding standard deviations.

Between-participant standard deviation of all the syllable-noise RPCs was averaged to calculate the between-participant standard deviation from which the standard error for various numbers of participants could be derived. Within-participant standard errors were calculated differently for each method. In rhythm adjustment, the adjusted asynchrony in each trial directly yields an RPC estimate for a given pair of sounds. The standard deviation of the trial estimates, calculated separately for each syllable-noise pair and participant, was used to calculate an average standard error for the method. In the tap asynchrony method the RPC for each syllable was calculated within each block as the difference between the mean asynchrony of the syllable trial and the reference noise trial in that block. The standard deviation of these estimates was averaged as before to calculate an average standard error. Finally, since the PCR method uses linear regression of observations from several trials to yield a single RPC estimate, it was not possible to directly measure the standard deviation of this RPC estimate. Instead, using bootstrapping with replacement (see

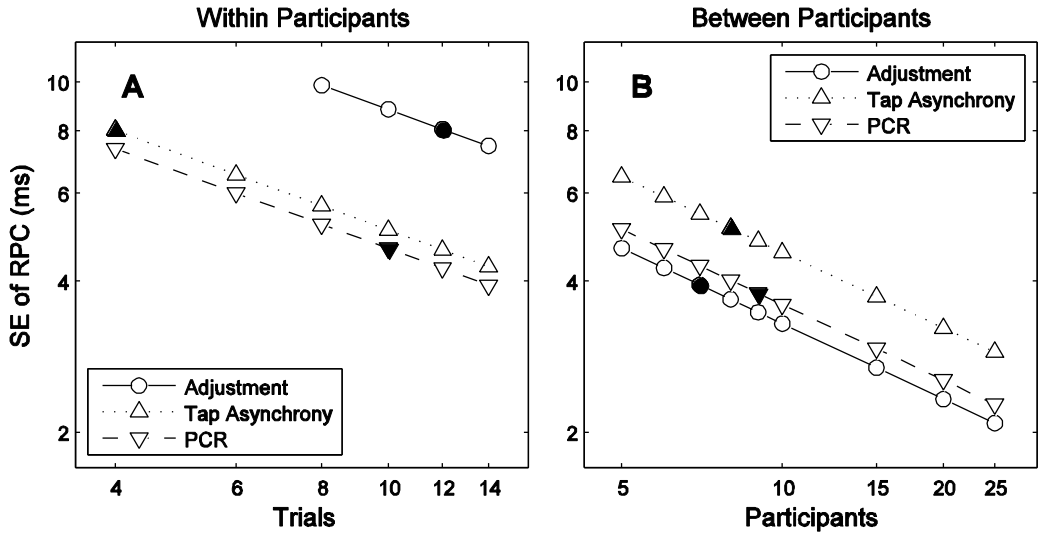


Figure 3.7 Standard errors of within-participant and between-participant RPC estimates. (A) Standard error of the within-participant RPC averaged across participants and syllable-noise pairs; (B) standard error of between-participant RPC averaged across syllable-noise pairs. Data derived (filled shapes) and extrapolated (empty shapes) standard errors are both shown.

Appendix B), the observations were resampled to estimate a standard error for each RPC and these were averaged as with the previous methods to calculate the average standard error.

The within-participant and between-participant standard errors of the RPC are plotted in Figure 3.7. It is clear that rhythm adjustment and the PCR method yield the most reliable between-participant RPC estimates with little difference between them.

Efficiency for each method depends primarily on the time requirements for each participant. The mean participant trial duration was rather similar for the adjustment and PCR methods, 48.2 and 43.0 seconds respectively. In both cases the time requirement for $N$ trials of $M$ test sounds (all paired with the same references sound) can be estimated simply as $N \times M \times T$, where $T$ is the trial duration (with some allowance for breaks between trials). It is worth noting, however, that the PCR method requires an initial investment of time in a pilot experiment to establish approximate RPCs and this time is not accounted for in the estimate. The mean trial duration for

simple tap asynchrony was 32.3 seconds, but unlike the two previous methods it requires trials for the reference sound to be performed separately from those of the sounds which will be referred to it. Thus for $N$ trials of $M$ test sounds the total time requirement can be estimated as $N \times (M + 1) \times T$.

The relatively small within-participant standard errors of the RPC for the simple tap asynchrony method suggested that this method could have used fewer taps per trial. The data were analyzed again using just the first 16 taps of each sequence. The averaged within-participant standard error for the experiment increased slightly (from 8.0 to 9.4 ms), but the between-participant standard error actually decreased slightly (from 5.1 to 4.7 ms). Using just the first 16 taps (18 pacing sound presentations), each trial could be completed in just 16.9 seconds.

In a similar manner, the PCR method data were reanalyzed using just 6 of the original 11 EOS levels, namely, -50, -30, -10, 10, 30, and 50 ms. Although the averaged within-participant standard error increased to 7.2 ms, the between-participant RPC estimates differed from the originals by less than 3 ms in all cases and the averaged between-participant standard error changed little (from 3.8 to 4.4 ms). Using these EOS levels, each PCR trial could be completed in just 23.5 seconds, providing a very useful reduction in participant time required.

## 3.8 Discussion

### *3.8.1 P-centre measurement*

This research evaluated three P-centre measurement methods using a common stimulus set so that P-centre estimates from each method could be compared. Experiment I used the most commonly applied method, rhythm adjustment, which acted as the control against which the other methods could be compared. Experiment II used the tap asynchrony method, and

Experiment III used the PCR method. The study had several objectives: to determine whether or not the methods produce consistent RPC estimates, to evaluate the efficiency of each method, to confirm or disconfirm the P-centre context independence hypothesis, and to investigate P-centre specific effects on the PCR (in particular the PCR function slope). Each of these questions will be considered separately.

#### *3.8.1.1 Consistency of estimates*

Estimated P-centres varied significantly among stimuli as expected in all three experiments, and all methods produced similar mean RPC estimates. The results also showed that the PCR method and rhythm adjustment are consistent, indicating that they measure the same percept. This finding is important because there does not appear to be any previous study which has explicitly compared P-centre measurement methods. Instead, a variety of measurement methods have been used with no evidence that they all measure the same percept. In fact, specific problems reported with other methods such as the existence of multimodal P-center distributions (Gordon 1987; Wright 2008) and underestimated interval durations (Fox & Lehiste 1987b) would suggest that it is dangerous to simply assume that all measurement methods are equally valid and comparable.

The rhythm adjustment method produced RPC estimates with the smallest between-participant variability. The variability obtained from the PCR method was a little higher while the simple tap asynchrony method exhibited the largest variability of all. It seems likely that much of the within-participant standard error difference between the PCR method and tap asynchrony is due to differences in the music skills of the participants and the number of participant trials that were run for each sound or pair to be tested. Between participants, however, the tap asynchrony method variability differs very obviously from that of the other two methods. Tap anticipation bias does differ between individuals, with some individuals tapping consistently earlier or later than others, but this alone should not

explain the greater dispersion of tap asynchrony RPC estimates because the RPC calculation method cancels out any constant within-participant bias. Of course, if the anticipation bias is not constant then it will not be cancelled out completely though the magnitude of its effect may be reduced.

In addition to exhibiting the highest RPC standard errors, the tap asynchrony method tends to underestimate RPC values (relative to the reference noise) when compared with the other methods in this study. The underestimation, which can also be interpreted as a reduced range of RPCs, is not constant (it is most pronounced for SPA and SPLA) and is therefore not due to a later P-centre in the reference noise sound alone. Surprisingly, this underestimation is clearly exhibited even relative to the methodologically very similar homogeneous EOS sequence tap asynchrony. The asynchronies in both methods should be similar unless the disturbance introduced by the EOS has an effect. If the EOS levels are approximately centred on the point of subjective isochrony for each test sound (as they must have been for homogeneous EOS sequences) then there should be no systematic effect on the asynchrony mean, though the variance about the mean may increase.

As noted previously, the two methods used independent participant groups with different music skills at different laboratories. Music skill may affect the variance of tap asynchronies and slightly reduce anticipation bias. Nonetheless, as long as the anticipation bias remains constant between trials, it is not clear how either effect could cause systematically underestimated RPCs. Another difference between the methods is, of course, the nature of an EOS itself: The EOS may disturb or reset some aspect of a participant's internal timing and synchronization mechanism, thus causing the observed RPC differences.

There is an important difference between the tap asynchrony method and the rhythm adjustment and PCR methods. The latter two methods always present two different sounds in a sequence and this allows RPCs to be measured directly. In contrast the tap asynchrony method only measures

EPCs and must derive RPCs assuming P-centre context independence. If this assumption is violated, the RPCs will not be correct. To give the observed data, the P-centre would have to depend on context such that it was earlier for homogeneous sequences and later for mixed sequences or sequences including EOS perturbations. This argument is examined again later when considering P-centre context independence.

The final explanation to consider is that EPCs measured may not be consistent with the RPCs measured in other methods, not because the P-centre depends on context (specifically, the identity of the previous sound), but because of a non-constant anticipation bias. If, for example, participants responded to sounds with less clear P-centres (SPA and SPLA) by increasing their anticipation this would give the observed results. In summary, further investigation is required to determine the cause of the RPC underestimation and the specific scenarios in which it occurs.

#### *3.8.1.2 Context independence*

P-centre context independence predicts that the RPC for a pair of sounds will be unaffected by their order, except for sign. Experiment I and Experiment III upheld this prediction, finding no significant effects of order for the rhythm adjustment or PCR methods. (Experiment II could not address this issue.) Context independence also predicts that direct and indirect RPC estimates should be equal. Experiment I and Experiment III explicitly compared direct and indirect RPCs, and neither experiment provided evidence of significant differences between them. Therefore these results support previous P-centre context independence findings for rhythm adjustment (Eling, Marshall & van Galen 1980; Marcus 1981) and extend those findings to the PCR method. The most important implications of this context independence for P-centre measurement are that indirect RPC estimates may be calculated from averaged direct measures and that, as a consequence, RPC estimates from different experiments or studies

using these methods may be compared provided the stimulus sets share at least one sound in common.

There is, however, one more context independence prediction to be considered. The mean tap asynchrony, measured in Experiment II and Experiment III, provides a biased estimate of the EPC for any given sound and it was predicted that this estimate should not be affected by the context in which the sound occurs. There were two potential violations of this prediction: The first, underestimation of RPCs by the tap asynchrony method, has already been discussed; the second involves homogeneous and mixed EOS sequence tap asynchrony. In EOS sequences, the base sound tap asynchrony should be independent of the particular test sound, as long as perturbations of the test sound are correctly centred about the point of subjective isochrony. It is therefore interesting that some EOS sequences involving N and LA did in fact exhibit a small effect of the test sound on the base sound mean asynchrony. One possibility is that, for at least some sounds, the average timing of the test sound P-centre did not correspond to the point of subjective isochrony relative to the base sound. If the mean perceptual timing deviates from the point of subjective isochrony, then the mean tap asynchrony will tend to be slightly biased in the same direction as the deviation[21]. There was not much evidence that EOS values were incorrectly centred however. On the contrary, the final data agree well with pilot RPC estimates.

It was previously suggested that a non-constant anticipation bias may explain the reduced range of RPC estimates resulting from the tap asynchrony method. Non-constant anticipation bias might explain the limited interaction of test and base sounds in EOS sequences. It is striking that homogeneous EOS sequence mean asynchrony for each sound is always

[21] Although asynchronies tend to return to their baseline within a few taps of an EOS, the speed of return depends on the $\alpha$ parameter. When EOS values are not correctly centred about the point of subjective isochrony, the calculated mean asynchrony will be biased in the direction of the mean PCR.

more positive than the corresponding mixed EOS sequence mean asynchrony (cf. Table 3.4 and Table 3.5). This does suggest that participants anticipate slightly more when a change of sound is introduced into the sequence. Although there is no reported precedent for this behaviour, perhaps participants find mixed sequences more difficult than homogeneous sequences (due to spectral and envelope changes between sounds) and tap more conservatively by anticipating more as a result. It has previously been shown that lighter (perhaps more hesitant) taps result in more negative asynchrony than forceful taps (Aschersleben, Gehrke & Prinz 2004).

A final possibility that cannot be completely ruled out is that P-centres are not context independent after all. The only potential evidence for this arises from tap asynchrony measures; neither the rhythm adjustment nor the PCR methods appear to exhibit any context dependence. Even with tap asynchrony measures, it is possible that there are distinct mechanisms at work in simple tap asynchrony (with an isochronous sequence) method and EOS sequence tap asynchrony.

In conclusion, there does not appear to be strong evidence of P-centre context dependence, at least within the constraints of this study, that is, for approximately equally loud isolated sounds with an interval of 650 to 700 ms between consecutive P-centres. Although context independence should hold at other intervals, it remains an empirical question to determine the range of intervals over which this is the case. For example, it seems likely that any substantial overlap between sounds (caused by intervals shorter than the sound duration) would affect the RPC. Despite the broad support for P-centre context independence, both the tap asynchrony RPC underestimation and EOS sequence tap asynchrony test and base sound interaction effect warrant further study.

#### *3.8.1.3 Efficiency (accuracy and duration)*

Within participants, the standard error of the RPC is notably higher for the adjustment method than the two tapping methods. This is reasonable, however, since each adjustment RPC estimate is derived from a single observation, whereas RPCs for the simple tap asynchrony and PCR methods are derived from multiple observations each. With 10 trials per sound, 5 sounds to be tested against a reference sound, and using the time-optimized method variants discussed in the method comparison the rhythm adjustment method could be expected to achieve an average standard error of 8.8 ms with an average 49 minute session for each participant. The corresponding estimates for simple tap asynchrony and the PCR method are 5.9 ms in about 10 minutes and 7.2 ms in about 28 minutes respectively. Clearly, the tap asynchrony method appears very attractive based only on the within-participant estimates.

However, it is apparent that within-participant RPC standard errors do not translate to corresponding between-participant standard errors in a straightforward manner (see Figure 3.7). With 10 participants and once again using the time-optimized method variants (not those plotted in the figure) the rhythm adjustment, tap asynchrony, and PCR methods would be expected to achieve average RPC standard errors of 3.3, 4.2, and 4.4 ms respectively. Once again, the tap asynchrony method would seem to be the most attractive were it not for its unexplained underestimation of RPCs. The PCR method is next most efficient, although time for at least one and possibly several participants to run a rhythm adjustment pilot experiment must be factored in when planning to use this method.

One final comment worth making is that the PCR method was tested with highly skilled musicians and it is likely that results would be less reliable with less skilled participants; higher within-participant variability could be expected and more participants may be required to achieve an equivalent between-participant standard error.

### *3.8.2 Phase Correction Response*

The phase correction response is primarily characterized by the parameter $\alpha$, estimated as the slope of the PCR function, which determines the magnitude of the correction in response to a phase change and hence the time (or number of taps) required to completely adjust to a new phase. The parameter $\alpha$ determines the weight given to the timing of external pacing events relative to internally planned tap events; it is an index of the strength of sensorimotor coupling. It can be interpreted as indicating how confidently a participant perceives the P-centres of pacing events. If the P-centre is difficult to locate accurately, then a participant cannot attribute it much confidence and should instead rely more on continuation of their established internal timing. On the other hand, if the pacing P-centre can be located accurately then responding quickly to any perturbations in the pacing sequence is a better strategy for staying synchronized.

The results of Experiment III show a clear effect of sound on $\alpha$ for homogeneous sequences. It is largest for the N sound; participants adjust their taps most confidently and rapidly to phase perturbations of this sound. In contrast, $\alpha$ is smallest for SPLA, the most complex syllable with one of the latest EPC estimates. Participants appear to adjust more tentatively and slowly when this sound is perturbed from perceptual isochrony. To explain these results, the subjective precision of the P-centre percept must be considered in more detail.

Some sounds have subjectively well defined and clear P-centres. Short sounds, percussive sounds, and the N sound in this study fall into this category. The P-centres of sounds with longer and more gradual or more complex onsets seem to have P-centres that are somewhat more ambiguous, or at least more difficult to detect accurately. This phenomenon is generally not reported in the literature with the possible exception of Rasch (1979), who suggested that "shorter and sharper rises of notes make better synchronization both necessary and possible" (p. 128). In particular the phenomenon does not appear to have been formally identified to date,

nor have there been any detailed studies examining it. As a consequence, the term P-centre clarity is introduced here to describe the subjective precision of a P-centre.

Although P-centre clarity was not formally investigated as part of this study, it seems that, for homogeneous sequences at least, $\alpha$ may be directly related to the perceived clarity of the P-centre. For mixed sequences, however, the situation is more complex. The perturbed test sound had a significant effect on the PCR function slope whereas the base sound did not appear to have an effect. The direction of the effect was generally the same as that of homogeneous sequences, suggesting that the PCR slope of mixed sequences was related to the perceived clarity of the test sound's P-centre. Mean slopes for mixed sequences appeared to be smaller than those of homogeneous sequences for each test sound, but this effect did not reach significance. Nevertheless, a reduction in slope, which can be interpreted as reduced confidence in localizing the test sound P-centre, suggests that a change of sounds results in a perceptual penalty. (There was also a suggestion of this penalty in the mean asynchrony data for EOS sequences.) A possible explanation for the penalty is the increased cognitive load when perceptual expectations, spectral and temporal, created by the repeated base sound are suddenly violated by the inserted test sound.

These results raise an interesting question: Is $\alpha$ constant throughout a sequence, or does it adapt to changes? Before the first EOS is encountered there is no difference between a homogeneous and a mixed sequence, so it would be natural to expect that the initial value of $\alpha$ in a sequence would be identical for both sequence types. After the first EOS, it is possible that there is a step change in $\alpha$ for mixed sequences which remains approximately constant thereafter. An alternative hypothesis is that $\alpha$ adapts gradually but continuously throughout the sequence. Yet another alternative is that that $\alpha$ depends only on the identity of the most recent pacing sound and therefore may change after each sound. Unfortunately, the experiments in this study cannot easily distinguish between the hypotheses. Certainly, the possibility

that the strength of sensorimotor coupling is continuously variable warrants further investigation.

## 3.9 Conclusions

The PCR method was shown to be a useful new method for measuring relative P-centres. It is essentially interchangeable with the more commonly used rhythm adjustment method both in terms of the mean and variability of RPC estimates that result, indicating that both methods measure the same percept. The PCR method's compelling advantage is that it does not require conscious decision making by participants, an advantage when some of the P-centres to be measured are relatively unclear. It also appears that the PCR method can be executed in less time than rhythm adjustment (though this should be confirmed with less musically skilled participants) and this is a definite advantage if trying to assemble a large corpus of P-centre labelled data. Despite the subjective difficulty reported by participants for some sound combinations, the data do not appear adversely affected and the rhythm adjustment method may be used if desired. In the context of RPC measurement, the main advantage of this method is its simplicity, both in terms of apparatus and subsequent data analysis.

The simple tap asynchrony method seems very attractive for several reasons: shorter participant time required, the subjective ease of the task, and the subsequent ease of data analysis. Unfortunately this method appears to exhibit differences from the rhythm adjustment and PCR methods and there is currently no explanation which would allow data resulting from the tap asynchrony method to be used in an interchangeable manner with data from these other methods. Further investigation would be required to determine why it is that simple tap asynchrony, which relies on asynchrony differences between trials, and the PCR method, which uses asynchrony differences between consecutive sounds, appear to yield different RPC estimates.

The data does not provide any evidence of P-centre context dependence for the rhythm adjustment and PCR methods, at least when changes in context are restricted to presenting different preceding or succeeding sounds. This finding is important because the assumption of P-centre context independence is the foundation on which RPC comparison within and between experiments using any of the methods in this study relies. The data on tap asynchrony is less definitive on this point and further investigation is required.

The term P-centre clarity was introduced to describe the subjective precision with which an event's P-centre is perceived. Though not specifically manipulated in this study, clarity seems closely related to both the abruptness of the event onset and the lateness of the P-centre relative to the event's onset. When sounds with relatively unclear P-centres are approximately isochronously timed, the dispersion of acceptable points of subjective isochrony might be expected to be wider than for sounds with clear P-centres. However the data appears to exhibit just one potentially reliable effect of P-centre clarity: the slope of the PCR function gets shallower for sounds with more complex onsets and less clear P-centres.

The final intriguing question raised by this study is how the strength of sensorimotor coupling (measured by $\alpha$) depends on the nature of the sequence and may change (or not) throughout the sequence. In particular the difference in tap asynchronies observed between simple isochronous and EOS perturbed homogeneous sequences deserves further attention.

Naturally it would be valuable to verify that the results of this study can be generalized to alternative sound sets including musical sounds, synthetic sounds, and alternative reference sounds. Perhaps the most efficient way to achieve this and yet meaningfully advance the state of P-centre research is to begin the process of building a P-centre labelled corpus and embed the generalization test within that effort.

# Chapter 4

# Neuroelectric correlates of the P-centre

Behavioural methods suitable for efficiently and consistently measuring event-local P-centres and relative P-centres were evaluated in Chapter 3. However, none of the currently known methods can objectively measure the absolute P-centre (the true moment at which the event perceptually occurs) directly. This limitation affects event-local P-centre (EPC) measures, which are biased by the inclusion of some unknown constant as a consequence, but not relative P-centre (RPC) measures. When measuring the temporal pattern of a sequence of events, any constant bias in the EPC will be cancelled out by the difference operation used to calculated intervals between any pair of events in the sequence. Nevertheless, EPC and RPC measures can only be applied to sequences of events with unambiguous onset times (see Chapter 3). Therefore, inability to measure absolute P-centres prevents accurate measurement of perceived temporal patterns in naturally produced event sequences including natural speech, many kinds of music performance, animal vocalizations, gestures and movements. Clearly it would be beneficial to develop a method of measuring absolute P-centres directly.

Although it would be preferable to measure absolute P-centres in an efficient and straightforward manner, even an inefficient or difficult method

could be used to develop, refine, and evaluate a general P-centre model applicable not only to isolated events with unambiguous onsets, but also to continuous event sequences. If the model's predictions are reliable, then model-predicted P-centres can be substituted for subjective P-centre measurements in many situations, just as psychoacoustic model predictions can often be substituted for subjective listening tests (e.g. ITU-T 2001). A general model should also be based on, or at least informed by, an understanding of the sensory and neurophysiological mechanisms that underlie perception of event timing.

For all these reasons it seemed appropriate to look for measurable neurophysiological correlates of the P-centre. Such correlates, if found, could elucidate the mechanisms of P-centre perception and provide an objective method of absolute P-centre measurement.

### *4.1.1 The basis for neurophysiological measurement*

The central nervous system consists of a very large, highly connected network of neurons that is neither anatomically nor physiologically homogenous. Neurons differ in details of their morphology and connectivity, and parts of the network exhibit functional specialization. All high level functions, including sense and perception, action control, and cognition, ultimately manifest as activity in this neural network. Measuring this activity yields particular insight into the otherwise invisible internal operation of perceptual and cognitive tasks. In the specific case of the P-centre, measuring such activity could potentially allow the moment at which an event occurrence is perceived to be detected.

To understand the basis for neurophysiological measurement, it is necessary to briefly examine the operation of the central nervous system at the cellular level of neurons. Each neuron can receive input signals from many sources and transmit its own output signal to many targets. Cellular outgrowths, particularly the axon, facilitate communication over distance.

The terminal interface between neurons is the *synapse*, comprising the presynaptic terminal, a small gap (the synaptic cleft), and postsynaptic receptor sites.

At rest, a neuron has a slightly negative potential. When an input signal activates a synapse, the potential in the postsynaptic region of the cell membrane changes (and this change may last over 100 ms). If the synapse is inhibitory, a transient change in chemistry makes the membrane potential more negative. Excitatory input, in contrast, makes the membrane potential less negative, partially depolarizing the cell. If the neuronal membrane is depolarized beyond a critical threshold a rapid change in cell chemistry results in the *action potential*, an electrical impulse, lasting about 1 ms. This flows as a wave of excitation over the cell membrane, and in particular along the axon toward other neurons. The action potential is an all-or-none signal (which does not vary in amplitude) and it is the basic information signal of the central nervous system.

Non-invasive detection and measurement of neural activity can be achieved with a variety of techniques. *Electroencephalography* (EEG) is based on measuring the very small potential differences (10 to 100 microvolts) that appear between electrodes connected to the scalp with conductive gel. These potential differences are thought to result from current in extracellular space produced by summation of the postsynaptic potentials from a large number of neurons, and not from the very brief action potentials (Fisch 1999). *Magnetoencephalography* (MEG) measures extremely weak magnetic fields generated by electric current within the brain. Unlike EEG, the field is thought to originate from intracellular currents flowing within the dendrites of neurons during synaptic transmission. As with EEG, synchronised changes in a large number of neurons are required to result in a measurable signal. *Functional Magnetic Resonance Imaging* (fMRI) measures the haemodynamic response associated with increased neural activity rather than any electrical or magnetic aspect of the neural activity itself. Neurons require more energy

when active and the blood supply is dynamically regulated to provide more energy where it is required. Near infrared spectroscopy (NIRS) also measures haemodynamic response but only relatively close to the brain surface.

Of the methods listed above, EEG and MEG have the best temporal resolution whereas spatial resolution is best with fMRI. When the objective is to correlate function with anatomical structure, spatial resolution is important. In this work, however, the primary objective was to identify neural activity temporally correlated with behaviourally measured P-centres. Only EEG and MEG are appropriate for this task and EEG has the benefit of somewhat more readily available and inexpensive equipment.

### *4.1.2 Neuroelectric correlates of sound and timing*

EEG recordings provide evidence of ongoing oscillatory activity in several frequency bands, named for the order in which they were first described. These frequency bands are: *delta*, 1–4 Hz; *theta*, 4–8 Hz; *alpha*, 8–13 Hz; *beta*, 13–30 Hz; and *gamma* which is variously interpreted as 36–44 Hz or an expanded range of approximately 20–60 Hz.

An *evoked potential* is a systematic change in ongoing EEG activity following presentation of a stimulus. It depends primarily on physical properties of the stimulus and is time-locked to it. Of particular relevance to this study is the *auditory evoked potential* (Davis 1939), the neuroelectric response to a sound stimulus. The evoked potential is sometimes called a signal-related potential or exogenous potential to signify its external dependence. In contrast, systematic EEG changes which depend on internal events that a participant generates in response to circumstances, state, and stimulus (for example detection of omission or mismatch in a sequence) are collectively called *event-related potentials* (ERPs), or endogenous potentials. Examples include P300 or P3, a large positive wave at a latency of about 300 ms occurring when a participant must respond to infrequent stimuli

interspersed with a larger number of frequent stimuli, and mismatch negativity (MMN), a negative deflection at a latency of 150–275 ms occurring when a participant detects a signal which does not match those that came before it (Gelfand 1998). Although the ERP traditionally referred to later response components associated with cognitive function, its definition is sufficiently broad to encapsulate evoked potentials also. Therefore the term ERP will be used in a general sense to refer to all event-related activity whether exogenous or endogenous.

The EEG is spatially imprecise and records potentials which are summed across a great many neurons with potentially diverse function (Goldstein & Aldrich 1999). Thus a single trial response can be difficult to discriminate from ongoing EEG activity. The traditional conceptualization of evoked potentials and ERPs is that individual components (waves) indicate neural activity bursts that are time locked to the eliciting event. This neural activity is superimposed on, and additive to, ongoing background EEG which is assumed to be independent of the eliciting event and can be modelled as noise. Using this model, the background or baseline EEG activity can be reduced by averaging time aligned response epochs to produce an average evoked potential (AEP) or ERP. The signal to noise ratio (SNR) of this averaged response improves with the number of epochs averaged and various methods for estimating the presence and quality of a signal in the averaged response exist (see for example Elberling & Don 1984; Stürzebecher, Cebulla & Wernecke 2001; Wong & Bickford 1980).

The traditional view of the ERP has been challenged in recent years. An alternative proposal is that the ERP results not from additive activity, but from phase resetting (and possibly amplitude modulation) of ongoing neural oscillation in response to experimental events (Klimesch et al. 2006; Klimesch et al. 2007; Makeig et al. 2002; Mäkinen, Tiitinen & May 2005; see also Sauseng et al. 2007 for a review). This alternative view has inspired additional analysis techniques. In particular, time-frequency analyses permit examination of *evoked power* (power in the EEG components that

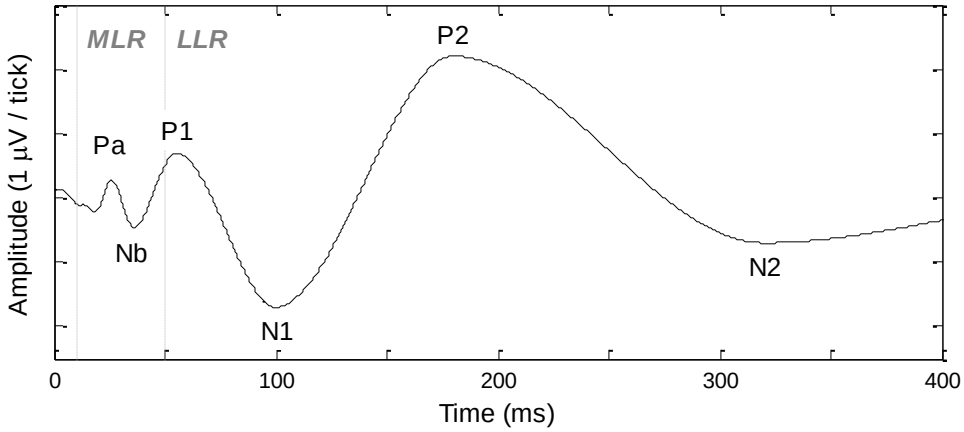


Figure 4.1 A synthetic auditory evoked potential (AEP) illustrating main features of the response to a very brief click. The AEP is divided into three post stimulus time spans: the early latency response (0–10 ms, not labelled in figure), the middle latency response (MLR, 10–50 ms), and the late latency response (LLR, 50–500 ms). The signal was filtered at 70 Hz (as is typical for enhancing and smoothing the main LLR and later MLR components) and the main positive and negative waves remaining are given their standard labels.

are phase locked to event onset), *induced power* (amplitude modulation that is time-locked to the event onset though the underlying EEG oscillations are not), and the *inter-trial phase coherence* (across trial consistency of phase angles at each time and frequency with respect to event onset). Roach and Mathalon (2008) provide a good overview of these techniques (see also Delorme & Makeig 2004; Tallon-Baudry et al. 1996).

The auditory AEP to very brief stimulus (usually a click) has a canonical morphology consisting of a number of identified positive and negative peaks at latencies up to 500 ms or so post stimulus onset, as shown in Figure 4.1. Early latencies (0–10 ms) are associated with activity in brainstem and these components are often called the *brainstem auditory evoked potential* (BAEP) or *auditory brainstem response* (ABR). Both the magnitude and latency of early components are used clinically, for example in threshold audiometry. Longer sounds (for example tones) cannot generally be used to study early latency components because neural activity is insufficiently synchronized to identify the response (Mason 2004). Furthermore, brainstem activity is associated with sensation rather than

perception, so it does not seem likely that early latency components can play a useful role in acoustic P-centre measurement.

The *middle latency response* (MLR), between 10 and 50 milliseconds, is associated with a succession of positive (P) and negative (N) waves: N0, P0, Na, Pa, Nb, and Pb (P1). The earliest of these are probably generated by subcortical sources but subsequent peaks have been associated with the auditory cortex (Eggermont & Ponton 2002). These peaks are often of rather small amplitude but, in the 10–60 Hz frequency range, seem relatively unaffected by stimulus repetition rate (Snyder & Large 2004).

The *long* or *late latency response* (LLR), between 50 and approximately 500 ms, is dominated by three large amplitude waves: P1-N1-P2. These waves appear to result from temporally overlapping components originating from different neuron populations (Eggermont & Ponton 2002). The response seems largely due to cortical activity and for that reason the term *cortical auditory evoked potential* (CAEP) is used synonymously with the LLR (Gelfand 1998). Unlike the early response and MLR, the LLR is particularly susceptible to variability dependent on participant state. For example, LLR amplitude depends on a participant's alertness and whether they attend to or ignore stimuli (Goldstein & Aldrich 1999). Snyder and Large (2004) also showed that the LLR for tones diminished in amplitude as repetition rate increased and essentially disappeared when tones were repeated at short random intervals (375–750 ms).

Although many auditory AEP studies use brief, spectrally homogenous, and simple stimuli (mainly clicks and short tones), auditory evoked potentials may also be elicited using more complex stimuli, including speech. Potentials elicited by speech are reliably reproduced and distinct tokens elicit distinct response waveforms (Tremblay et al. 2003). Responses to short, rapid onset, speech stimuli exhibit a single P1-N1-P2 complex not unlike responses to clicks and tones. However stimuli which are of longer duration or incorporate intensity and frequency changes elicit a response which consists of multiple overlapping P1-N1-P2 responses (Martin,

Tremblay & Korczak 2008). When a P1-N1-P2 response occurs in response to stimulus change (including offset) it is termed the *acoustic change complex* (ACC). In short consonant-vowel (CV) syllables, for example, the ACC is elicited by the transition from consonant to vowel. Burger et al. (2009) have shown that this response can be approximated by two overlapping tone responses separated by the voice-onset time.

A recent study found that the N1 and P2 latencies were shorter for /ta/ than for /da/ (Digeser, Wohlberedt & Hoppe 2009), a finding which contrasts with previously reported P-centres for these sounds (Harsin 1997). In addition to N1 (with latencies around 100 ms), a later broader wave, termed N250, whose latency varied in a manner that does not appear to match P-centre data (namely, earliest for a 250 ms tone, later for a speech syllable, and later again for 50 ms tone) has also been reported (Vidal et al. 2005). Finally, Sanders, Newport, and Neville (2002) found that N100 (N1) amplitude increased at the onsets of nonsense words in continuous speech after learning, concluding that N100 amplitude indexes speech segmentation. Thus, N100 seems important to the process of speech segmentation and perhaps acoustic event segmentation.

Although many of the studies listed appear to focus on the latency and amplitude of LLR components, there is evidence that time-frequency analysis may be particularly relevant. Recent studies have shown that phase resetting in the alpha and theta bands may be an important contributor to the P1-N1-P2 complex (Gruber et al. 2005; Kim & Han 2006; Low & Strauss 2009). Additionally, Luo and Poeppel (2007) found that theta band phase reliably discriminated spoken sentences and suggested that the theta period (125–250 ms) acted as a temporal segmentation window that reset as necessary to track continuously changing speech dynamics. Barry (2009) found that early exogenous components of the ERP arise substantially from phase resetting of ongoing EEG activity (in the delta, theta, and alpha bands) whereas later endogenous components result from evoked activity. Fuentemilla, Marco-Pallares and Grau (2006) investigated the attenuation

of N1 amplitude occurring during repeated presentation and concluded that the attenuated N1 resulted from transient phase coherence, whereas the initial non-attenuated N1 had an additional evoked power component.

Oscillatory activity in the gamma band has been implicated in the formation of coherent object representations, including those of auditory objects (Knief et al. 2000; Tallon-Baudry & Bertrand 1999). Palva et al. (2002) showed that evoked gamma band responses to speech and non-speech differed as early as 40–60 ms after stimulus onset, though frequencies below 20 Hz did not exhibit differences at this early stage of processing. They suggest that evoked gamma band activity may be sensitive to high level properties of the stimulus. Rodriguez et al. (1999), investigating recognition of faces, proposed that an early peak in induced gamma band activity (about 230 ms after stimulus onset) corresponded to the moment of perception itself. Recent research has also found that 40 Hz gamma band activity selectively enhances interactions between the auditory cortex, cerebellum, and thalamus (Pastor et al. 2002; Pastor et al. 2008). More significantly, a number of recent studies have shown that gamma band activity is associated with metrical and rhythmic expectancy (Snyder & Large 2002, 2005; Zanto et al. 2005; Zanto, Snyder & Large 2006). In particular, induced gamma band power peaks appear to be associated with temporal expectancy whereas evoked power peaks are associated with actual occurrence. In addition to signalling expectation of occurrence, an induced gamma band power peak at around 200 ms may indicate detection of omission (Gurtubay et al. 2006).

There is evidence that auditory rhythms activate areas outside the classical auditory system; notably the premotor areas are activated and appear to be involved in stimulus prediction (Bengtsson et al. 2009). Premotor activation may indicate readiness or preparation to synchronize. While it may respond to a rhythmic pattern of neurally coded P-centres it seems unlikely that it serves a primary function in their encoding. At different scales, interval timing seems to rely on different neural mechanisms and brain structures,

specifically the cerebellum for millisecond timing of discrete events (having a P-centre) and the basal ganglia for longer continuous events (Buhusi & Meck 2005).

In summary, the search for a neuroelectric correlate of the auditory P-centre should focus on features that occur at latencies corresponding to the MLR and LLR. Although stimulus onset and change appears to be associated with evoked activity, examination of induced activity and phase coherence may provide additional insight. All frequency bands require examination, but the most likely candidates for P-centre related activity seem to be the theta, alpha, and gamma bands.

## 4.2 The present study

The primary objective of the present study was to identify a neuroelectric correlate of the P-centre. If such a correlate were found, it would provide the first objective, non-behavioural method of measuring the P-centre. Perhaps more importantly it could identify the actual moment at which an event's occurrence is perceived. To date, no method of measuring this moment (the absolute P-centre) has been available. Thus the central hypothesis to be tested was that the P-centre has some measurable neuroelectric correlate and the corresponding null hypothesis was that it has none.

As discussed previously there is evidence that non-auditory areas are activated by rhythmic stimuli. For example, it is certainly the case that some coordination (direct or indirect) between auditory and motor areas must take place in order to perform a sensorimotor synchronization task, such as tapping in synchrony to a regular stimulus (see Repp 2005). Nevertheless, this study was based on assumption that it is the function of the auditory system to detect and encode features of an acoustic stimulus and that this function includes neural encoding of the auditory P-centre. For that reason

the experiments in this study confined electrode placements to those recommended for recording auditory evoked potentials.

The study was executed as two experiments, beginning with a pilot. The purpose of the pilot was to examine the responses elicited by a range of speech and non-speech stimuli. The speech stimuli were based on monosyllabic words that had been used in previous P-centre studies (Marcus 1981; Scott 1993; Villing, Ward & Timoney 2003). The non-speech stimuli were ramped tones with varying rise times. The P-centre of similar tones had been shown to depend on rise time (Vos, J. & Rasch 1981) and a dependency of cortical neuron first spike latency on rise time had also been demonstrated (Heil 1997) suggesting a possible link between the two. It was intended that collectively these speech and non-speech stimuli would exhibit a range of P-centre and acoustic features. Although there were several methodological questions to be answered by the pilot, the most important question was naturally: would there be any evidence of a response feature that correlated with the behaviourally measured P-centres?

The second experiment's purpose was to refine the methodology of the pilot and examine the response to stimuli whose P-centres had been behaviourally measured in Chapter 3. Specifically, previous experiments had suggested that there may be a difference between P-centres measured using tap asynchrony and those measured using rhythm adjustment. Therefore there was an additional question to be answered: would neuroelectric activity correlate more closely with P-centres measured using tap asynchrony (which uses a similar repeated homogeneous stimulus presentation paradigm) or those measured using rhythm adjustment (which uses alternating stimuli)?

Both traditional averaged response and modern time frequency analysis techniques were used to identify candidate features in both experiments. Contour features of sound are those which mark changes and transitions, including onsets, offsets, voice onset time, and amplitude modulation with

rates up to 20–30 Hz. Eggermont (2001) proposes that contour features of sound modulate the ongoing activity of neurons, controlling, for example, the degree of neural synchrony. Empirical P-centre measurements, generally exhibit dependence on contour features of sound and therefore may be correlated with modulations of neural synchrony. Whether such modulations would become apparent in the evoked activity, induced activity, or phase coherence remained to be seen.

The primary measures investigated were the latencies of local extrema in the processed signals (AEP frequency bands, evoked and induced power, and inter-trial phase coherence). Although the amplitude of EEG responses often varies in systematic ways with stimulus, such amplitude changes did not seem to be good candidates for correlating with a temporally precise P-centre measure. Latencies which exhibited systematic variation between stimuli would be selected as candidate neuroelectric predictors of the P-centre.

Candidate predictors would be evaluated by regression against the behaviourally measured P-centres. If the P-centre is associated with sufficient specific neural activity at the moment of its perception, then a candidate predictor with a regression slope close to 1 should be found. Alternatively, if the neural representation of the P-centre is indirect, then a predictor may still be found but its slope will not be 1.

## 4.3 Experiment IV Pilot

The aim of the pilot experiment was to measure the EEG response to speech and non-speech stimuli and identify candidate features that may correlate with behaviourally measured P-centres. The experiment incorporated both the behavioural P-centre measurement and EEG response measurement elicited by the repeated presentation of the same stimuli.

### *4.3.1 Methods*

#### *4.3.1.1 Participants*

Only the author participated in P-centre measurement. EEG recordings were also made with just one male participant (aged 21 at the time of the experiment). This participant had no formal musical training and had not been involved in previous P-centre experiments[22].

#### *4.3.1.2 Materials and equipment*

Two different stimulus sets were investigated in this experiment: natural speech and synthetic tones. Speech stimuli comprised the monosyllabic digits, "one", "two", "five", and "six", produced naturally by three speakers (one female, speaker A, and two male, speakers B and C). Speakers were asked to produce the digits at a speaking rate corresponding to a "marching pace" while ensuring that there was a separation between words. This ensured that digit durations were relatively short ($M$ = 441 ms) and could easily be isolated within the recording (see 0). Speech recordings were single channel with a sampling rate of 11025 Hz and 16 bit resolution. Individual digit productions were trimmed for length, but otherwise not edited. Therefore there was some natural deviation in both the peak level (up to 5dB between speakers[23]) and acoustic realization of each digit.

Synthetic stimuli consisted of six equal duration 1 kHz tones. Each tone had a cosine shaped ramped onset (20, 40, 60, 80, 120, or 160 ms), a cosine shaped damped offset (fixed at 80 ms), and a constant amplitude mid portion whose duration compensated that of the onset such that total

[22] EEG recordings were conducted by Chris Soraghan. The author was not present for the recordings but conducted all analyses presented herein.

[23] The peak level of each digit was compared between speakers using an exponentially weighted moving average (with a 125 ms time constant) of the instantaneous peak level calculated in accordance with (ITU-R 2006).

duration of each tone was 240 ms. Tones were synthesized with a sampling rate of 16 kHz and 16 bit resolution and the peak level of all tones was equal.

The mixed harmonic and noise reference sound described in Chapter 3 again served as the common reference for rhythm adjustment. For convenience the reference sound is hereinafter referred to as N, the synthetic tones as T*rt* (where *rt* is the rise time in ms, for example T20), and the speech sounds as S*Sd* (where *S* is the speaker identifier and *d* is the digit, for example SA1).

*Behavioural measurement*. Sounds were paired for P-centre measurement and formed two groups. Speech pairs consisted of all 12 speech stimuli paired with the reference sound in both orders, giving 24 unique permutations to be tested. Synthetic pairs consisted of the 6 tones again paired with the reference sound in both orders, giving 12 unique permutations.

Custom java software (see Appendix B) running under Windows XP on a personal computer implemented the rhythm adjustment method as described previously for estimating relative P-centres (cf. Chapter 3). The range of adjustment was ±400 ms and adjustments were possible with a 1 ms resolution. All stimuli were digitally resampled to 44100 Hz for rhythm adjustment. The digital audio for each sequence was mixed in real time at this rate, and then converted to analogue by an M-Audio USB Duo 2 audio interface connected to a notebook computer via USB. Stimuli were presented monaurally (right ear) using the Eartone 3A insert earphone (driven from the headphone output of the audio interface) in a quiet room. A sound attenuating earplug was used in the contralateral (left) ear to block low level environmental noise. The listening level was adjusted for comfort once, and then fixed for the duration of the rhythm adjustment experiment.

*EEG measurement*. EEG signals were recorded with a Biopac MP100 system and ERS100C amplifier module (Biopac Systems, Inc., Goleta, CA).

The positive electrode was attached at the forehead (Fpz), the negative electrode to the right, ipsilateral earlobe (A2), and signal ground to the left, contralateral earlobe (A1). Although use of the vertex (Cz) is more conventional for evoked potential audiometry, Fpz is an accepted alternative when attachment to Cz is difficult or unreliable (Goldstein & Aldrich 1999). The ERS100C amplifier gain was 50000. Signals were filtered with a 1 Hz high pass filter (6 dB/octave roll-off) and sampled at a rate of 2 kHz. The evoked potential measurements were made in a quiet room that was not electrically shielded. Sound was presented identically to the rhythm adjustment experiment except that the sound level was not fixed for the experiment duration. Repetition (looping) of the stimulus was controlled by Goldwave (Goldwave Inc., St. Johns, NL, Canada) and the evoked potential recordings were synchronised to the stimulus presentation by means of an embedded trigger signal and custom hardware (see Appendix B for details).

#### *4.3.1.3 Procedure*

P-centres were measured behaviourally using the rhythm adjustment method, following the same general procedure as described in Chapter 3. In this case, the mean inter-onset interval was 700 ms and the cycle duration was 1400 ms. The test sound asynchrony at the start of each trial was chosen randomly from the discontinuous range -200 to -100 ms and 100 to 200 ms for reasons explained previously. Speech pairs (24 trials) and synthetic pairs (12 trials) were tested in separate blocks and the order of trials was randomized each time a block was tested. Each block was presented 6 times over the course of two sessions on consecutive days.

For EEG acquisition, the participant first seated themselves comfortably with eyes closed and then listened passively to the stimulus sequence. Each sequence comprised 500 isochronous repetitions of a single stimulus presented with an inter-onset interval of 1518 ms. The EEG was recorded in

400 ms epochs triggered to begin 20 ms before each stimulus origin[24]. Sequences were grouped in blocks for presentation. Speech blocks, comprising all digits for a single speaker, were presented three times for speakers A and B and twice for speaker C giving (3 + 3 + 2) × 4 (repetitions × digits) speech sequences in all. Synthetic stimulus blocks, comprising all tones, were presented twice so there were just 2 × 6 (repetitions × tones) synthetic stimulus sequences.

#### *4.3.1.4 Analysis*

Rhythm adjustment results were analysed in the usual manner to estimate the P-centre of each stimulus relative to the common reference sound (see Chapter 3).

EEG recordings from each stimulus set were pre-processed differently. For speech stimuli, artefact rejection and averaging had been performed by the MP100 system during acquisition and only the resulting AEP was available for subsequent analysis. This precluded estimation of the AEP signal quality and induced power estimation. In contrast, individual EEG epochs were saved for each synthetic stimulus sequence. In this case artefact rejection and averaging were performed in MATLAB. An epoch was rejected if any of the following conditions were met: the absolute value of the amplitude exceeded 50 μV, the absolute value of the amplitude gradient (difference between consecutive samples) exceeded 50 μV/sample, or the within-epoch amplitude range exceeded 80 μV. The artefact free epochs for each sequence were averaged to yield the within-block AEP for a single stimulus.

Where single EEG epochs had been saved, the quality of each AEP was evaluated in two ways. First, the modified single point variance ratio, $F_{SP}^{*}$

[24] In this context the stimulus origin refers to the beginning of the digitized data for the stimulus and not to the physical or perceptual onset of the sound. In particular, the sound onset for speech data always occurred after an initial period of "silent" background which varied between sounds.

(Stürzebecher, Cebulla & Wernecke 2001), was evaluated using data from latencies of 50–250 ms (referred to stimulus origin). This provides a more reliable estimate than the original single point $F_{SP}$ of Elberling and Don (1984). Whether using the modified or original version, values of $F_{SP}$ over 3.1 indicate the presence of a signal[25] with $p < .01$. This is a lower limit for response detection (useful in threshold audiometry for example), however signal reproducibility improves as $F_{SP}$ gets larger and so larger values are desirable. The signal to noise ratio (SNR) of the AEP was derived from $F_{SP}^{*}$ and the corresponding lower limit for response detection is 3.2 dB.

Consistency between block AEPs was evaluated by calculating the correlation coefficient between each possible pair of blocks for a given stimulus. Values close to 1 would indicate that individual block AEPs match each other closely. Subsequently, the block AEPs for each stimulus (2 or 3 for each speech sound, 2 for each tone) were combined to form a single stimulus AEP. This AEP was filtered with zero phase shift band pass filters (using MATLAB's `filtfilt` function) corresponding to the delta, theta, alpha, beta, and gamma bands of the EEG.

Time-frequency analyses of EEG activity were performed using the continuous wavelet transform[26] and the complex Morlet wavelet with a constant bandwidth approximately 27% of its centre frequency (for details see Delorme & Makeig 2004; Roach & Mathalon 2008; Snyder & Large 2005). The evoked response power spectrum was calculated by wavelet transformation of the AEP. The induced response power spectrum could only be calculated where individual epoch data had been saved. Each artefact-free epoch was individually transformed and the resulting power spectra were then averaged between epochs.

[25] The degrees of freedom for the *F* test are chosen conservatively for the numerator because consecutive signal values are correlated. In this case, $F(5,500)$, was used.

[26] Torrence and Compo's wavelet software (1998) was used.

AEP bands, evoked power, and induced power were all subject to exploratory analysis seeking candidate features for neuroelectric measures of the P-centre. Candidates were identified when the latency of local extrema exhibited systematic variation between stimuli. All candidate features were subjected to simple linear regression against the measured RPC. Both the slope and coefficient of determination ($R^2$) were assessed as indicators of predictor quality.

### *4.3.2 Results and discussion*

#### *4.3.2.1 Behavioural measurement*

There were some subjective difficulties with the stimuli used in this experiment. The speech sound recordings incorporated various noise qualities (including a background hiss and speaker breath noise) which tended to stream apart at the rather short repetition interval of the rhythm adjustment task. The sound quality of the tones was very different to that of the reference sound and, while the difference was not so great as to induce streaming, alignment was subjectively more difficult than for most speech sounds. The main results of the rhythm adjustment experiment are shown in Table 4.1.

It is apparent from the standard errors that some of the pooled RPC estimates, notably those for SA1 and SC5, are less reliable than others. Nevertheless, the results generally exhibit good consistency between orders ($d$ < 15.1 ms) for all stimuli except SA1 ($d$ = 34.7 ms). The speech sounds exhibit a wide range of relative P-centres but it must be noted that in some cases this is due to delayed onset of sound energy in the stimulus recording rather than a late P-centre within the sound (see 0 for detailed waveforms for all stimuli). Although a wide range of rise times (20–160 ms) were used to synthesize tones, the relative P-centre estimates span a range of less than 20 ms. This result is consistent with Scott's findings using a ramped synthetic vowel (Scott 1998), but somewhat less of an effect than found by

Vos and Rasch (1981). Additionally it appears that the relative P-centres are clustered around two dichotomous values: one around 2 ms for T20, T40, T60 and T80 (implying that there is essentially no P-centre difference between these sounds) and another near 18 ms for T120 and T160.

**Table 4.1 Relative P-centres obtained by rhythm adjustment**

| | RPC | | Pooled RPC | |
|---|---|---|---|---|
| Stimulus | Fwd | Rev | *M* | *SE* |
| SA1 | 112.2 | 77.5 | 94.8 | 7.9 |
| SA2 | 22.2 | 20.7 | 21.4 | 5.3 |
| SA5 | 19.8 | 20.0 | 19.9 | 4.7 |
| SA6 | 80.8 | 87.7 | 84.3 | 4.3 |
| SB1 | 13.3 | 7.7 | 10.5 | 4.0 |
| SB2 | 21.7 | 25.3 | 23.5 | 3.9 |
| SB5 | 13.3 | 23.5 | 18.4 | 3.8 |
| SB6 | 65.5 | 62.3 | 63.9 | 5.6 |
| SC1 | 203.2 | 188.5 | 195.8 | 4.8 |
| SC2 | 51.0 | 51.2 | 51.1 | 2.7 |
| SC5 | 124.3 | 109.2 | 116.8 | 7.7 |
| SC6 | 113.0 | 110.0 | 111.5 | 5.8 |
| T20 | 6.2 | -4.3 | 0.9 | 3.1 |
| T40 | 3.3 | 5.7 | 4.5 | 3.8 |
| T60 | 7.3 | -3.3 | 2.0 | 3.6 |
| T80 | 1.0 | -0.8 | 0.2 | 4.4 |
| T120 | 20.7 | 14.3 | 17.5 | 3.8 |
| T160 | 30.2 | 7.7 | 18.9 | 4.1 |

*Note*—All RPC estimates are expressed in terms of the stimulus relative to the reference sound, N. In the forward order (Fwd) N was the base and the stimulus was the test sound, whereas in the reverse order (Rev) the roles were reversed. Pooled RPC = RPC estimates were pooled between orders. Both the mean (*M*) and standard error (*SE*) are shown.

#### *4.3.2.2 EEG Measurement*

Data analysis uncovered a minor problem with the trigger apparatus which affected EEG data collected for speech stimuli. Specifically, epochs for 17 of the 32 recording blocks were triggered late, 18 ms after stimulus origin instead of 20 ms before stimulus origin as intended. For analysis and display, therefore, all blocks were time aligned to the stimulus origin. As a consequence, stimulus AEPs obtained by combining block AEPs variously spanned the ranges -20 to 380 ms (all block AEPs triggered on time), -20 to 418 ms (some AEPs triggered on time, some triggered late), and 18 to 418 ms (all AEPs triggered late).

Between-block correlation coefficients for individual stimuli ranged from 0.76 to 0.96 for all sounds except T60 (.58) and SB1 (.33). The root mean square error (RMSE) between each block AEP and its corresponding stimulus AEP (averaged between blocks) did not exceed 0.53 microvolts. For synthetic stimuli, $F_{SP}^{*}$ ranged from a rather marginal 3.4 up to 11.0; the corresponding SNR ranged from 3.8 to 10.0 dB.

For each of the synthetic tones a summary of the AEP results, including band pass filtering and time frequency representations, can be seen in Figure 4.2.

Several observations can immediately be made. Despite variations in the rise time of the tones, the morphology of the AEP appears quite consistent. The large negative deflection more than 100 ms after tone onset almost certainly corresponds to the N1 component of the click AEP. The timing of the waves before and after this is consistent with a P1-N1-P2 complex. The progression between stimuli is not entirely consistent, however. In particular the negative wave following P2 does not reach the same depth as the one before it for either T60 or T80 but does for both shorter and longer rise times.

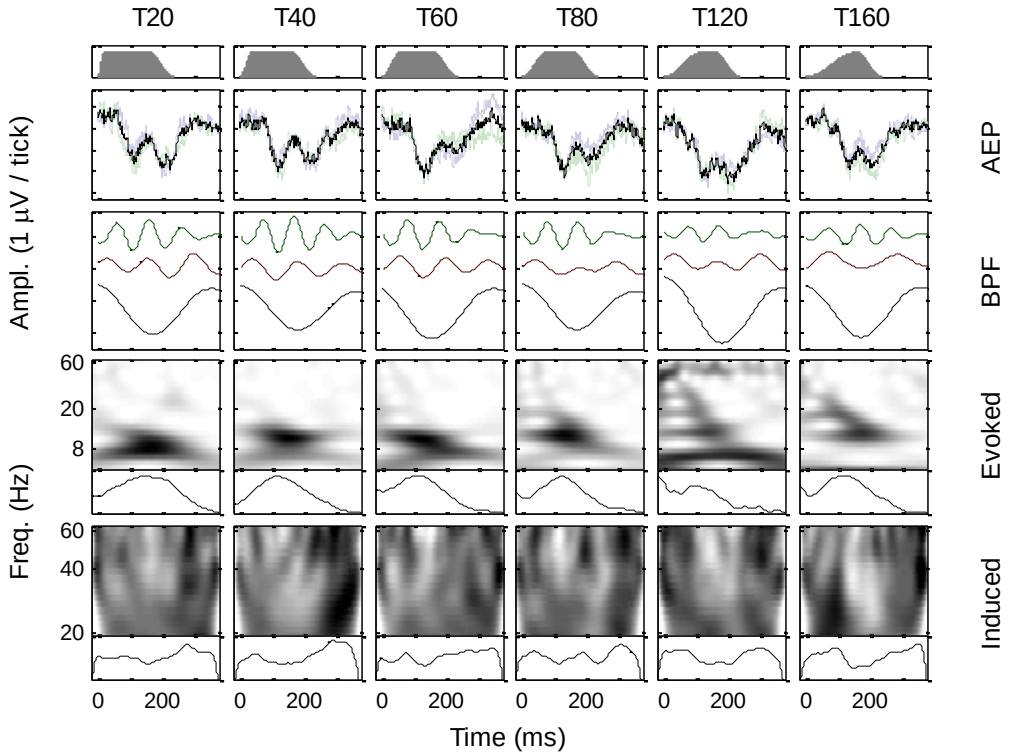


Figure 4.2 Processed ERPs for synthetic tones. For each sound (columns) the panels from top to bottom are: the stimulus amplitude (arbitrary units); the AEP for Fpz-A2 (heavy line = stimulus AEP, light lines = block AEPs); the filtered AEP (top to bottom: alpha, theta, delta band); the evoked response time frequency and time average plot (5–60 Hz); and finally, the frequency normalized induced response time frequency and time average plot (20–60 Hz).

Filtering the AEP reveals a single negative deflection in the delta band whose phase appears to change slightly between stimuli. The theta and alpha bands both exhibit distinctive oscillations that reach a maximum soon after stimulus onset and decay over the remainder of the stimulus duration. (Part, but not all, of the initial increase and subsequent decay in amplitude can be attributed to data edge effects associated with filtering. These effects have been minimized but cannot be removed completely with the available epoch durations.) Both the phase and amplitude of the oscillations vary with stimulus, though the amplitude appears largest for the tones with shortest rise times.

The evoked response spectrogram excludes delta band frequencies both because this band contains considerable energy which would dominate the spectrogram and because at this low frequency boundary effects of the wavelet transform make the energy estimates unreliable (see for example Addison 2002, pp. 56-62). Each evoked response features time-frequency regions of significant energy (which appear darker in the figure). For rise times up to 80 ms this energy appears in the alpha band whereas the two longer rise times also show considerable AEP energy at higher frequencies.

The induced response spectrogram has been normalized relative to the average power in each frequency. The level of induced gamma band activity does not vary much; essentially all activity is confined to a range of ±15% around the average and this is what the figure shows. While there are some activity extrema which do not appear in the evoked spectrogram there does not appear to be a systematic pattern.

AEP data for speech stimuli were processed in a similar manner to those of synthetic tones except that no induced response could be evaluated for reasons described above. The processed data for each of the speakers are shown in Figure 4.3, Figure 4.4, and Figure 4.5 respectively.

Similar to the data for synthetic tones, a number of observations can be made. Despite fairly large differences between speaker productions of the each of the four digit tokens, there are obvious between-speaker similarities in the morphology of the AEP for each token. Once again, the most consistent feature among the AEPs is the large negative wave occurring about 150–300 ms after stimulus onset.

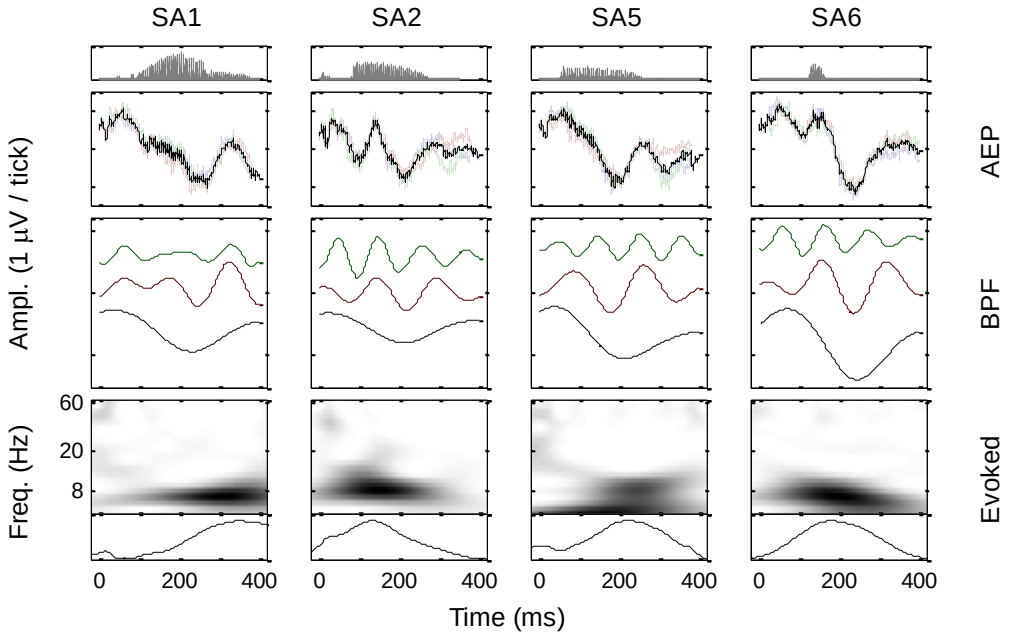


Figure 4.3 Processed ERPs for speaker A speech sounds. Panels are organized as Figure 4.2 with no induced response panels.

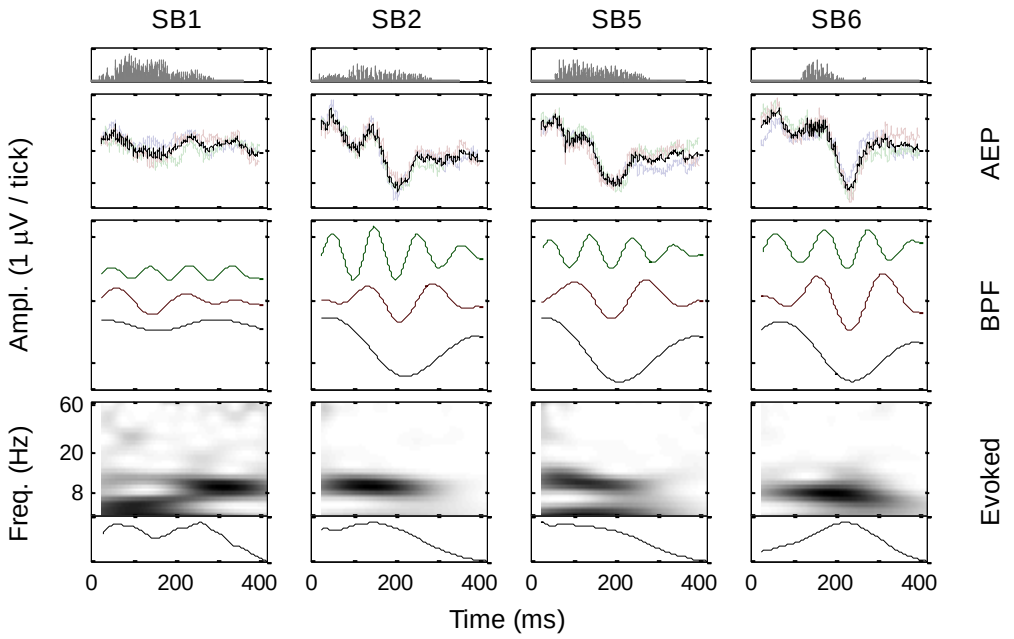


Figure 4.4 Processed ERPs for speaker B speech sounds. Panels are organized as Figure 4.2 with no induced response panels.

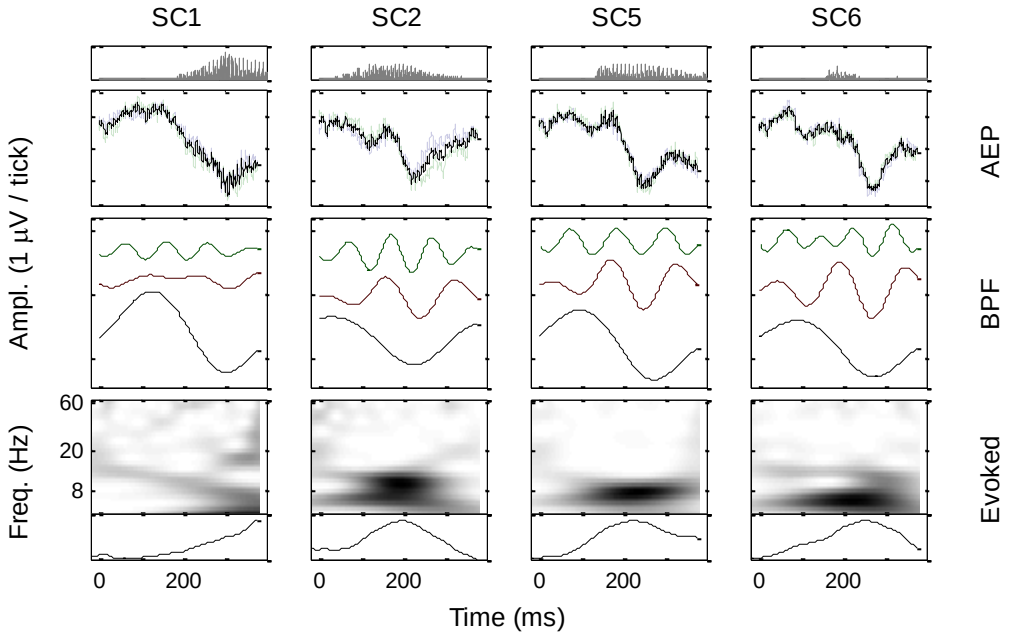


Figure 4.5 Processed ERPs for speaker C speech sounds. Panels are organized as Figure 4.2 with no induced response panels.

Again it is the delta band of the filtered AEP that appears to vary systematically between stimuli. The amplitude of this band varies somewhat and stimulus SB1 in particular elicited a very low amplitude response. There is no evidence that this low amplitude was due to recording conditions; the amplitude was replicated across three blocks recorded at different times. There is apparent evidence of phase resetting (or entrainment) of the delta band activity to the stimuli and, in all cases, this can be measured using the latency of the local minimum. The theta and alpha bands for speech AEPs show a similar pattern to those of tones, except that the peak oscillation amplitude occurs somewhat later, particularly in the theta band. No consistent phase pattern is apparent by inspection.

Each of the evoked response spectrograms displays an energy peak in the alpha band and several also reveal an energy peak in the theta band. In at least two cases (SA5 and SB1) the lower frequency peak occurs earlier.

Based on the observations above the following candidate features were identified: the signal minimum in each of the delta, theta, and alpha bands

of the AEP; the maximum of the evoked power in each of the theta, alpha, and gamma bands; and finally, the signal maximum of the gamma band induced power.

The result of the linear regression fitting the latency of these candidate features to RPC can be seen in Figure 4.6. The alpha band of the AEP is not included in the figure because it was a very poor fit ($R^2$ = 0.10). The theta band fit was also poor and not significant, $F(1,16) = 2.351$, $p = .14$. In this case however, it appeared that the fit for stimuli with RPCs larger than approximately 50 ms would be better, and indeed that proved to be the case: $y = -309.4 + 1.6\,x$, $R^2 = 0.87$, $F(1,5) = 34.026$, $p < .01$. Whether the

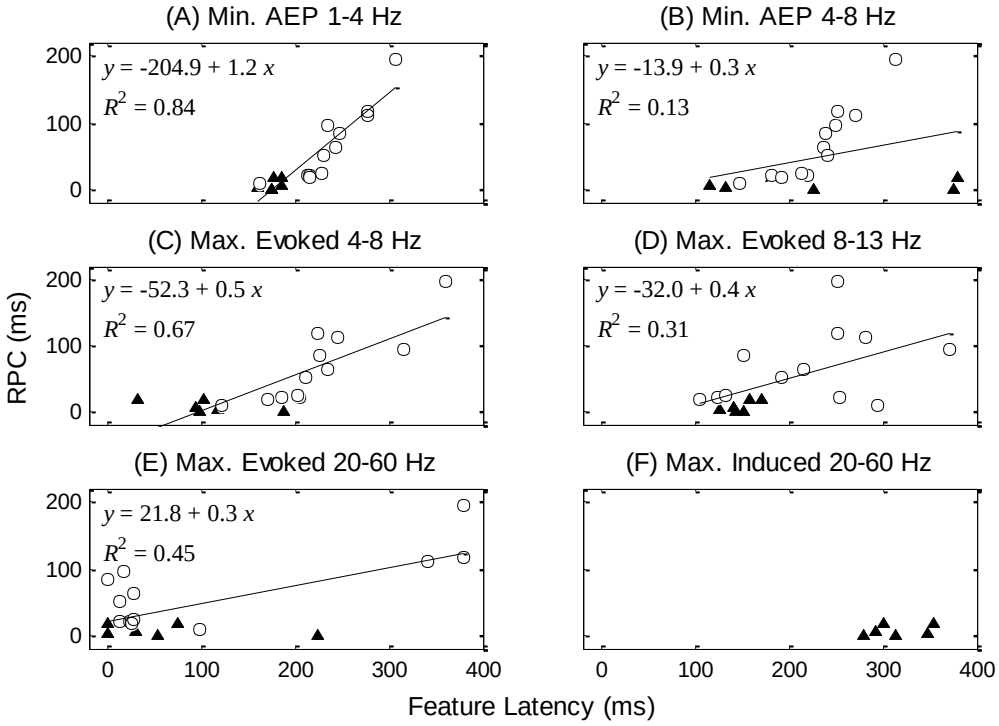


Figure 4.6 Simple regression of candidate feature latency against relative P-centres (RPCs) of each stimulus. (Filled triangles = synthetic tones, open circles = natural speech.) Candidate features examined in each panel are as follows: (A,B) latency of the filtered AEP minimum in delta (1–4 Hz) and theta (4–8 Hz) bands; (C–E) latency of the evoked power maximum in the theta, alpha (8–13 Hz), and gamma (20–60 Hz) bands; and (F) the latency of the normalized induced power maximum in the gamma band. The regression fit, regression equation, and coefficient of determination ($R^2$) are also shown in all panels except the last where the data was judged insufficient to warrant a regression fit.

existence of this improved theta band fit for later P-centres represents a switch between neural mechanisms, a limitation of the feature detection algorithm, or a statistical anomaly is not clear.

The best fit is found in local minimum timing of the delta band AEP and this fit was highly significant, $F(1,16) = 75.413$, $p < .001$. However, the slope of the fit is greater than 1 indicating this feature changes somewhat more slowly than the RPC. This suggests that phase resetting or entrainment of the delta band is strongly influenced by the P-centre but does not directly encode its moment of occurrence.

This was an encouraging result which suggested at a minimum, that neuroelectric activity may be affected in a systematic and predictable manner by the P-centre of the auditory stimulus. Nevertheless, this pilot experiment also highlighted a number of methodological improvements to be made before attempting to replicate the results. Specifically, the pilot used just a single participant, and this should naturally be extended to several participants before any conclusions regarding the generality of the findings could be made. Although the synthetic tones had the benefit of varying along a single parameter dimension, the narrow range of P-centres measured argued against continued use of these stimuli. Furthermore, for reasons already noted, a change to better quality speech stimuli was also warranted. Finally, the pilot had demonstrated that recording short EEG epochs causes problems for subsequent analysis due to edge effects (where the filter or frequency domain transform runs out of data). Recording longer epochs was therefore a requirement for any subsequent experiment.

## 4.4 Experiment V

The purpose of the second experiment was to repeat the general approach of the pilot, incorporating improvements based on that experience. One of the main questions to be answered was: could the results of the pilot experiment be replicated using alternate stimuli and multiple participants?

This experiment used alternative EEG equipment which allowed continuous EEG recording, thereby circumventing the short epoch problems of the pilot. As a side effect, all epochs would be saved enabling full time frequency analysis after. Both the equipment and recording environment were more sophisticated than those used in the pilot, suggesting another question: would this change in equipment noticeably improve the quality of the EEG recordings or clarify reproducible features of the EEG response?

Several additional changes were also incorporated. Stimuli whose P-centres had been previously measured using multiple techniques by several participants were used (see Chapter 3). Because P-centres measured using tap asynchrony and rhythm adjustment exhibited significant differences, candidate predictors would be regressed against each set of P-centre measures separately. Friberg and Sundberg (1995) showed that the just noticeable anisochrony increases with the stimulus inter-onset interval (IOI). It therefore seems possible that the temporal precision of the P-centre may depend on the IOI and for that reason this experiment used the same IOI during EEG recording as had been used for behavioural measurement (700 ms rather than the ~1500 ms of the pilot experiment).

### *4.4.1 Methods*

#### *4.4.1.1 Participants*

In addition to the author, there were three unpaid volunteer participants (one male and two female, all 21 years old). All but one of the volunteers had previously taken part in Chapter 3 and none had any known hearing deficiencies. All participants were native speakers of English and had various levels of musical training, though none were highly trained.

#### *4.4.1.2 Materials and equipment*

Five sounds for which P-centres had previously been measured were used again (see Chapter 3 and 0). These were the mixed harmonic and noise reference sound, N, and four naturally produced monosyllables, BA, PA, SA, and SPA. These syllables were chosen because they exhibited a wide range of P-centres relative to N. Values for BA, PA, SA, and SPA measured with the rhythm adjustment method (6, 47, 110, and 186 ms respectively) and tap asynchrony method (4, 46, 101, and 158 ms respectively) were both used.

Sound presentation was nearly identical to the pilot experiment. Stimuli were again presented monaurally (right ear) using the Eartone 3A insert earphone driven from the headphone output of an M-Audio USB Duo 2 audio interface (see Appendix B). A sound attenuating earplug was used in the contralateral (left) ear to block low level environmental noise. The audio interface headphone output was set at the same level that had been used for the measurement experiments in Chapter 3 and this level was fixed for all participants and for the duration of the experiment. To mitigate differences between the listening conditions[27] used during P-centre measurement and those used here, digital amplification (9 dB) was applied to all stimuli. Stimulus repetition (looping) to form sequences was controlled by Goldwave and EEG recordings were synchronised to the stimulus presentation by means of an embedded trigger signal and custom hardware (see Appendix B).

EEG signals were instrumented using a Brainvision QuickAmp-136 (Brain Products GmbH, Gilching, Germany) connected via USB to a Windows XP based personal computer. Electrodes were always attached at the vertex (Cz), the ipsilateral and contralateral earlobes (A2 and A1 respectively), and at the forehead (Fpz). The impedance of all electrodes was maintained

[27] The P-centre measurement experiment used diotic listening (which is approximately twice as loud as monaural listening) and HD 280 Pro headphones (which sound louder than the Eartone 3A with identical input because of sensitivity and frequency response differences).

at less than 5 kΩ. Electrode signals were not filtered and were digitized (at 2 kHz) relative to an average reference, then recorded and saved using the Brain Vision Recorder software. Participants were seated in a darkened, electrically shielded room during each recording session.

#### *4.4.1.3 Procedure*

Participants were asked to listen to stimulus sequences passively and were not given a task to control for vigilance. Each sequence comprised 500 isochronous repetitions of a single stimulus presented with an inter-onset interval of 700 ms[28]. In each session participants first listened to a short click sequence (with a 200 ms IOI) to validate the experimental setup. This was followed by two experimental blocks. Each block comprised five sequences, one for each sound, and the order of sequences was randomized in each block. All but one participant took part in two sessions, each yielding 2 × 2 × 5 (sessions × blocks × sounds) EEG recordings; the remaining participant took part in just one session.

#### *4.4.1.4 Analysis*

EEG recordings were analysed using custom MATLAB scripts. Because high frequency components would not feature in subsequent analyses, EEG signals were first down-sampled by a factor of 4 to 500 Hz. Next, the signals were digitally filtered, in all cases using MATLAB's `filtfilt` function, a zero phase shift filter implementation which doubles the effective filter attenuation. EEG signals were first filtered with a 1 Hz high pass filter (12 dB/octave) which removed slow DC drift. All subsequent filtering was applied to these DC corrected signals.

[28] Behavioural P-centre measurement for the sounds used in this experiment had used an IOI of 650 ms with the rhythm adjustment method and 700 ms with the tap asynchrony and PCR methods.

The main analysis band (1–70 Hz) was filtered from the DC corrected signals using only a low pass filter (70 Hz) whereas all other bands used band pass filters. Effective filter cut-off slopes were 24 dB/octave in all cases. Frequency bands initially filtered included the standard delta (1–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (12–30 Hz), and gamma (20–60 Hz) frequency bands. Eventually this set was supplemented by others, detailed in the results section. (By eliminating edge effects, filtering the continuous EEG signals at this early stage of processing provided a more accurate signal representation than later filtering of single trial or averaged epochs).

For subsequent analysis the filtered EEG signals were divided into epochs, time locked to each stimulus in the sequence (signalled by trigger instances during recording). Each epoch extended from 700 ms before its stimulus origin to 1400 ms after it, thus spanning three inter-onset intervals. This long epoch duration improved time-frequency analysis for reasons described previously. Epochs containing artefacts were excluded from further processing. Artefact detection was performed using the DC corrected (but otherwise unfiltered) EEG signals. An artefact was identified if any of the following conditions were met: the absolute value of the amplitude gradient (difference between consecutive samples) exceeded 100 μV/sample; the amplitude range within a given segment exceeded 200 μV; or the RMS amplitude (in any 2 ms window) exceeded 50 μV.

All artefact free epochs for each stimulus sequence were averaged to yield the within-block AEP (main analysis band) for that stimulus. The quality of this AEP was evaluated using the modified single point variance ratio, $F_{SP}^{*}$, (evaluated at latencies of 50–350 ms relative to stimulus origin) and the corresponding SNRs were also calculated.

Consistency between block AEPs for each participant was evaluated by calculating the correlation coefficient between each possible pair of blocks for a given participant and stimulus. For each stimulus, block AEPs were

averaged within-participant to form the participant AEP. The mean square error (RMSE) between the block AEPs and the participant AEP indicates the absolute size of any inconsistency and this measure provided additional insight. Consistency between participants was assessed in a similar manner. Correlation coefficients were calculated between all pairs of participant AEPs for a given stimulus, and the RMSE between these participant AEPs and their average, the stimulus AEP, was also evaluated.

Time-frequency analyses of EEG activity were performed broadly in line with Experiment IV with two exceptions. First, the relative induced power, an extension to the induced power spectrum, was obtained by normalizing relative to the mean power spectrum of a baseline period which in this case spanned the 250 ms just prior to the stimulus origin. Second, the inter-trial phase coherence calculation was calculated using the same approach as had been used for the induced power spectrum: the magnitudes of all time-frequency coefficients for each individual epoch were normalized (to 1) and then averaged between epochs; the phase coherence is then given by the magnitude of the average at each time-frequency point. Where phases broadly align between epochs, the phase coherence will be close to 1, but where phases are randomly distributed, the phase coherence will be close to zero.

As in the pilot experiment, candidate predictors of the P-centre were identified by exploratory analysis of the AEP bands, evoked power, induced power, and inter-trial phase coherence. As before, the regression slope and coefficient of determination ($R^2$) were assessed as indicators of predictor quality.

### *4.4.2 Results and discussion*

Approximately 14% of epochs met the artefact detection criteria and were excluded from further analysis. Examination of EEG consistency between blocks (within-participant) revealed more variability than had been

observed in the pilot experiment and this can be seen in Figure 4.7 which shows all within and between participant AEPs.

Correlation coefficients between blocks for each participant and sound ranged from 0.34–0.94 ($M$ = 0.73). These values were somewhat lower that those of the pilot and the least consistent AEPs were those in response to the PA sound for participants S1, S2, and S4. The maximum RMSE between the within-participant block AEPs and their corresponding stimulus AEP was 0.56 microvolts ($M$ = 0.39). The value of $F_{SP}^{*}$ averaged between all block AEPs was 5.0 and the corresponding SNR was 6.0 dB.

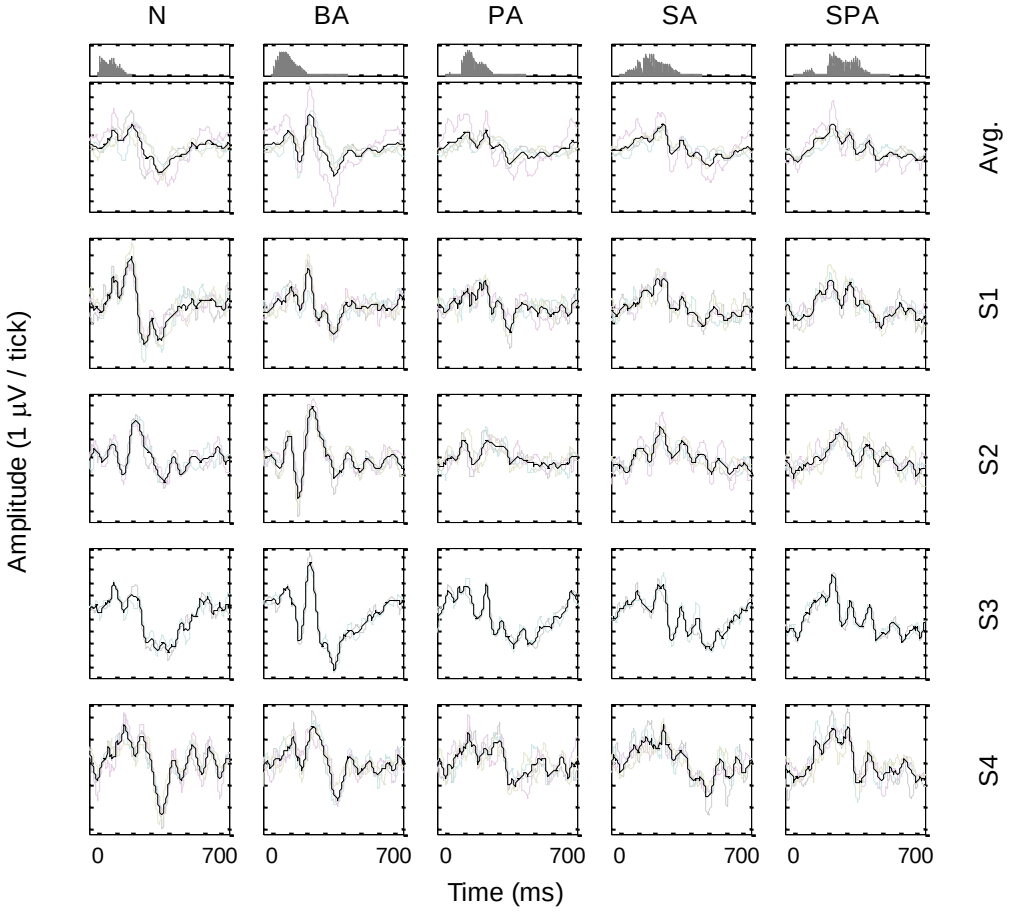


Figure 4.7 Within and between participant AEPs (Cz-A2). For each stimulus (columns) the panels from top to bottom are: the stimulus amplitude (arbitrary units); the summary AEP (heavy line = between-participants, light lines = within-participant); and each participant's AEP (heavy line = between-blocks, light lines = within-block).

Unlike the pilot experiment, which had just one participant, this experiment allowed consistency between participants to be examined also. Comparison of within-participant AEPs for each sound yielded correlation coefficients which ranged from 0.27–0.80 ($M$ = 0.66). The RMSE between the within-participant and between-participants AEP for each sound peaked at 1.12 microvolts (for sound BA, $M$ = 0.60 microvolts). In this case, however, there is another factor to consider. The AEP magnitudes for individual participants were not normalized in any way before being combined into the between-participants (grand) average, yet the response magnitudes for participant S3 in particular, are notably larger than those of the other participants and this inflates the apparent RMSE.

Figure 4.8 shows a summary of the main data processing (averaged between participants). The prominent negative wave of the AEP observed in the results of Experiment IV can be seen here in the AEPs for N and BA but it is not as clear for the remaining stimuli. All the AEPs exhibit clear oscillations with a period around 100 ms, consistent with increased alpha band synchronization during "eyes closed" EEG recording.

The AEP sub-bands exhibited a number of distinctive features. The delta band of the AEP began with a positive wave in all cases. In the pilot experiment, this positive wave was not present for tones and was distinctive only for speech sounds with RPCs larger than about 50 ms. The local minimum which followed was narrowest and deepest for N and BA, but broader and shallower for the remaining sounds. This seemed to indicate frequency modulation or interacting components within the delta band which differed between sounds. Similar to the pilot experiment, both the theta and alpha bands of the AEP featured oscillations which increased to a maximum and then decayed over the course of the sound. The peak amplitude of the theta band was approximately synchronous with (or slightly leading) that of the alpha band for all sounds except SPA. Furthermore there was some evidence of both phase continuity changes and period changes in these bands.

The evoked power featured a distinctive peak in the alpha band for all sounds. A power peak of similar size can also be seen in the theta band for BA. In all cases this power peak appears to be associated with change in frequency, rising for N, BA, SA, and falling for SPA.

Inter-trial phase coherence averaged between participants was low to

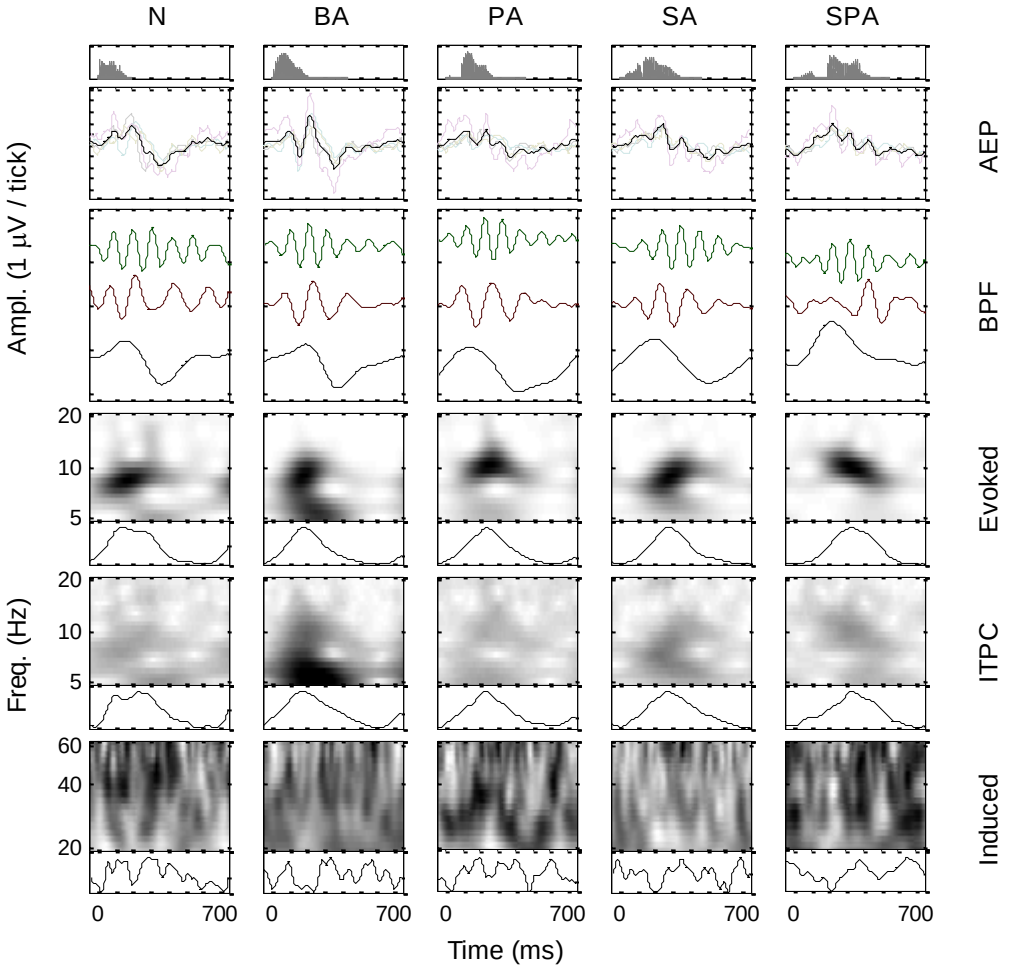


Figure 4.8 Between-participant summary of processed ERPs (Cz-A2) for all stimuli. For each stimulus (columns) the panels from top to bottom are: the stimulus amplitude (arbitrary units); the AEP (heavy line = between-participants AEP, light lines = within-participant AEPs); the AEP sub-bands (top to bottom: alpha, theta, and delta bands); the evoked power spectrogram and power plot (5–20 Hz); the inter-trial phase coherence (ITPC) and average coherence plot (5–20 Hz); and finally, the baseline-relative induced power spectrogram and power plot (20–60 Hz).

moderate in general (99% of values ≤0.30). Nevertheless there was a clear peak in phase coherence for all sounds and the pattern of phase coherence closely matched the pattern of evoked power in the corresponding frequency band. For this reason, it appeared that the inter-trial phase coherence did not contribute any new insight or information not already available from the evoked power.

The baseline-relative induced power featured numerous peaks and valleys. However it was difficult to discern any systematic pattern that would be amenable to automated feature detection. Furthermore, there did not appear to be any clear evidence of peaks indicating rhythmically or metrically anticipated beats as previously reported (Snyder & Large 2005; Zanto, Snyder & Large 2006).

In the pilot experiment there was just one participant and candidate features could be identified directly from inspection of the summary data. In this case, examination of the data had already revealed some differences between participants and thus final selection of candidate features required more detailed within-participant examination and comparison. Figure 4.9 shows between-participants and within-participant AEP bands. Because the between-participants delta band showed signs of interacting components it was subdivided into two ranges: 1-2 Hz and 2–4 Hz.

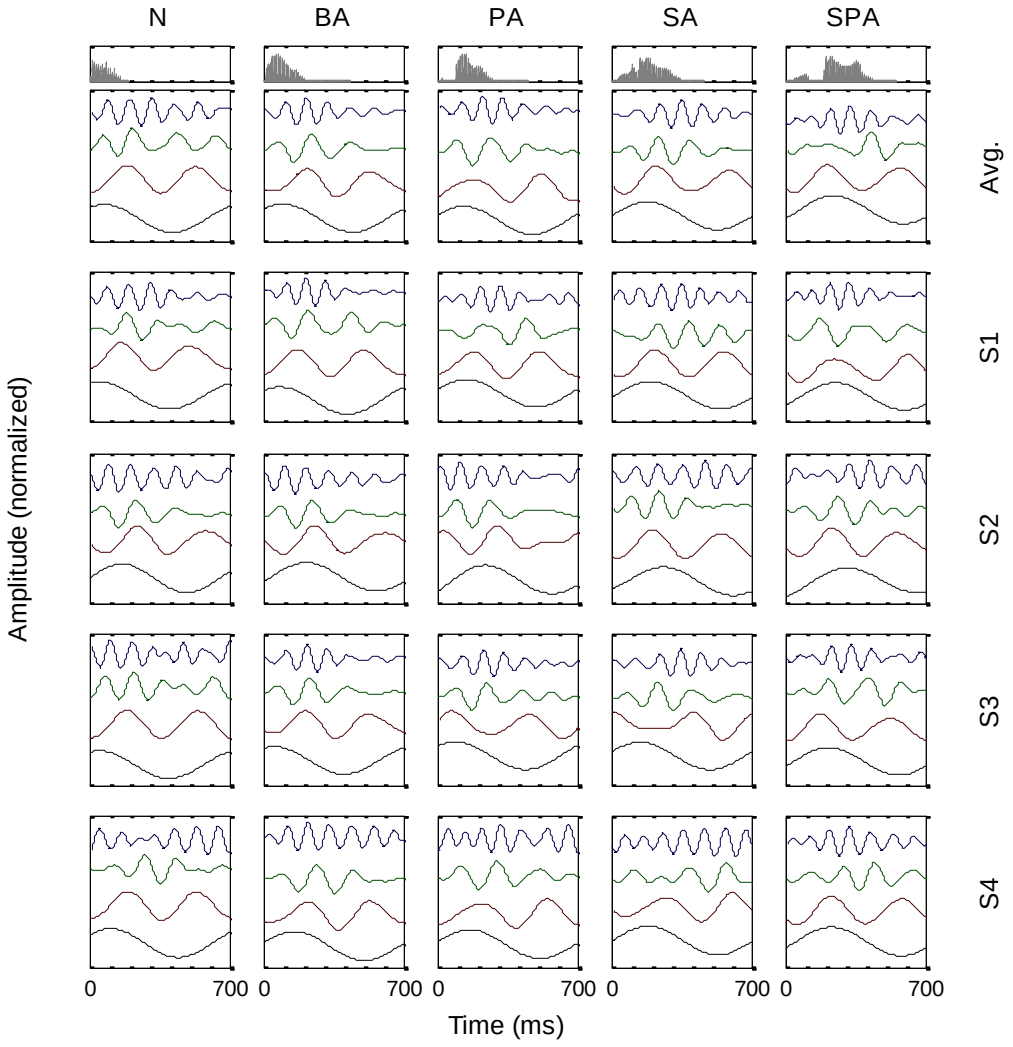


Figure 4.9 Within-participant AEP sub-bands (Cz-A2). For each stimulus (columns) the panels from top to bottom are: the stimulus amplitude (arbitrary units); the between-participants AEP sub-bands; and the within-participant AEP sub-bands for each participant. Amplitudes were normalized within each sub-band and the bands shown are the alpha, theta, upper delta (2–4 Hz), and lower delta (1–2 Hz) bands.

The instantaneous frequency of lower delta band (1–2 Hz) oscillation (calculated using the Hilbert transform) was very close to 1.42 Hz throughout the epoch as expected (this frequency corresponds to the 700 ms IOI at which stimuli were presented). The phase of this oscillation varied in a systematic manner between sounds. Although the absolute starting phase of the waveform varied between participants, the within-participant phase delay (relative to N for that participant) grew progressively larger for sounds BA, PA, SA, and SPA respectively.

Oscillations in the upper delta band (2–4 Hz) varied in instantaneous frequency over a narrow range around 2.8 Hz (the second harmonic of the presentation rate[29]). There was also some amplitude modulation but no consistent pattern.

In the theta band, the instantaneous frequency of oscillations varied more substantially both between and within individual participant and stimulus combinations. There was little evidence that oscillation at any of the presentation rate harmonics within this band (4.2, 5.6, or 7 Hz) was dominant. The instantaneous frequency also exhibited some dramatic variations (consistent perhaps with a phase reset), but again no systematic pattern was evident.

Observations are similar for the alpha. The instantaneous frequency was not related to a harmonic of the presentation rate. There was evidence of phase reset, but no consistent pattern, except perhaps that all participant responses to SPA seemed to exhibit an alpha band reset in the first 50–150 ms.

Within-participant evoked power was generally consistent with the between-participant summary shown in Figure 4.8, featuring peaks in the alpha band with latencies that apparently varied systematically for all participants except S4. This pattern was not reliably repeated in either the beta or gamma bands. The within-participant inter-trial phase coherence was very similar to the within-participant evoked power.

The baseline-relative induced power had not exhibited obvious systematic variation when summarised between-participants and no additional insight was provided by the examination of this feature within participants. Therefore this feature was excluded from further analysis.

[29] The fundamental frequency is also known as the first harmonic.

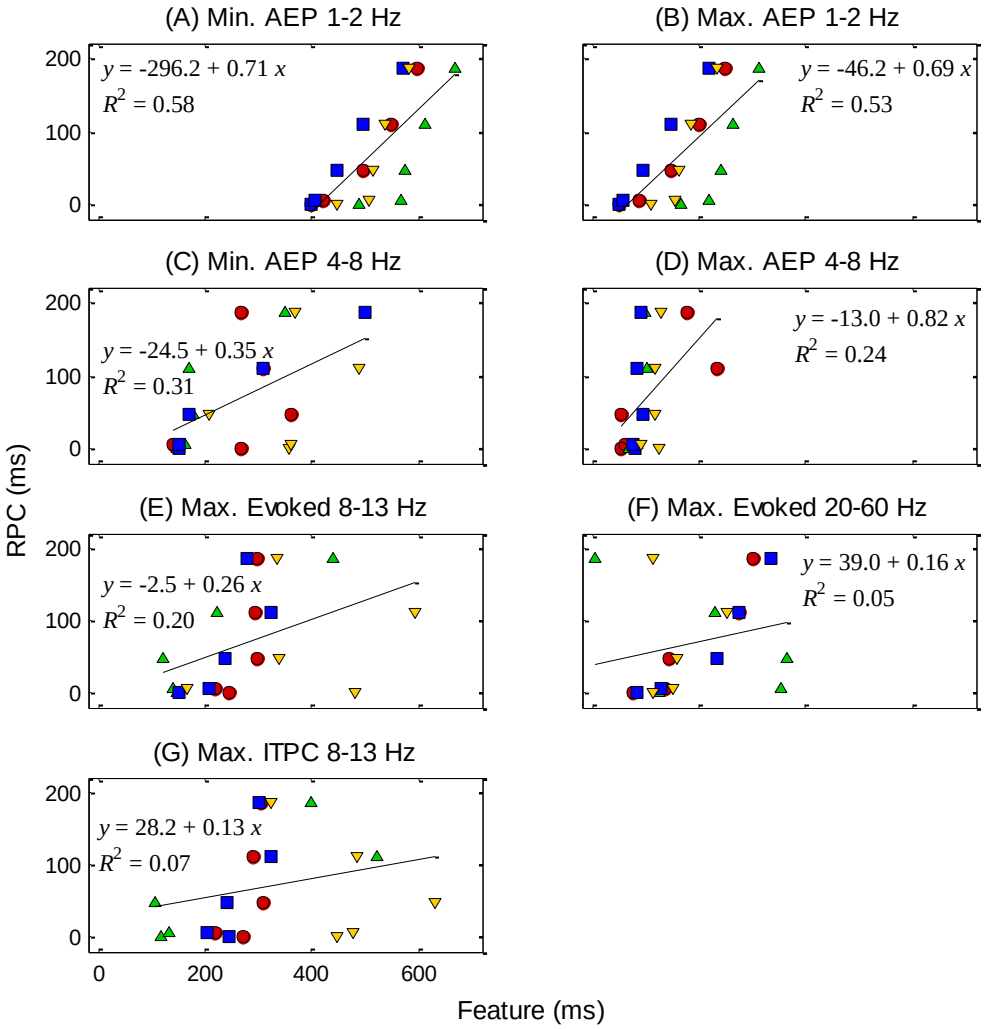


Figure 4.10 Simple regression of candidate predictors against relative P-centres (RPCs) of each stimulus. (Each colour symbol combination identifies a participant.) Candidate predictors in each panel are as follows: (A,B) latency of the lower delta band (1–2 Hz) AEP minimum and maximum; (C,D) latency of the theta band (4–8 Hz) AEP minimum and maximum; (E,F) latency of the evoked power maximum in the alpha band (8–13 Hz) and gamma band (20–60 Hz); and (G) latency of the inter-trial phase coherence peak in the alpha band (8-13 Hz). The regression fit, regression equation, and coefficient of determination ($R^2$) are shown in each panel.

In summary then, the candidate P-centre predictors to be evaluated were as follows: the latency of the maximum and minimum of the AEP in a number of bands (lower delta, upper delta, theta, and alpha), the peak evoked power latency in the alpha, beta, and gamma bands, and the peak inter-trial phase

coherence in the the alpha, beta, and gamma bands. Figure 4.10 shows the main results; predictors not included in the figure featured $R^2$ less than 0.1, or negative slopes.

The best between participant fits found were the minimum and maximum of the lower delta band and both were highly significant, $F(1,18) = 25.271$, $p < .001$, and $F(1,18) = 20.639$, $p < .001$ respectively. The regression slopes were less than 1 in both cases, indicating that the feature changed more quickly than the RPC. Individual within participant fits for these same predictors were much better ($R^2 > 0.81$). Although it was anticipated that the main participant effect would be on the regression constant, the slopes also varied between participants and were steeper than the between participant fit ($0.91 < \text{slope} < 1.65$).

Individual within-participant regression fits for predictors in the theta band of the AEP were again better than the between participant fit, but in this case slopes were very inconsistent ($0.28 < \text{slope} < 4.37$). Individual fits to the latency of maximum evoked power in the alpha band were similarly variable.

The gamma band evoked power maximum latencies did not fit well overall. However, the individual fit for two of the participants (S1 and S3, both female) was much better ($R^2 = 0.90$ and 0.88 respectively).

### *4.4.3 General discussion*

The main result of both experiments was that the latency of the minimum (or maximum) of the lowest frequency AEP sub-band predicts the variation between P-centres relatively well. In practice it is unlikely that it is specifically the signal maximum or minimum that is relevant here, rather their latency is a reasonable proxy for phase delay of the entire oscillation. In Experiment IV the delta band AEP latency varied less than the RPC. In Experiment V this relationship appeared to reverse; the mean lower delta band latency varied more than the corresponding RPCs. When individual

participant latencies were regressed, however, the slopes matched Experiment IV more closely.

The delta band latency results may be explained in several ways. First, it may be the case that the P-centre does not cause synchronously increased activity in any neural population at its moment of occurrence specifically. In this interpretation, the neural coding of the P-centre would not be amenable to direct EEG measurement. If, however, secondary activity that was related to or affected by the P-centre was the main source of delta band latency, then a non-unity slope would simply indicate the indirect measurement of the P-centre itself. A related interpretation is that the neural population involved in processing the P-centre may be much smaller than those processing other features of the sound. Again in this case, it would only be secondary components affected by the P-centre that would be measurable in the EEG.

An alternative and perhaps more realistic possibility is to consider a mix of neural populations whose EEG activity waxes and wanes in response to each stimulus. It is worth considering two principal types of activity: oscillatory activity, which is more or less regular and ongoing (though its amplitude may change), and evoked activity (either a single wave or multi-wave complex) in response to some aspect of a stimulus.

If the activity of a population is mainly oscillatory, but has a frequency outside the delta band and steady amplitude, then it will have little effect on the phase of the delta band signal. Amplitude modulation of such activity, time-locked to the stimulus, can, however, affect the delta band phase. This suggests, for example, that whereas steady theta, alpha, beta, or gamma band activity will not contribute to the observed delta band phase delay, stimulus-locked amplitude modulations in those same bands will. Such amplitude modulations were present in the AEP results and their peak modulations corresponded to the evoked power peaks. However the latencies of these peaks did not correlate well with behaviourally measured relative P-centres.

A neural population which responds just once to some temporally anchored feature of the stimulus (such as the sound onset) and which does so with approximately constant processing delay will result in EEG components which are phase locked to the stimulus. All such phase locked components will contribute to cyclic EEG activity at the stimulus presentation rate. If there are components which are time-locked to the P-centre, then the latency of these components in the AEP response to different stimuli should vary by exactly the same amount as the corresponding P-centres. In the unlikely case that all AEP components were time-locked to the P-centre, then the phase delay of EEG oscillations at the presentation rate would co-vary with the stimulus P-centre and the regression slope relating them would be 1. This is not supported by the results, however. More likely is that some AEP components were time locked to the stimulus onset (whose time of occurrence was nearly identical for all stimuli) while others were time locked to the P-centre. In this case, the phase delay of EEG oscillations at the presentation rate would vary with the P-centre, but by rather less than the time difference between P-centres would suggest. The results exhibit exactly this relationship.

Although the delta band phase delay predicts the P-centre, the explanations above all suggest that this phase delay is not the primary response to the P-centre but rather that it arises in an indirect manner. Although several possibilities were considered, the results and analyses above are inconclusive in this regard and further empirical measurement would be required. It is interesting to note that the absolute delta band phase delay varies among participants although the relative delay between stimuli is relatively consistent within each participant's results. This absolute phase variation suggests that the temporal relationship of components which contribute to the phase delay (or perhaps the relative magnitudes of those components) varies among participants.

The explanations considered for delta band phase delay make it likely that there could be an interaction between presentation rate and the ability to

resolve P-centre related phase shifts in the AEP. However, longer inter-onset intervals, or even perhaps random inter-onset intervals could provide additional insight. There are some additional methodological issues that deserve attention based on the results obtained here. The pilot results, using the Fpz rather than Cz electrode site generated slightly larger potentials with better SNR. Nevertheless, both Experiment IV and Experiment V used very sparse electrode placement, typical for clinical AEP whereas much ERP research, and particularly the relevant research of Large and colleagues, uses a much denser electrode configuration. Two additional factors which could improve on the SNR obtained in this work (allowing weaker AEP components to be resolved) would be to use diotic stimulation and to set a task to control for participant vigilance.

### *4.4.4 Conclusions*

Two separate experiments revealed that P-centre differences between stimuli were correlated with measurable changes in the AEP to those stimuli. Specifically the phase delay of very low frequency components in the delta band reliably predicted the relative P-centre. It is very likely that such low frequency phase shifts are a secondary effect of some more primary neural correlate. Nevertheless, this is first time that any neural correlate of the P-centre has been detected. Furthermore, the indication of significant neural activity associated with the P-centre (such as there must be to have a measurable effect on an AEP) lends further credibility to the fundamental importance of this percept.

# Chapter 5

# P-centre models

To be truly useful, P-centre research must ultimately yield a model that can accurately predict the perceived timing of one or more events. There are two motivations for such a model. First, an accurate, reliable P-centre model can be used to measure or control the perceptual timing of heterogeneous events without requiring constant recourse to subjective experiments. For this purpose it is not strictly necessary for the model to be psychophysiologically realistic or complete, though it is likely that an accurate model would incorporate at least some psychophysiologically inspired elements. The second motivation for a P-centre model is that a model may give insight into and aid exploration of the psychophysiological processes underlying event timing perception. For this purpose, psychophysiological plausibility may be more important than having the smallest error when compared to a particular corpus of measured P-centres (though naturally, large errors are not desirable).

Although a homogeneous sequence of events can be easily timed using the intervals between any convenient corresponding time points, it is not possible to accurately measure or control the timing of heterogeneous events (either within or between sensory modalities) unless the corresponding P-centres are known. Although this limitation is generally not noted, it has an effect on many research questions that concern timing. For example, research into sensorimotor synchronization (see Repp 2005 for a review) is generally constrained to use homogeneous (or nearly

homogeneous) event sequences in order to avoid the potential effect of P-centre differences between events. Investigations of rhythmic timing and microtiming cannot adequately measure performances in which the P-centres of events in a sequence can vary substantially relative to each other. In particular, without knowledge of P-centres the rhythm of spoken language cannot be measured accurately and thus questions about the perceived timing of individual languages can be answered only on the basis of flawed or indirect data at best. A researcher who needs to prepare event sequences with specific perceptual timing for use in an experiment cannot use heterogeneous events if the event P-centres are not known. Indeed, the P-centre term originated when Morton et al. (1976) discovered that they could not easily construct a perceptually regular sequence of recorded words for a memory experiment. Moving beyond the domain of the research laboratory, a P-centre model has a key role in achieving expressive performance with speech and music synthesis and other temporally sensitive activities. Indeed it may well have a part to play in achieving natural interaction and gesture timing for anthropomorphic robot and virtual human models (for a suggestive example, see Murata et al. 2008).

## 5.1 Existing models

An acoustic P-centre model is just one aspect of a more general P-centre model that can be applied to events in any modality (or perhaps, just one of a family of models). However, there do not appear to be any studies evaluating the relationship of non-acoustic event features to the P-centre. Furthermore, there are substantial challenges that must be overcome to realize even an acoustic model in a comprehensive and reliable manner.

For separated, non-overlapping events, there is no a priori reason to assume that the P-centre is not located at the perceived event onset, that is, at the moment of event detection. In particular, musical notation encourages exactly this assumption: Rhythm is assumed to be specified by the timing of note onsets and not their durations or offsets (Rasch 1979).

However, Morton et al. (1976) failed to construct perceptually regular sequences of recorded words for their memory experiment when they made the word onsets isochronous; clearly the P-centre is not coincident with the onset of a word or syllable. (Although they did not specify it, it seems that Morton et al. used onset to mean the objective or physical onset rather than the perceptual onset. However, their Figure 1 does not demonstrate any alignment by a common threshold, a feature that would be expected if perceptual regularity resulted from perceptual onset isochrony.) Gordon (1987) similarly found that neither a simple absolute or relative onset threshold could accurately predict the P-centre of all the musical tones he had empirically measured. In fact, the P-centre of acoustic and speech events does not appear to reliably correspond to any obvious acoustic or speech specific feature. Numerous candidate features have been considered but shown to fail in at least some cases; these include local or global intensity peaks (Gordon 1987; Marcus 1981), the measured vowel onset (Marcus 1981), the number of initial consonants (Cooper, Whalen & Fowler 1986), and the vowel quality (Fox & Lehiste 1987b). For continuous stimuli, which may result in imprecise and overlapping event boundaries, the interaction between events in the vicinity of their onsets and offsets would also seem to argue against the P-centre corresponding to a single simple onset-related feature.

Nevertheless, though it may not correspond to a single simple feature, most auditory P-centre studies suggest that the P-centre is located in the vicinity of a sound's onset (for example Gordon 1987; Scott 1998; Vos, J. & Rasch 1981) or, for syllables, the syllable onset to nucleus transition (for example Allen 1972b; Cooper, Whalen & Fowler 1986; Fowler 1979; Janker 1996a). A number of P-centre models have been proposed (Gordon 1987; Harsin 1997; Marcus 1981; Pompino-Marschall 1989; Schütte 1978; Scott 1993; Vos, J. & Rasch 1981). Although the models vary both in general approach and specific details, in each case the model developer has reported results indicating that the model predicts behaviourally measured P-centres with

little error. Unfortunately these results have not generally been replicated or independently verified.

A problem that is apparent from the literature is that each P-centre model has been developed and tested or trained with a relatively sparse corpus of sounds with P-centres measured by the researcher who developed the model. Surveying the models, all appear to have been tested with P-centres measured using the rhythm adjustment method, but with a number of detailed differences including presentation conditions, cycle duration and rhythm, and sequence length. It is not known whether these differences are significant. Furthermore some models have been tested only (or at least mainly) with speech sounds (for example, Harsin 1997; Marcus 1981; Scott 1993) while others have been tested exclusively with non-speech sounds (for example, Gordon 1987; Schütte 1978; Vos, J. & Rasch 1981). Taken together these issues make it difficult to know whether the results obtained by any individual researcher on any single test corpus can be generalized.

### *5.1.1 Overview of models*

In general there is little indication that any of the individual P-centre models was developed by evolving or refining those models which preceded it. In particular there does not appear to have been any significant analysis of prior models in order to determine which sound types were problematic and therefore how those specific sound types should be addressed[30]. Despite the lack of clear model lineage, certain patterns and recurring ideas can be discerned.

Acoustic P-centre models can be divided into two broad categories: *onset models* which make use of local onset features only, and *global models* which predict the P-centre using some integration of global features of the

[30] Several researchers do test a small subset of prior models for comparison purposes when testing their proposed model. However the performance of prior models with various sound types appears to have been examined only after the proposed model was developed.

sound. The onset models are those of Rapp (as described by Marcus 1981), Gordon (1987), and Scott (1993). While the specific definition of what constitutes an onset varies, the common feature of onset models is that their P-centre predictions cannot be affected by sound features which occur after the onset. In particular onset models are unaffected by secondary onsets within the sound, by the nature of the sound offset, or indeed by the duration of the sound. Most onset models are primarily threshold detectors and thus insensitive to supra-threshold variation. Additionally onset models tend to be simpler than global models and focus on amplitude changes either within the whole signal or some narrower sub-band; in both cases the model will tend to be fairly insensitive to changes in pitch, timbre, or frequency. Perhaps the strongest argument in favour of the onset model approach is that all the information necessary to determine the P-centre is available before the sound has ended. This seems to reflect subjective experience: the rhythmic beat of a musical note may be felt even while the note is sustained.

In contrast, global models are affected by sound features which occur after the onset, though such features may be attributed less importance than those which occur during the onset. The global models are those of Marcus (1981), Howell (1984; 1988), Pompino-Marschall (1989; 1990), and Harsin (1997). The model of Marcus is based on speech specific notions such as the time of vowel onset, which would not appear to be the most promising approach for a general acoustic P-centre model. Howell described a modelling approach—calculating the P-centre as the centre of gravity of some features of the whole sound—rather than a specific model. This centre of gravity notion was subsequently adopted by Pompino-Marschall and Harsin who both used partial events as the elementary sound features to be integrated (though their identification and weighting of partial events differed). However, it is not clear that the criteria chosen to identify partial events are perceptually salient. The most significant argument against the global models is the implausibility of being unable to identify the P-centre until after the sound has ended.

A problem encountered with almost all models was that the descriptions are incomplete or inconsistent on certain details. It is not possible to implement the models without making certain assumptions and as a consequence it is not possible to review the models in detail without grounding the review in a specific implementation which includes those assumptions. For that reason, the following subsections serve both as review of the individual models and as detailed description of the implementations used in this work. (The commented MATLAB code implementing each of the models is listed in Appendix C.) The models are reviewed in chronological order of first publication.

### *5.1.2 Marcus (and Rapp-Holmgren)*

Marcus's model is a global model formulated in terms of speech specific features and tested only with speech sounds. No doubt influenced by the test corpus which primarily comprised CV and CVC syllables, the model predicts P-centres based on two durations: the time between acoustic (syllable) onset and vowel onset; and the time from vowel onset to acoustic (syllable) offset. As the model only uses timing features it is insensitive to (possibly large) differences between sounds which do not affect the point of onset, offset, or vowel onset.

Marcus also described a variant of the model which he attributed to Rapp-Holmgren (1971) and which differed from his model only in the specific parameter values used. Therefore a single implementation can generate the predictions of both Marcus's and Rapp-Holmgren's models.

In this work, the model implementation details were as follows:

1. Marcus originally fitted his model to sound data sampled at 20 kHz. For this reason, the signal sample rate is first resampled to this rate, if necessary. This resampling excludes higher frequency components against which the model had never been tested.

2. Next, the signal is divided into frames. Both the duration and inter-frame interval are 10 ms (200 samples). A single-sided power spectral density estimate is obtained for each frame using a 512 point FFT with a rectangular window.

3. All frames whose power (summed from the power spectral density) exceeds a threshold are considered audible and the signal onset and offset times ($t_1$ and $t_2$) are identified as the midpoints of the first and last of these frames respectively. The threshold was chosen to be a relative level 30 dB below the signal maximum in this implementation. Marcus did not explicitly specify the nature of this threshold or its level, describing it only as a "fixed criterion". If the rate of onset or offset is particularly slow then a different threshold might change the detected onset and offset time enough to alter the P-centre prediction significantly. Nevertheless 30 dB seems to be a reasonable relative threshold level.

4. In each frame, summing the power spectral density across FFT bins from 500–1500 Hz yields the mid band power. The vowel onset is indicated by the most rapid increase in this mid band power. Again, there are two ways of calculating this: the *absolute increase* is the difference in linear power between consecutive frames, whereas the *relative increase* is the difference in dB (log) power between consecutive frames. Early testing indicated that the largest absolute and relative increase did not always co-occur, and this discrepancy can affect the model predictions. In this implementation, the vowel onset time ($t_V$) was taken to be the midpoint of the frame exhibiting the largest relative power increase.

5. The general form of the model has two parameters, α and β. For Marcus's model, the fixed compromise values are used, .65 and .25 respectively. For the Rapp-Holmgren model the values are .50 and

0 so that the model degenerates to a one parameter model. Finally the P-centre (*PC*) is calculated according to equation 5.6 with times expressed in milliseconds. (The constant *k* is unknown but cancels when relative P-centres are calculated.)

$$PC = \alpha\left(t_V - t_1\right) + \beta\left(t_2 - t_V\right) + k \tag{5.6}$$

Figure 5.1 shows the main processing stages and key elements of Marcus's model. It should be apparent that a sound whose main energy (and energy changes) lies outside the frequency limits of the Marcus's mid band is likely to cause problems for this model. Furthermore, the specific location of the vowel onset may also be sensitive to minor fluctuations in mid-band power

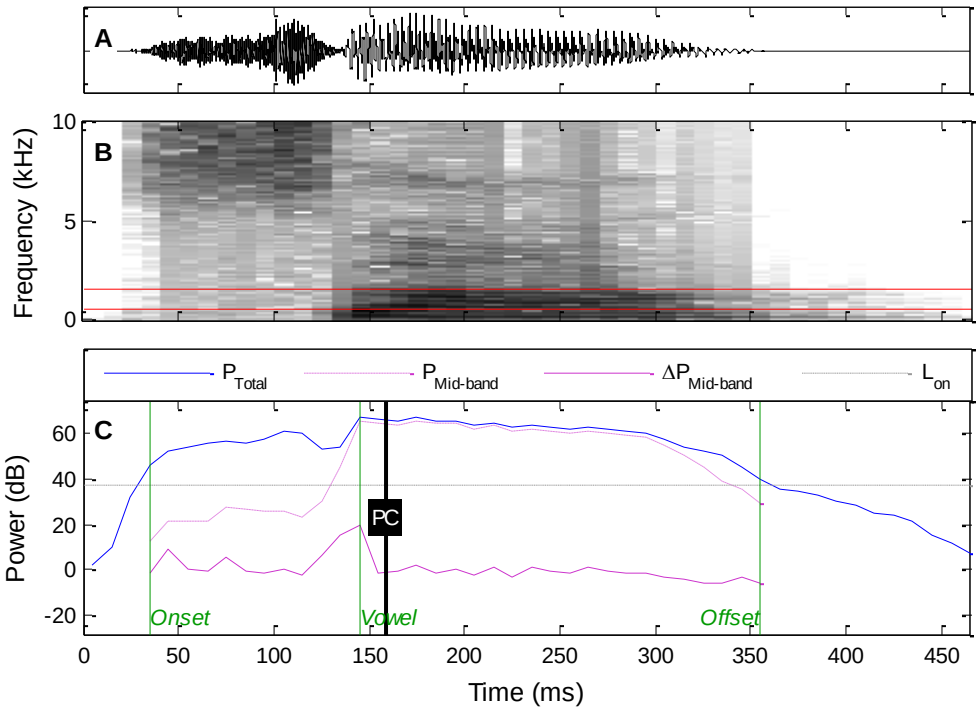


Figure 5.1 The model of Marcus applied to the sound /sa/. (A) the sound waveform; (B) the spectrogram (power in dB, mid band frequencies between the solid lines); and (C) the main processed signals and time points of the model. The key time points are the onset, offset, vowel onset, and predicted P-centre (PC). These time points are derived from the total power ($P_{Total}$), mid band power ($P_{Mid\text{-}band}$), the relative change in power ($\Delta P_{Mid\text{-}band}$), and the perception threshold level ($L_{on}$), all measured in dB. (The unknown constant, *k*, in equation 5.6 is assumed to be zero for the purpose of indicating a P-centre location.)

in the case where the mid-band power is nearly constant throughout the signal. It is also worth noting that this model was fitted only to wideband natural speech so its applicability to other sounds (for example musical sounds) is unknown. Finally, the use of simple threshold detection for onsets and offsets makes the model sensitive to background noise and recording imperfections.

### *5.1.3 Vos and Rasch*

The Vos and Rasch model is an onset model operating as a relatively simple threshold detector. The main distinguishing feature of the model is that it uses a relative threshold that depends on the signal level. Furthermore, the model was designed to fit P-centres obtained with simple envelope shaped sawtooth tones so its applicability to more complex sounds including speech is unknown.

The operation of Vos and Rasch's model is as follows:

1. As the model depends on the sensation level of the sound above a masked or absolute threshold, this sensation level must first be determined. If not specified for a particular sound, the sensation level ($L_{\mathrm{SL}}$) is estimated as the peak RMS level (dB, exponentially averaged with 125 ms time constant), less the masker level (dB). The masker level is assumed to be 0 dB if not specified.

2. Next the signal envelope is estimated. Vos and Rasch developed their model with sawtooth tones whose envelope was known whereas for a general model, the envelope must be estimated for each sound. In this implementation, the envelope was estimated by applying a low pass filter (100 Hz, Butterworth, order 2) to the full wave rectified amplitude.

3. The P-centre threshold ($L_{\mathrm{PC}}$) is established relative to the peak level. Vos and Rasch did not specify whether the threshold was

relative to the sensation level ($L_{SL}$, which they estimated from long duration continuous tones rather than short stimuli) or the peak envelope level ($L_{Peak}$) of the signal. In this implementation the latter was used. Vos and Rasch's results indicated that the relative threshold should range approximately 7–15 dB below maximum for sensation levels from 20–70 dB. A linear regression fit to the exact results yielded the following expression for relative threshold:

$$L_{PC} = L_{Peak} - 3.18 - 0.17\,L_{SL}, \quad L_{SL} > 20 \tag{5.7}$$

4. Finally, the predicted P-centre is the moment at which the envelope first exceeds the P-centre threshold, $L_{PC}$.

Figure 5.2 illustrates the main elements of the Vos and Rasch model, namely the envelope and thresholds used. Because the model makes no attempt to

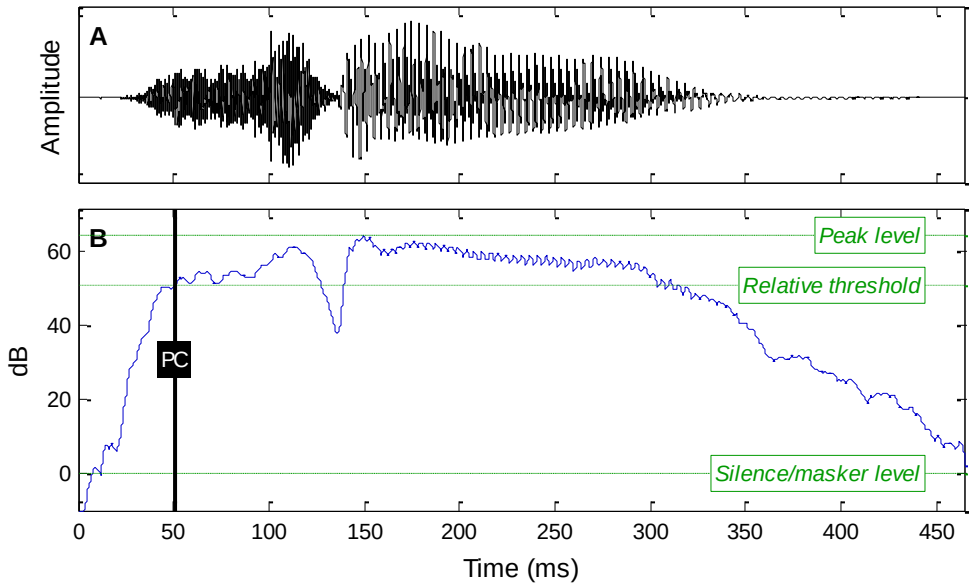


Figure 5.2 The Vos and Rasch model applied to the sound /sa/. (A) The sound waveform; and (B) the sound envelope. The relative threshold (with respect to the peak level) is based on the difference between the maximum signal level and the silence or masker level. The P-centre is the moment at which the envelope exceeds the relative threshold.

incorporate psychoacoustically realistic features such as equal loudness weighted frequency response, the estimated envelope may not approximate the perceived envelope well for sounds with significant high or low frequency energy (cf. Figure 5.1 where the difference between the total and mid band power shows the obvious effect of excluding or attenuating high frequency components).

### *5.1.4 Gordon*

Gordon implemented and evaluated a variety of models applied to amplitude, power, and loudness envelopes. Rather than re-evaluate each of these models, only the best performing model in Gordon's tests, *normalized with rise*, was tested.

This model operates as follows:

1. First the amplitude envelope of the signal is estimated. In Gordon's original implementation, which used tones with a fixed fundamental frequency, this envelope was obtained by interpolating between fundamental period waveform peaks. Such an approach cannot be reliably applied to general sounds, so this implementation instead applied a low pass filter (100 Hz, Butterworth, order 2) to the full wave rectified amplitude. The resulting envelope was then resampled with a sample period of 1 ms.

2. Next the amplitude envelope is converted to a power (intensity) envelope by squaring it. This envelope is then normalized to its maximum.

3. The slope of the normalized envelope is then calculated. Using Gordon's method a line is repeatedly fitted to a 19 sample (19 ms) window of data, advanced in steps of 1 ms throughout the

envelope. The slope of this line provides the envelope slope estimate at the centre point of the data window.

4. The time points which delimit the rise time of the envelope are then identified. Gordon defines the rise time as the duration over which the slope of the normalized envelope exceeds the slope threshold ($0.36 \times 10^{-3}$). This duration is delimited by the rise time beginning ($t_1$) and the rise time end ($t_2$).

5. Finally the P-centre is calculated according to the following equation:

$$PC = t_1 + 0.08(t_1 - t_2) \tag{5.8}$$

Although Gordon's modelling data was based on sounds presented at approximately 90 dB(A), normalization within the model (Step 2 above) allows Gordon's parameter values to be applied regardless of sound level. Nevertheless, Vos and Rasch (1981) found that their threshold parameter varied with presentation level and it is possible that the same would be true for Gordon's parameters. If the level dependence was significant, then Gordon's model could be expected to predict less well than others the P-centres of sounds presented at typical speech levels (60–70 dB SPL).

Figure 5.3 illustrates the main features of Gordon's model. The most striking feature is the dramatic underestimation of the rise time apparent in the signal envelope. Naturally, any underestimation of rise time would affect the P-centre prediction. Unlike the instrumental tones used by Gordon, the onset of the natural speech sound shown is not a monotonic rising function; the slope is both positive and negative at various times in the sound onset. Gordon proposed a modification to his model to handle special cases where the slope crossed the threshold twice, but this modification requires a somewhat arbitrary weighting factor and does not handle more than one threshold crossing. (As a consequence, the modification was not applied to the model implementation in this study.) In

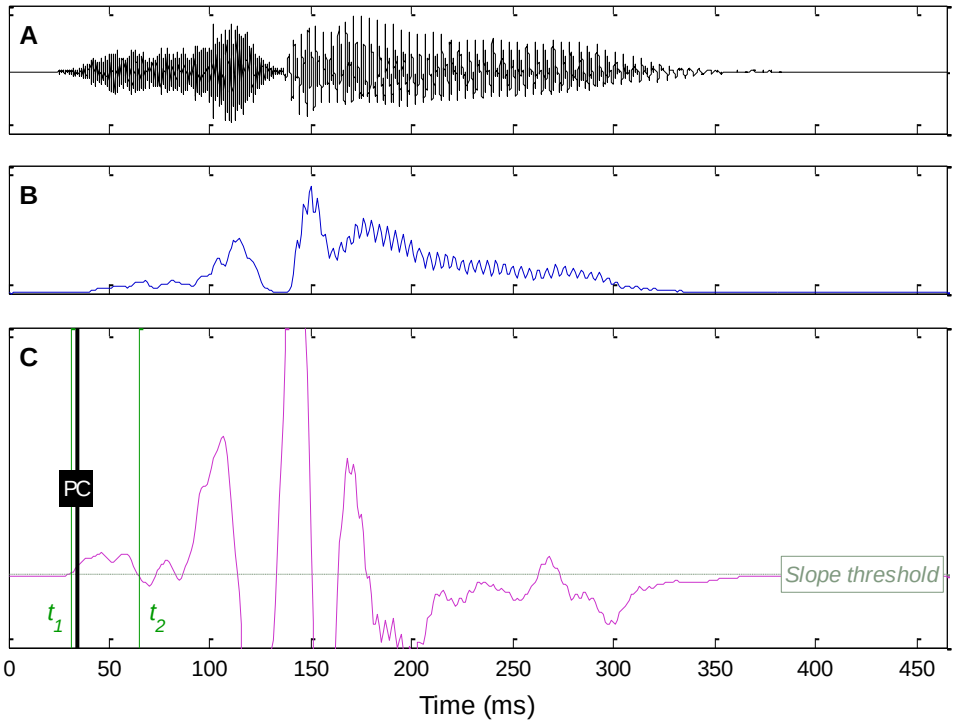


Figure 5.3 Gordon's normalized with rise model applied to the sound /sa/. (A) The sound waveform; (B) the power envelope obtained by applying a low pass filter (Butterworth, order 2, 100 Hz cutoff) to the squared amplitude; and (C) the envelope slope (smoothed with a 20 ms moving window). The P-centre (PC) is the rise time beginning ($t_1$) delayed by a fraction of the rise time ($t_2$ - $t_1$), the duration over which the envelope slope exceeds the slope threshold.

practice it is not clear how the model should be modified to handle sounds with non-monotonic rise functions: should the beginning of the rise time occur only where a monotonic rise exists (for example, around 130 ms in Figure 5.3)? Alternatively, should the beginning stay where it is and the end of rise occur only when the slope falls below threshold for the last time before the envelope maximum? Without attempting to fit data, these questions cannot be answered, but it does seem likely that Gordon's model will not yield good P-centre predictions for sounds with complex onsets.

### *5.1.5 Howell*

The Howell model is a global model, but as noted previously, Howell proposed a general model approach and did not describe a specific model

implementation or its parameters. Scott, however, did implement and evaluate a model she referred to as Howell's model and it was this implementation that was used as a basis for the Howell model in this work.

The operation and implementation of Howell's model is detailed in the following steps:

1. First the envelope is estimated so that the perceptual onset and offset can be determined. This step, omitted in Scott's implementation, is necessary if the sound signal incorporates any preceding or succeeding "silent" portions. The envelope is obtained by full wave rectification followed by low pass filtering (25 Hz, Butterworth, order 2). Based on the approach used in Marcus's model, the onset and offset are identified as the time at which envelope exceeds a threshold for the first and last time respectively. In this implementation, a relative threshold 30 dB below envelope maximum was used.

2. Next, the "weight" signal is generated. Based on Scott's implementation but modified in line with Howell's own description (1984; 1988), this was generated by full wave rectification of the input signal followed by low pass filtering (25 Hz, Butterworth, order 2). Though this is the same as the envelope calculation used in step 1 this is a coincidence and the two processing stages are independent of one another.

3. Finally the P-centre (PC) is estimated using the usual centre of gravity calculation, interpreting the weights ($w_i$) as the values of the weight signal and using times ($t_i$) instead of distances as shown in equation 5.9. Scott's implementation did not in fact calculate the centre of gravity but instead calculated the weight midpoint (where exactly half the weight lies either side of the midpoint). A similar calculation was used by Fowler et al. (1988). The weight midpoint is not the same as the centre of gravity,

however, because it ignores the effect of distance (time). The centre of gravity seemed closer in intent to Howell's original descriptions (particularly Howell 1988) and thus it was the centre of gravity calculation that was used.

$$PC = \frac{\sum w_i t_i}{\sum w_i} \tag{5.9}$$

Figure 5.4 shows the main elements of Howell's model. It is clear that this model implementation is simplistic and may not produce good P-centre predictions. In particular, the model may be too sensitive to the distribution of energy in time, so that a sound which gradually gets louder would tend to have a very late P-centre.

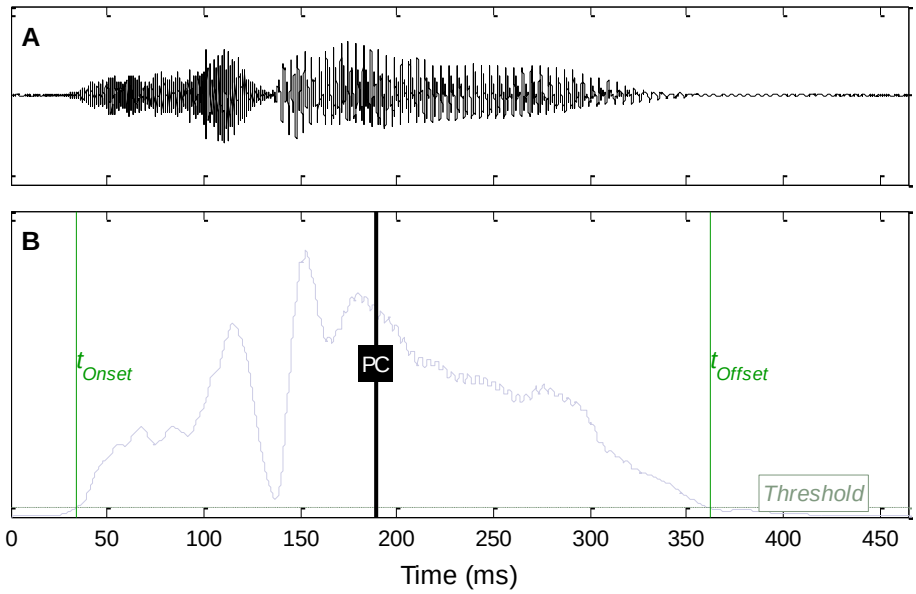


Figure 5.4 The Howell model applied to the sounds /sa/. (A) The signal waveform; and (B) the amplitude envelope including the relevant time points and threshold. The P-centre is centre of energy between the onset and offset.

### *5.1.6 Pompino-Marschall*

Pompino-Marschall's model is a global model which incorporates psychoacoustically plausible loudness and amplitude modulation sensitivity

processing. Despite the availability of a Fortran code listing, Pompino-Marschall's model was undoubtedly the most complex model to implement. It is also the most complex model in operation as illustrated by the following description:

1. First the signal waveform is resampled if necessary. The sample frequency used by Pompino-Marschall was 20 kHz (Pompino-Marschall 2007).

2. Next, a time-frequency representation of the signal is generated using a multi-resolution short term Fourier Transform (STFT) analysis. So that spectral energy would grow smoothly from zero, 60 ms of zeros are first prepended to data. Analysis frames are then extracted from the signal every 15 ms (so that the frame rate is 66.67 Hz). Each frame is shaped with three different duration Hanning windows: 60, 30 and 15 ms. Thereafter, each window is transformed using the DFT to yield spectra with different effective frequency resolutions (16.67, 33.33, and 66.67 Hz from window lengths of 1200, 600, and 300 points respectively). A multi-resolution power spectrum is obtained by combining different frequency bands from each DFT (16.7–500 Hz, 533–1500 Hz, and 1533–5267 Hz at the finest, medium, and coarsest frequency resolutions respectively). In each analysis frame, windows are aligned by their first sample and not their temporal centre (Pompino-Marschall 1990, pp. 207-10; 2007). As a consequence low frequency components of the multi-resolution STFT which use a long time window appear 45 ms earlier than high frequency components and 30 ms earlier than mid frequencies.

3. From the multi-resolution spectrum for each frame, estimates of the mean power spectral density in critical bands are derived. There are 19 critical bands, with Bark scale centre frequencies of 1–19 Bark and a bandwidth of 1 Bark each (refer to the code in Appendix C for specific centre and edge frequencies.) The mean

power spectral density is calculated as the mean of the power spectral bins whose frequencies fall within the 3 dB bandwidth of the critical band.

4. Within each critical band the power envelope is filtered to model temporal integration and masking effects. First each envelope ($x_3$) is linearly filtered with a first order filter (equation 5.10) to provide a subtle 0.05 dB gain to low frequency modulations, and an equivalent attenuation to high frequency modulations (the crossover between gain and attenuation occurrs at 16.7 Hz). The resulting envelope ($x_4$) is filtered again, but in this case decreasing and non-decreasing regions are filtered differently (equation 5.11). Non-decreasing envelope regions are filtered with a first order low pass filter (-3dB at 30 Hz) whereas decreasing envelope regions are filtered with a non-linear first order low pass filter.

$$x_4(n) = x_3(n) + 0.0067\, x_3(n-1) \tag{5.10}$$

$$x_5(n) = \begin{cases} 0.15\, x_4(n) + 0.85\, x_4(n-1), & x_4(n) \geq x_4(n-1) \\ x_5(n-1)\exp\left[0.21 \log\left(\dfrac{x_4(n)}{x_5(n-1)}\right)\right], & x_4(n) < x_4(n-1) \end{cases} \tag{5.11}$$

5. The filtered power envelope is converted to dB; then the specific loudness in each critical band is estimated and smoothed using the loudness calculation of Paulus and Zwicker (1972)—their Fortran code was translated into a functionally equivalent MATLAB implementation (see Appendix C). The loudness is calculated assuming a free field response and a 0.2 Bark frequency sampling which is subsequently averaged within each critical band (Pompino-Marschall 1990, p. 211).

6. At this stage, partial onset and offset events are identified within each critical band. For each critical band, *i*, and partial event, *j*, the measures to be used by the model are evaluated, namely the time ($t_{ij}$) and specific loudness difference ($\Delta L_{ij}$) associated with the

event. This loudness difference is always measured relative to the endpoint of the last detected partial event (or zero if none). Thus the difference is positive for onset events and negative for offset events. A partial event is detected whenever the loudness difference exceeds 12% of the maximum loudness and the partial event endpoint is the next local maximum or minimum (for onset and offset events respectively). The partial event's time is associated not with the start point, but with the (linearly interpolated) moment at which the loudness crosses a relative threshold set at 40% of the specific loudness increase or decrease as appropriate.

7. A sequence of contiguous partial onsets defines a rising flank and correspondingly, a sequence of contiguous partial offsets defines a falling flank. Between the rising and falling flanks lies a peak. A sound may have more than one peak, for example, a short speech syllable can have one peak associated with a consonant and a second associated with a vowel. Therefore, partial events on the rising and falling flanks surrounding each peak are first weighted in preparation for subsequent integration. The weight for each partial onset is calculated according to the time difference between it and the peak onset ($t_{\mathrm{peakon}}$), the onset just before the peak (equation 5.12). Onsets occurring early on the rising flank are attributed less weight than those occurring later. The weight for partial offset events is calculated similarly. In this case, there is an additional scaling by 0.5 to signify that onsets are more perceptually salient and the weight is calculated using the time difference between each partial offset and its corresponding peak offset ($t_{\mathrm{peakoff}}$), the first offset after the peak (equation 5.13). Offsets occurring late on the falling flank are attributed less weight than those occurring early.

$$w_{ij} = \Delta L_{ij} \exp\left[ -\frac{\left( t_{\text{peakon}} - t_{ij} \right)}{\tau} \right], \quad \Delta L_{ij} > 0 \tag{5.12}$$

$$w_{ij} = 0.5 \Delta L_{ij} \exp\left[ -\frac{\left( t_{ij} - t_{\text{peakoff}} \right)}{\tau} \right], \quad \Delta L_{ij} < 0 \tag{5.13}$$

8. On each rising flank, partial onsets are integrated to form a single peak onset event. Similarly on each falling flank, partial offsets are integrated to form a single peak offset event. Finally, matched peak onset and offset events are integrated to form peak events. The time of integrated events is calculated using the normal centre of gravity calculation (equation 5.14). Because weights can be negative for partial offsets, the absolute value of the weight is used. The calculation of integrated event weight used by Pompino-Marschall is unusual however (equation 5.15). Normally the integrated weight associated with a centre of gravity is simply the sum of the weights. In this case, the integrated weight calculation has the effect scaling the integrated weight according to how closely in time the constituent weights occur. Furthermore the calculation is not time invariant: the same temporal distribution shifted by a constant offset will result in a different integrated weight. Nevertheless, this was the calculation used by Pompino-Marschall (1990, p. 218) and is therefore the calculation used in this implementation of the model.

$$t_{ik}^{\text{Integrated}} = \frac{\sum \left| w_{ij} \right| t_{ij}}{\sum \left| w_{ij} \right|} \tag{5.14}$$

$$w_{ik}^{\text{Integrated}} = \frac{\sum \left| w_{ij} \right| t_{ij}}{\sum t_{ij}} \tag{5.15}$$

9. Subsequently all peak onset events in each critical band are integrated to form a single critical band onset event. Similarly all peak events in each critical band are integrated to form a single

critical band peak event. Finally all critical band onset events are integrated to form the *syllable onset* and all critical band peak events are integrated to form the *syllable centre of gravity*. In all cases the integration calculation is identical to before. Eventually, the P-centre can be estimated as the time of the syllable centre of gravity.

The criteria for identifying partial events and the subsequent calculations to weight and integrate these partial events seem somewhat arbitrary. In particular the main publication of the model (Pompino-Marschall 1989) does not provide much explanation of the chosen values and calculations. Furthermore, Pompino-Marschall notes that the various scaling factors and integration factors have yet to be determined experimentally.

Figure 5.5 shows the main stages of processing in the operation of Pompino-Marschall's model. It is obvious from Figure 5.5 (B) that high frequency energy is ignored, even if substantial. In a sound dominated by high frequency content (typically sibilant or noise-like sounds), it seems likely that the model predictions may not reflect subjective experience. In the same figure panel the time advancement of low frequencies relative to high frequencies is also clearly visible and it appears that this could distort the P-centre calculation except in cases where there is little or no low frequency energy. Figure 5.5 (E–G) shows that integrated onset and peak events appear to be quite insensitive to offsets, and it could be questioned whether the complexity of identifying and integrating offset events is warranted.

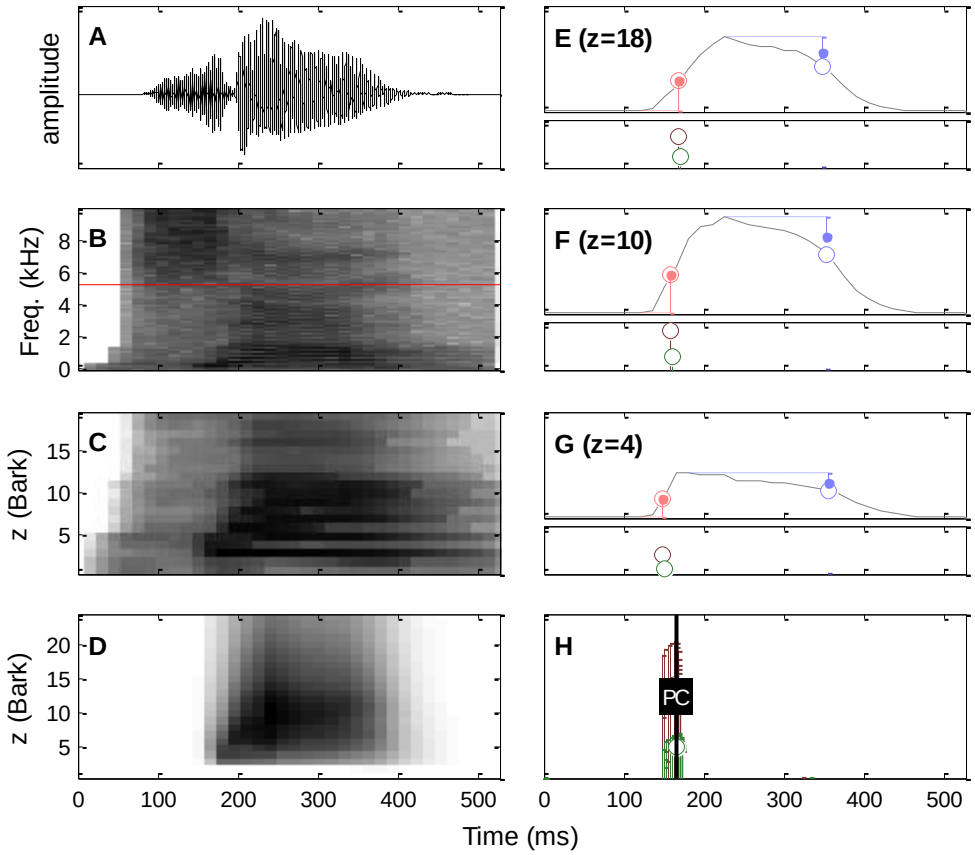


Figure 5.5 Pompino-Marschall's model applied to the sound /sa/. (A) The sound waveform; (B) The multi-resolution spectrogram (power in dB; frequencies above the solid line play no part in subsequent processing); (C) power (dB) in critical bands after within-band linear and log-linear envelope filtering; (D) the estimated specific loudness (sones) within each critical band; (E–G, upper) the loudness envelope, partial onsets and partial offsets in three example critical bands (18, 10 and 4 Bark); (E–G, lower) the integrated peak onset and peak events in those same critical bands; and (H) the integrated channel onset and channel peak events from all bands (lines), and (lines with open circles) the syllable onset, and the syllable centre of gravity. In this case the syllable onset and syllable centre of gravity are almost simultaneous and overlap on the figure; the P-centre is the time of the syllable centre of gravity.

### *5.1.7 Scott*

Scott's *Frequency dependent Amplitude Increase Model* is an onset model that operates essentially as a threshold detector. Like the models of Vos and Rasch and Gordon the threshold is relative to the signal maximum, but in

contrast Scott applies the threshold to a sub-band rather than the entire signal.

Scott's model is straightforward to implement and operates as follows:

1. First the signal envelope is estimated so that the perceptual onset can be determined. This step was not included in Scott's description but is necessary to prevent distortion of the P-centre estimate by an initial "silent" segment in the sound waveform. The envelope is obtained by full wave rectification followed by low pass filtering (25 Hz, Butterworth, order 2). Based on the approach used in Marcus's model, the onset is identified as the time ($t_{\mathrm{Onset}}$) at which the envelope exceeds a threshold (which in this implementation defaults to a relative threshold, 30 dB below envelope maximum).

2. Next the signal is band pass filtered with a Gammatone style filter[31] (578 Hz, 4 ERB bandwidth) to yield a single sub-band.

3. The envelope of the sub-band is estimated by applying the approach of step 1 to the sub-band signal. Then the time ($t_{\mathrm{SubAmp}}$) at which a relative threshold (half the sub-band envelope maximum or about 6 dB below the peak envelope level) is crossed is identified.

4. Finally the P-centre is calculated according to the following equation which incorporates both Scott's regression fit parameters and a correction for the (possibly delayed) signal onset time.

$$PC = t_{\mathrm{Onset}} - 11.2 + 0.407\left( t_{\mathrm{SubAmp}} - t_{\mathrm{Onset}} \right) \tag{5.16}$$

[31] The implementation described by Slaney (1998) was slightly modified to allow a non-standard bandwidth to be specified for the gammatone filter.

As specified above Scott's relative threshold is about 6 dB below maximum level. Though Scott's threshold is applied to just a sub-band of the signal, it is interesting to note that her threshold level is very close to the threshold set by Vos and Rasch for sounds just 20 dB above background masker level. Like Marcus's model, Scott's model also operates primarily on a sub-band of the signal. However her sub-band (nominally 420–731 Hz) is somewhat lower than Marcus's (500–1500 Hz) and will tend to be dominated by first formant energy in speech, rather than first and second formant energy in the case of Marcus's band. The use of Gammatone style filter with non-standard bandwidth is curious, but may have a significant effect on the model behaviour as, for example, a second order Butterworth band pass filter with the same cut-off frequencies would result pass less low frequency energy and more high frequency energy.

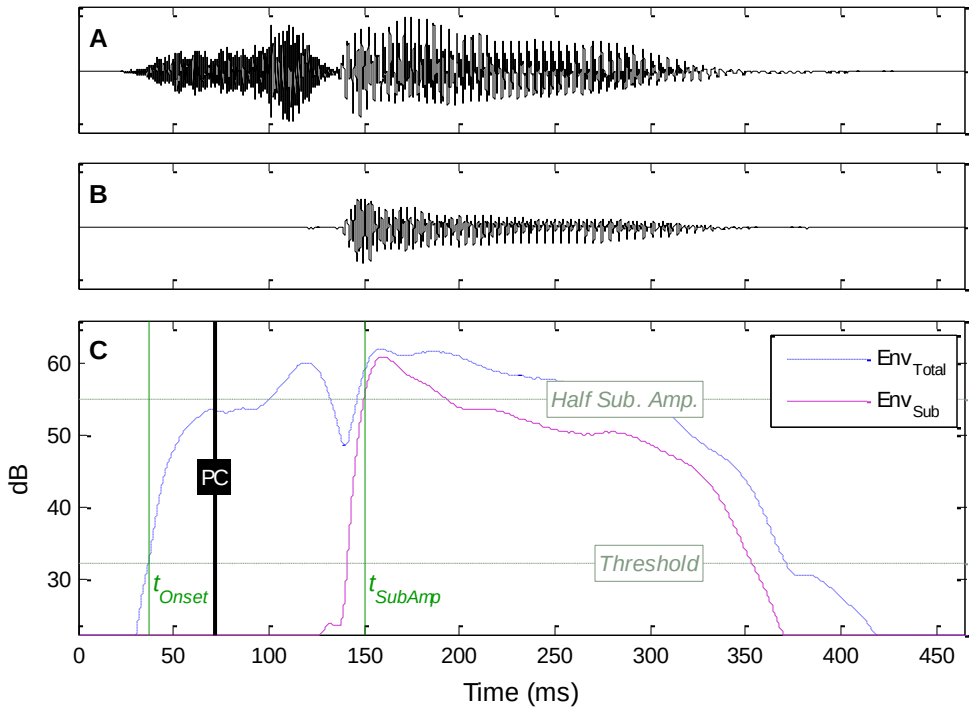


Figure 5.6 Scott's Frequency dependent Amplitude Increase Model applied to the sound /sa/. (A) The sound waveform; (B) the sub-band waveform obtained by bandpass filtering (Gammatone filter, 4 ERB bandwidth, centred at 578 Hz); and (C) the signal and sub-band envelopes, with relevant thresholds and time points indicated.

The key processing stages of Scott's model are shown in Figure 5.6. It is clear in Figure 5.6 (B and C) that the high frequency fricative energy of the /s/ in /sa/ is completely excluded from processing. This does suggest at least one possible weakness of the model, namely its insensitivity to energy in frequencies outside the rather narrow sub-band. In particular the P-centre of sounds dominated by high frequency energy may not be well predicted by this model.

### *5.1.8 Harsin*

Harsin's model is another global model with at least some relationship to that of Pompino-Marschall's model. The key differences are that the temporal integration calculation, loudness calculation, and method of identifying partial events are quite a bit simpler than in Pompino-Marschall's model. Harsin also introduces the concept of a *psychoacoustic envelope* into the model, though it is not clear whether this is a better or simply different representation of the hearing process.

Harsin's model operates as described in the following steps:

1. First the sound data is resampled to 10 kHz, giving sufficient bandwidth for narrowband speech.

2. Next, the signal is filtered (Butterworth, order 2) into 6 bands, namely: 366–659 Hz, 1073–1293 Hz, 1635–1928 Hz, 2172–2586 Hz, 2904–3514 Hz, and 3956–4758 Hz.

3. Within each band, an envelope is estimated. The processing steps are as follows: full wave rectification, low pass filtering (100 Hz, Butterworth, order 3), downsampling to 400 Hz (a factor of 25), low pass filtering the downsampled signal (100 Hz, Butterworth, order 3), and finally clipping negative values (caused by filter ringing) to zero.

4. Each envelope is scaled to approximate human loudness perception. Harsin specified that this should be achieved by raising each envelope value to the power 0.3 (Harsin 1993, p. 40; 1997). However, this value is appropriate only for intensity (power) signals. The envelope, which is an amplitude signal, should be raised to the power 0.6 to approximate loudness scaling (see for example Gelfand 1998). Nevertheless, this implementation uses the value specified by Harsin.

5. Modulations in each loudness envelope are analysed into four modulation bands: 3.1–5.5 Hz, 6.3–11.7 Hz, 12.5–23.5 Hz, and 24.2–46.9 Hz. The processing steps are as follows: each envelope is prepended with 512 zeros; starting from the first sample and advancing 4 samples (10 ms) each time, frames of 512 samples are extracted; the modulation power spectral density is estimated for each frame with a 512 point FFT (rectangular window); finally, spectral power is summed in each of the modulation bands and the square root taken to yield a modulation (magnitude) envelope.

6. Next each set of four modulation bands is weighted and combined to form a psychoacoustic envelope. Weights for the four modulation bands are 1.00, 0.80, 0.45, and 0.20, from lowest to highest frequencies respectively. Because Harsin appears to use the terms power and magnitude interchangeably it is not clear whether these weights should be applied to modulation power or modulation magnitude. This implementation assumed the latter (See also Zwicker & Fastl 1999). The magnitude of each modulation band is weighted and then squared to yield a modulation power. Modulation powers are summed across bands before taking the square root to yield the psychoacoustic (magnitude) envelope.

7. The velocity of the psychoacoustic envelope is calculated as the first difference of the envelope and the measures used by the

model are extracted at each velocity peak. Specifically, for each band, *i*, and peak, *j*, these measures are the time of the peak ($t_{ij}$), the peak velocity ($v_{ij}$), and the within-band magnitude change ($\Delta m_{ij}$). Fundamentally, the magnitude change is the difference between the envelope magnitude at one velocity peak and the magnitude at the previous velocity peak within the same band (if any, otherwise 0). Once again, there is some ambiguity regarding this measure. Harsin uses the term magnitude increment, suggesting that its value should always be positive, but later describes it as the amount of change (Harsin 1997, p. 249). The implementation choices which appear most compatible with Harsin's description seem to be to accept negative magnitude changes, to clip negative changes to zero, or to take the absolute value of the change. Although none of these options is entirely satisfactory (see Figure 5.8 and associated discussion for details), this implementation uses the absolute value approach by default.

8. Finally, the P-centre prediction is calculated as a temporal "centre of gravity" of the magnitude (change) weighted velocity, according to the following equation:

$$PC = \frac{\sum_i \sum_j \Delta m_{ij}\, v_{ij}\, t_{ij}}{\sum_i \sum_j \Delta m_{ij}\, v_{ij}} \qquad (5.17)$$

In addition to the ambiguity and assumptions described above there are some points to be made. The modulation analysis window is extremely long (1280 ms). Thus a single modulation will continue to affect the psychoacoustic envelope more than 1 second later. The modulation weighting (step 6 above) is applied to the loudness envelope implying that it is sensitivity to loudness modulations rather than amplitude modulations that is being modelled. This does not appear to be in keeping with the data on fluctuation sensitivity (Zwicker & Fastl 1999).

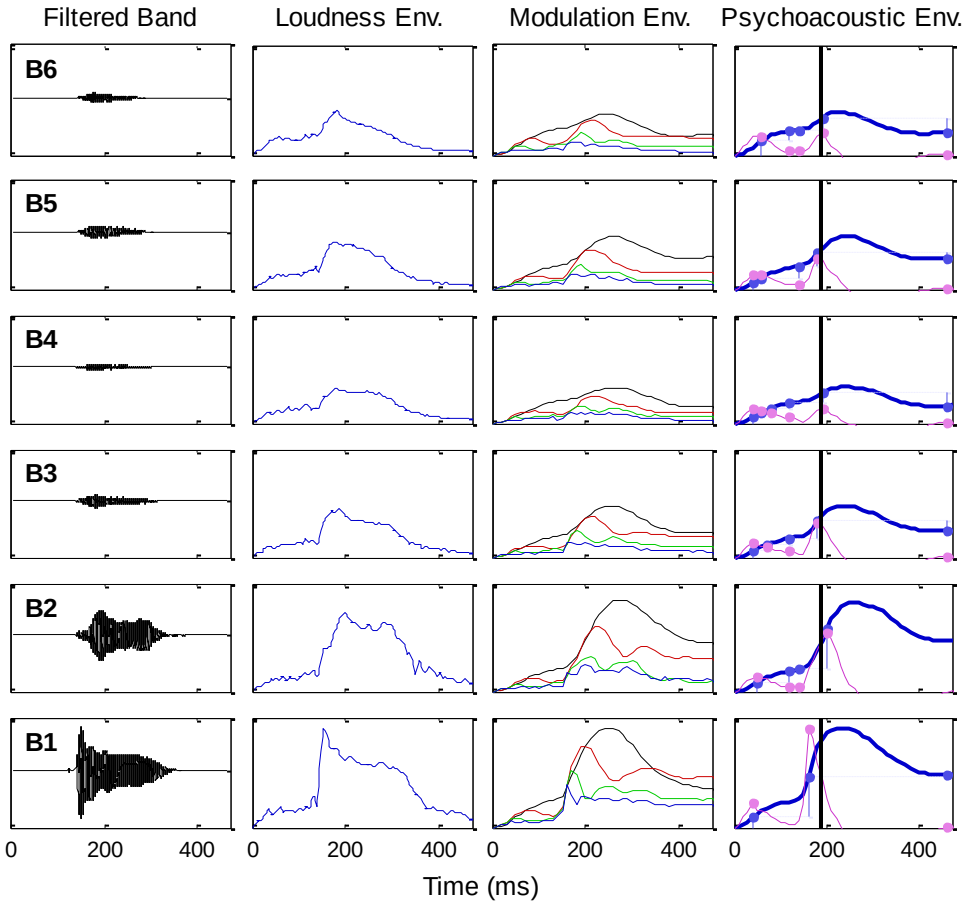


Figure 5.7 Harsin's per band magnitude-weighted velocity model applied to the sound /sa/. Each row of the figure corresponds to a frequency band (B1 to B6, low to high) approximately 2 critical bands wide. Across each row, the panels are as follows: the filtered waveform; the loudness scaled envelope; modulation envelopes in 4 separate sub-bands; and finally the psychoacoustic envelope (heavy line) and its corresponding velocity (light line). Velocity peaks and envelope magnitude changes between velocity peaks are also shown. The P-centre (heavy vertical line) is the "centre of gravity" of the magnitude-change-weighted velocities.

Figure 5.7 depicts the main processing stages in the model. A side effect of the long analysis window is that a modulation envelope settles to a constant or approximately constant value (see the modulation envelopes in Figure 5.7, band B3 for example) once the modulated portion of the signal is entirely within the analysis window—this constant value persists until subsequent modulations in the signal (if any) or the initial modulated portion of the signal clears the analysis window more than one second later.

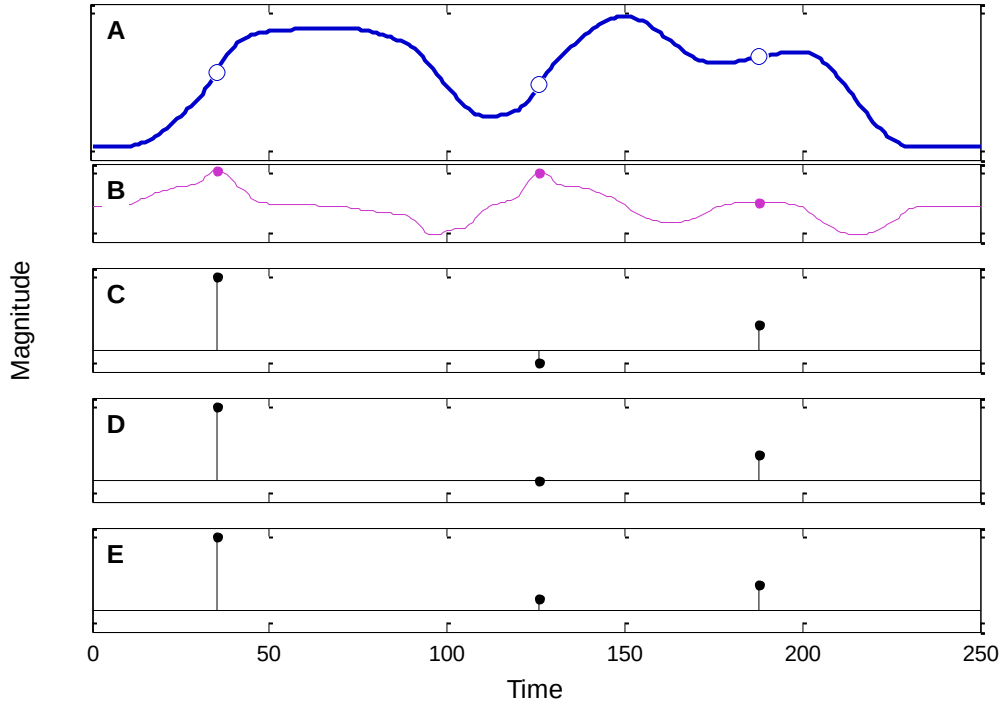


Figure 5.8 Three different methods for calculating "magnitude increments" in Harsin's model. (A) The magnitude envelope (symbols indicate partial events signalled by peak velocity times); (B) the envelope velocity (symbols again indicate the times of peak velocity); (C) magnitude increments calculated as the difference in magnitude between consecutive partial events; (D) magnitude increments calculated as in (C) but negative values are clipped to zero; (E) magnitude increments calculated as in (C) and then converted to their absolute values. Arbitrary units were used for both time and magnitude.

As noted in step 8 of the model description there is ambiguity over how to implement the calculation of magnitude increments. The three possibilities that appear to be at least somewhat consistent with Harsin's explanation are illustrated in Figure 5.8 If magnitude changes are calculated normally as the simple difference in magnitude between consecutive events, then some of these differences can be negative (see Figure 5.8, C). This is despite the fact that all (positive) velocity peaks must by definition occur during partial onsets. The distortion that a negative magnitude change would introduce into the centre of gravity calculation does not seem appropriate and therefore signed magnitude changes were not considered further.

The second option considered was to retain only positive magnitude changes. The simplest way to implement this is to clip all negative changes to zero (see Figure 5.8, D). In the example given in the figure it can be seen that this has the effect of retaining the third partial onset while suppressing the second. However, inspection of Figure 5.8 (A) would suggest the third partial onset would be less perceptually salient than the second one. Therefore, keeping only positive magnitude changes does not seem to be the appropriate behaviour either.

The final option considered was to use only the absolute value of the magnitude change (see Figure 5.8, E). This retains all events and does not distort the centre of gravity calculation with negative weights. It does however assign importance to decreases in magnitude between events and this behaviour also seems rather difficult to justify. In the end, none of these options is satisfactory and it may well be that a more complex approach to calculating magnitude changes is warranted, based perhaps on the approach of Pompino-Marschall. Nevertheless there is no evidence that Harsin implemented or investigated any more complex approach and thus the absolute magnitude change was chosen as the least unsatisfactory option.

## 5.2 Present study

The main objective of the present study was to evaluate and compare the existing P-centre models in a comprehensive manner. Such a comparison has not been reported in any of the literature to date, though the most recent model, that of Harsin (1997), was published more than a decade ago. As a consequence, there is currently no clear direction or recommendation that can be given in answer to the question: which model should one use? In particular it is not clear if the models perform more or less similarly or if, alternatively, there some models which perform well, and others which perform badly.

It may seem more direct to simply determine which single model performs best, but this overlooks a number of important details. First, if the only metric of performance is the correspondence of P-centre model predictions to behavioural measurements, then clearly there is considerable dependence on the test corpus used. In fact, by this metric, each model has already demonstrated excellent performance when tested against its own test corpus. Of course, using a single common corpus provides a more objective basis for performance measurement, but it is still the case that another corpus may yield a different result. Second, the implementation complexity of the models varies substantially. It may be acceptable to trade minor performance degradation for a simpler model. Third, the model which happens to fit a particular test corpus most closely may not give much insight into the underlying mechanism and psychophysics of the P-centre phenomenon and event perception in general. Finally, all the existing models assume discrete events with well defined boundaries. Not all models will be equally suitable for extension to predicting the P-centres of continuous event sequence (such as ordinary continuous speech).

For the reasons just given, the models were first compared against one another without any reference to behaviourally measured P-centres. A large corpus of discrete sounds comprising speech, instrumental, and synthetic material, was used. This large corpus, hereinafter called the "consistency corpus", seemed more likely to yield results that would generalize to other sound sets. Furthermore, none of the models had yet been tested on a wide ranging corpus and in particular certain models had been tested only with speech sounds and others only with non-speech sounds. Therefore it seemed likely that models which had been tested with one sound category should give similar predictions for those sounds, but might yield rather variable predictions for other sound categories. For example models originally tested with non-speech might be expected to perform rather variably with speech. There were two main questions to be answered: first, how consistent would the model predictions be in general, and second,

would there be any subset of "problem" sounds for which the models were particularly inconsistent?

In the second evaluation, each of the model's predictions was compared with sounds for which P-centres had already been behaviourally measured (see Chapter 3 and Chapter 4), hereinafter called the "measured corpus". The measured corpus was a strict subset of the consistency corpus and was necessarily much smaller because of the time-consuming nature of behavioural P-centre measurement. Additionally, most sounds in the measured corpus were speech. As a consequence, the results of the second evaluation must be interpreted carefully. There were two related questions to be addressed: which model or models would provide the most accurate predictions, and would there be some models which perform particularly badly? Taking the latter question first, if there were models which performed badly with the measured corpus, it seemed appropriate to conclude that these should not be used by researchers in future (at least not without modification). If, alternatively, a model performed well on this corpus then it would certainly be a candidate for future consideration, particularly if the set of sounds were similar to the test corpus used here. However, candidate models would require further testing with a larger test corpus before definitive recommendations could be made (a point which will be explored in more detail in the discussion.)

## 5.3 Evaluation I—model comparison

The first evaluation compared all model predictions against each other and did not compare against behaviourally measured P-centres. This permitted a large corpus with a wide range of acoustic properties to be tested, including slow onset, fast onset, speech, non-speech, and synthetic sounds.

### *5.3.1 Materials and method*

The consistency corpus comprised 259 sounds in three broad categories: speech, musical, and synthetic. As the sounds came from a variety of sources they were first normalized to a common sample rate (48 kHz) and loudness (nominally 65 phon). Loudness equalization was achieved by adjusting the level of each sound until its peak loudness equalled that of a 1 kHz, 65 dB SPL tone. The loudness calculation was performed in accordance with ITU-R BS.1770 (ITU-R 2006) using an exponentially averaged RMS (with a 125 ms time constant).

Although Patel, Lofqvist and Naito (1999) made a database of their speech sounds available for P-centre research, the P-centres of sounds in this database were never measured. As such, the primary usefulness to the model evaluations undertaken here was that it provided a readily available database of discrete speech sounds suitable for P-centre measurement. However each recording in the database contained repeated sounds and it was necessary to extract just one instance of each for use in the consistency corpus. In each recording, certain productions were better (clearer or more intelligible) than others and because of this the sound selected for extraction was not always the first in the recording. For the consistency corpus a single production of each of the monosyllables /ba/, /cha/, /ha/, /la/, /lad/, /li/, /ma/, /pa/, /sa/, /spa/, /ta/, and /ya/ from one male and one female speaker (DY and LC respectively) were selected. All 24 sounds were originally sampled at 10 kHz and the mean duration was 540 ms.

The consistency corpus included all speech sounds recorded specifically for the work in this thesis. High quality studio recordings of the monosyllables /ba/, /la/, /pa/, /pla/, /sa/, /spa/, and /spla/ had been made with two male and two female speakers. (The productions from just one of these speakers were previously used in experiments described in Chapter 3 and Chapter 4.) The 28 sounds were originally recorded at 48 kHz and were of moderate duration ($M$ = 485 ms). A variety of additional monosyllables, produced again by two female and two male speakers, had been recorded in

a quiet room setting. The speech tokens included the digits one, two, five, and six, and the syllables /da/, /ta/, /ga/, /ka/, /na/, /ra/, /sa/, and the words "eel", "wheel", "you" and "you'll", though not all speakers produced all tokens. In all, 49 sounds were selected. The recorded sample rate was 11.025 kHz and once again the sounds were of moderate duration ($M$ = 432 ms).

Synthetic sounds in the consistency corpus included six ramped tones (see Chapter 4) and a harmonic tone and noise mixture (see Chapter 3) developed specifically for the work in this thesis.

Additional synthetic sounds were selected from a database of sounds created by Collins (2006). These included 25 tones, one at each of five octave spaced frequencies (128–2048 Hz) and five onset durations (0, 10, 20, 45, and 100 ms). The total duration of each tone was fixed at 200 ms and onsets were followed immediately by offsets, both of which ramped linearly on a dB scale (where the minimum was 90 dB below full scale). White noise sounds were synthesized with 25 onset durations (0–240 ms in 10 ms steps). The total duration of each noise was fixed at 240 ms and envelope shaping was as for tones. A further 10 sounds, consisted of very brief sound extracts (70–315 ms) taken from dance music. These sounds were typically percussive in nature and formed from a composite of several original sounds that had been subjected to heavy processing during mixing. All 60 sounds were synthesized or sampled at 44.1 kHz.

The final category of sounds in the consistency corpus, musical sounds, were all selected from the database of Collins. In all, 39 sounds were selected, of which 13 were percussion hits, 3 were vocal sounds, and the remainder were a variety of stringed, brass, and wind instruments. As might be expected, durations were shorter for percussion sounds ($M$ = 289 ms) than the other sounds ($M$ = 501 ms). All sounds were sampled at 44.1 kHz.

For convenience each of the sound sets above was given a short label and these were PSyl, VSyl, VSpeech, VSynth, CSynth, CDance, CInst, and CPerc respectively.

Each model was used in its default configuration as described previously. In cases where a researcher described several different models, or model variants, only the model or variant which the researcher found performed best on their own test corpus was subjected to further evaluation here. Again, for convenience short labels were associated with each model as follows: MCS (Marcus), VRH (Vos and Rasch), GDN (Gordon), SCT (Scott), HWL (Howell), PML (Pompino-Marschall), and HSN (Harsin).

Each model was applied to all sounds in the consistency corpus including the reference noise. Thereafter RPCs were calculated as normal by taking the difference between the P-centre predictions of each sound and the reference noise. This procedure made P-centres comparable between models (and would later be used to compare with measured values). The RPC predictions were then compared between models for each sound.

### *5.3.2 Results and discussion*

The main results demonstrating the level of consistency between models for different sounds and sound sets are shown in Figure 5.9. The standard deviation of RPC predictions between models for individual sounds varied considerably, ranging from 6 to 137 ms ($M$ = 36 ms).

Examining the data in detail, it can be seen that the predictions for synthetic sounds are generally very consistent between models (see Figure 5.9, E). This is an interesting result that suggests that synthetic sounds may not be suitable for testing P-centre models. However all the synthetic sounds in the corpus except for those in the CDance set had simple envelope shapes and essentially constant spectra. Perhaps more complex synthetic sounds would prove to be suitable for model evaluation. Nevertheless, based on the results obtained here, such suitability would have to be demonstrated.

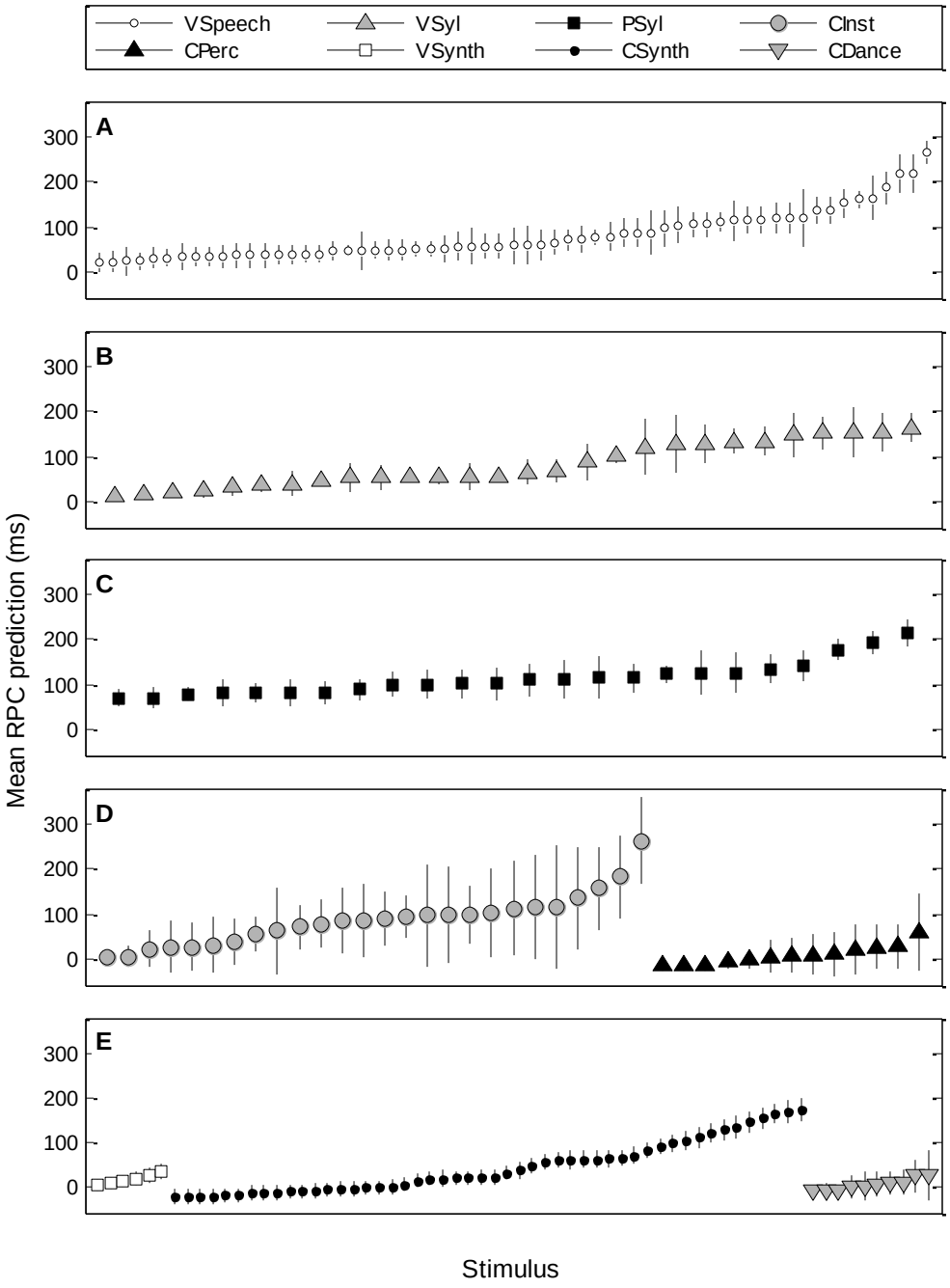


Figure 5.9 The consistency of model predicted RPCs for all sounds in the consistency corpus. Symbols mark the between-models mean RPC prediction and error bars show ±1 *SD*. Three sound categories were tested: speech (A–D), musical sounds (E), and synthetic sounds (F). Within each category the specific sound sets are identified by their short labels in the legend. Sounds within each set were sorted according to mean RPC prediction for presentation purposes.

Predictions are also quite consistent between models for many of the speech monosyllables in sound sets VSpeech, VSyl, and PSyl, particularly those with earlier mean RPC predictions (see Figure 5.9, A–C). Monosyllables with later mean RPCs do seem to elicit greater prediction differences between models although the trend is not reliable: certain late mean RPCs elicit predictions which are just as consistent as those of earlier RPCs.

Though by no means the least consistent sound set, some sounds in the CPerc set are associated with surprisingly inconsistent predictions. The inconsistency is surprising because these sounds, all percussion sounds with subjectively clear P-centres and short rapid onsets, should be straightforward to predict. Examination of some sounds in detail indicated that percussion sounds in particular tend to feature very high and very low frequency energy which can lie outside the frequency sub-bands used in some models. This in turn makes the predictions of those models unreliable with these sounds.

The least consistent predictions were elicited by sounds in the CInst sound set. In this set it appeared that in at least some cases the sounds had very late peak amplitude. This occurred for example with a slow bowed string sound. A related issue that arises with natural performance of sustained instrumental tones is that the level can drop in the middle of the sound before increasing again at the end. This envelope shape was observed for a sustained trumpet note. In all these cases the predictions varied according to how much importance each model attributed to later parts of the sound.

To gain additional insight individual model predictions were examined for the sounds with the greatest prediction inconsistencies. These individual predictions are shown in Figure 5.10. Because the model predictions are inconsistent it can be difficult to read this figure, but the main observations do not necessarily require very close reading. First, it is clear that inconsistency is not due any one problematic model. Furthermore, though there are some models which consistently predict early RPCs (notably the

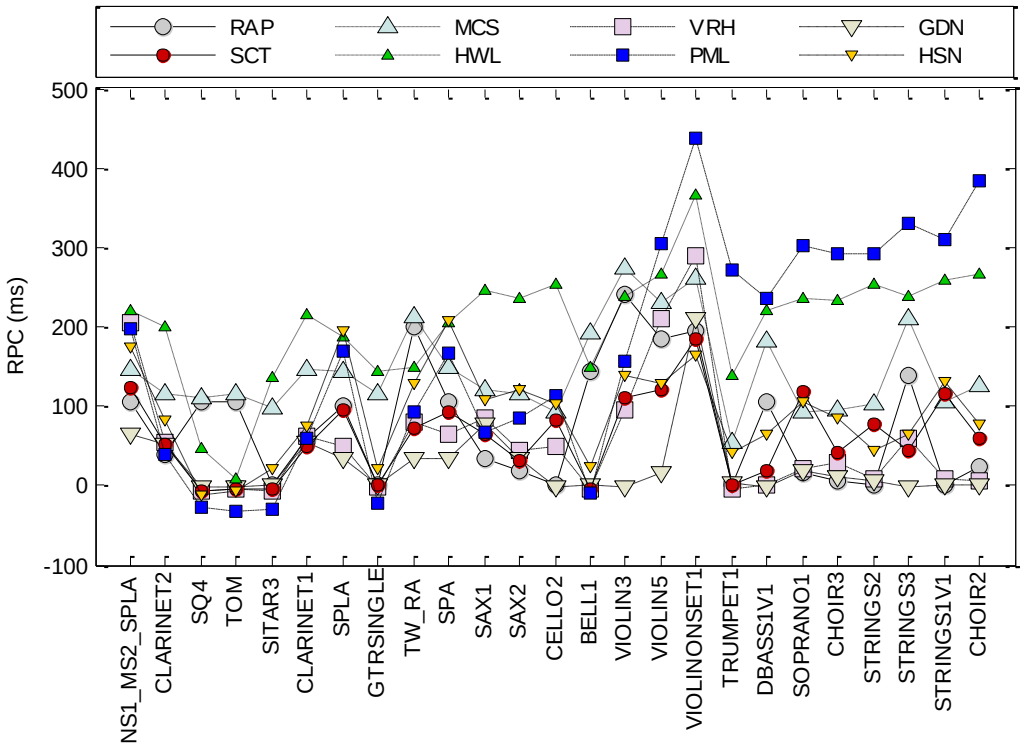


Figure 5.10 The 25 least consistent model predicted RPCs. These sounds are exclusively monosyllables (from the VSyl set) and instrumental sounds (from the CInst and CPerc sets).

models of Gordon and Vos and Rasch), the results of other models are more variable. For example, Pompino-Marschall's model makes early predictions for the speech sounds, but tends to make late predictions for the instrument sounds. It is also worth noting that the range of RPC predictions for each of these sounds is very large; differences even between models which appear clustered together in the figure may be detectable. The principle conclusion, then, is that unless at least one of these models can predict P-centres accurately in all cases then none of them can.

To conclude the objective comparisons in this evaluation, the RPC predictions of all models were subjected to pairwise correlation. The results of these correlations are shown in Table 5.1. The most similar model pairs were those of Scott and Harsin, Vos and Rasch and Gordon, and Vos and Rasch and Scott. Although the latter two pairs could have been expected based on similarities in model approach, the correlation between Harsin

**Table 5.1 Correlation of predicted RPCs between pairs of models**

| Model | MCS | VRH | GDN | SCT | HWL | PML | HSN |
|---|---|---|---|---|---|---|---|
| RAP | 0.840 | 0.793 | 0.732 | 0.753 | 0.545 | 0.590 | 0.749 |
| MCS | | 0.707 | 0.626 | 0.718 | 0.864 | 0.707 | 0.761 |
| VRH | | | 0.926 | 0.896 | 0.639 | 0.683 | 0.880 |
| GDN | | | | 0.850 | 0.549 | 0.572 | 0.842 |
| SCT | | | | | 0.723 | 0.762 | 0.949 |
| HWL | | | | | | 0.809 | 0.750 |
| PML | | | | | | | 0.728 |

*Note*—The short model labels are as described previously. All correlations were significant at the .001 level, *N*=258.

and Scott is surprising since these two models vary greatly in approach and complexity. The least similar models were those of Rapp and Howell while in general the models of Howell and Pompino-Marschall appear to be least similar to the other models on test.

It should also be noted that high correlation indicates that models tend to make predictions which vary in the same direction and by about the same normalized magnitude. This normalization is important because it hides the fact that predictions could still differ substantially without additional correction. Interestingly, several models, including Gordon's, Scott's, and Harsin's, include a final linear scaling stage. The coefficient values for this linear scaling were originally obtained by each researcher fitting their model's predictions to their own test corpus. It is possible that different coefficient values could make these models more consistent. Nevertheless the question remains: to what extent would any such coefficient values be specific to the corpus in use? If a model must be adjusted for each new corpus, it is clearly of very limited use.

## 5.4 Evaluation II—prediction accuracy

The second evaluation compared all model predictions against behaviourally measured P-centres. There were two main objectives: to determine which model or models make predictions that match measured P-centres most closely; and to determine whether any models make particularly inaccurate predictions.

### *5.4.1 Materials and methods*

Only sounds in the measured corpus were used in this evaluation. The measured corpus was a strict subset of the consistency corpus used in the first evaluation. All the sounds in the measured corpus had been used in previous experiments described in this thesis. Specifically, these were the usual reference noise sound, 6 tones from the VSynth set, and 19 speech sounds: 7 monosyllables from the VSyl set (one female speaker) and 12 digits from the VSpeech set (three different speakers).

The configuration and operation of each model was unchanged from the first evaluation. In this case, each model was applied to all sounds in the measured corpus including the reference noise. Thereafter, predicted RPC values were calculated and compared with corresponding measured values for each sound. Measured values for the VSyl and VSpeech sets were taken from the results of Experiment I and Experiment IV respectively.

### *5.4.2 Results and discussion*

To meaningfully compare predicted and measured RPC values, appropriate metrics must be selected. The most significant factor to consider is that people exhibit a certain amount of tolerance for timing deviations (as indeed they must since humans are generally unable to consistently perform rhythmic tasks with objectively precise timing). Madison and Merker (2002) found the threshold of anisochrony detection in an

approximately isochronous sequence was 3.5% of the nominal IOI. Friberg and Sundberg (1995), surveying previous research and integrating their own results, found that the just noticeable difference[32] from isochrony was 5% of the IOI (for IOIs larger than about 250 ms). Assuming the interval between sounds in an isochronous sequence was 700 ms then deviations from isochrony exceeding ±35 ms would be detectable[33]. Therefore it was assumed that RPC prediction errors exceeding this range would be significant. (With shorter intervals between sounds, such as those occurring in music or continuous speech, the range of acceptable error would get smaller.)

Several error measures were analysed. The root mean square error (RMSE) between a set of predicted and measured values gives an indication of the average error across all sounds. On its own, however, this is not sufficient. For example an RMSE that falls within the range of acceptable error could have several interpretations: perhaps all the RPCs were predicted with an acceptably small error but it could also be that one or a small number of prediction errors were large if they were compensated by a number of very accurate predictions. In summary, a small RMSE would be necessary but not sufficient to indicate an accurate model. In contrast, a large RMSE would automatically indicate poor predictions.

The second measure evaluated was simply the maximum error which could either be an underestimation (negative) or overestimation (positive) of the measured RPC. A completely accurate model should have a maximum error within the acceptable error range. The maximum error exhibited by a reasonably good model would not lie far outside the acceptable range. In

[32] There are various ways of measuring and estimating the just noticeable difference. The value used here is the value that Friberg and Sundberg indicated would be expected for the 50% correct level obtained with a two alternative forced choice method.

[33] The inter-onset interval (IOI) had been 650 ms in Experiment I and 700 ms in Experiment IV. The IOI used to calculate the jnd from isochrony was chosen to be 700 ms since more of the stimuli used here came from Experiment IV and the jnd is slightly larger for this IOI.

**Table 5.2 Errors between model predicted RPCs and measured RPCs**

| Model | RMSE (ms) | Max Error (ms) | Detectable (%) |
|---|---|---|---|
| RAP | 34.1 | -81.0 | 20.0 |
| MCS | 36.0 | 101.5 | 28.0 |
| VRH | 43.4 | -124.8 | 28.0 |
| GDN | 48.7 | -152.8 | 20.0 |
| SCT | 29.7 | -92.0 | 16.0 |
| HWL | 46.5 | 70.0 | 68.0 |
| PML | 19.0 | 39.4 | 8.0 |
| HSN | 33.7 | 103.5 | 20.0 |

*Note*—RMSE = root mean square error between model predicted and measured RPCs; Max Error = largest absolute error with the sign indicating whether the prediction underestimated (negative) or overestimated (positive) the measured RPC; Detectable = percentage of total sounds tested that exceed the acceptable error threshold (assumed to be ±35 ms). The reference sound was used to calculate RPCs but otherwise did not participate in the calculations ($N$ = 25).

such a case it is conceivable that the prediction may be acceptably close to a different sample of P-centre measurements.

The final measure examined was the percentage of detectably erroneous RPC predictions. Naturally this percentage is highly dependent on the specific sounds in the test corpus (mainly speech sounds in this case). Additionally, the size of the corpus was relatively small and it would therefore be unreasonable to generalize any result too far. Nevertheless, any prediction errors generated by a model with this corpus strongly suggests that errors could be expected with other sounds also.

Table 5.2 shows the main results obtained. Most of the models make similar numbers of detectable prediction errors (20–28%), but Pompino-Marschall's model performs better than most while Howell's model performs much worse. This latter result is not very surprising because the Howell model implementation was certainly too simplistic and did not even

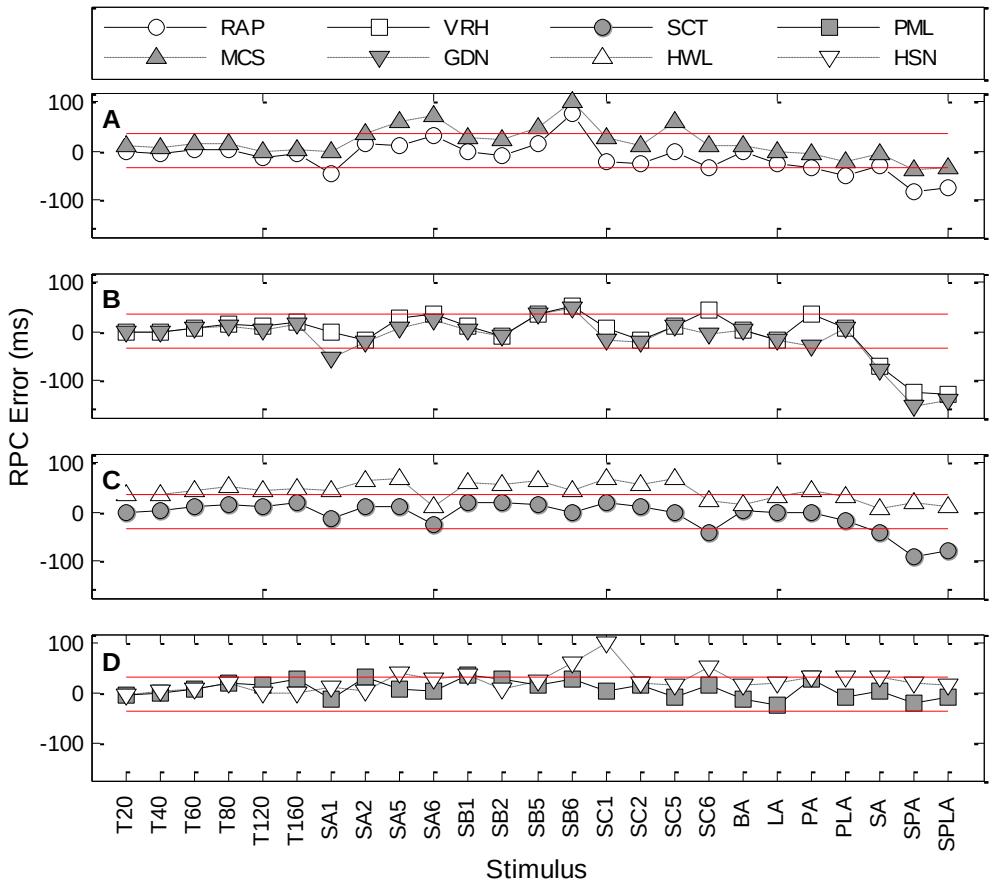


Figure 5.11 Errors between model predicted and measured RPCs. Each panel shows the errors measured for two models identified with their short labels in the legend. The horizontal lines indicate errors of ±35 ms, corresponding to the just noticeable difference from isochrony with an interval of 700 ms between P-centres. The reference sound, N, is not included in the figure because the RPC of N to itself is zero by definition and thus there can be no prediction error.

include a linear scaling stage to bring its RPC predictions closer to measured values. The maximum error results show that all models except those of Pompino-Marschall and Howell make prediction errors more than twice the detectable error threshold. Finally Pompino-Marschall's model also exhibits the smallest prediction RMSE.

Figure 5.11 presents an alternative view of the results from which the individual sounds eliciting prediction errors can be determined. It is clear that the models of Rapp-Holmgren, Vos and Rasch, Gordon, and Scott underestimate the RPCs for the sounds SPA and SPLA. Both these sounds

have complex onsets which the onset models appear unable to handle correctly. In contrast, the models of Howell and Marcus seem to overestimate several RPCs, at least half of corpus in the case of Howell's model. Harsin's model appears accurate for most sounds but significantly overestimates the P-centre of certain speaker-digit combinations (SB6 and SC1). Finally, it is clear that Pompino-Marschall's model generally predicts within or very close to the acceptable range of error for these speech and synthetic sounds.

In general there was no overall pattern of prediction errors that could be discerned which might indicate something problematic with the test stimuli or some problem shared by all models. In particular, the total number of prediction errors per sound exhibited no clear relationship to either the duration or the measured RPC of the sounds.

## 5.5 General discussion and conclusions

Several questions were posed by this study. Are the predictions of the existing models consistent? Are there any sounds that reveal particularly large inconsistencies? Which models predict measured RPCs most closely and least closely? Can any guidance be given to a researcher wondering which model they should use? Each of these questions will be dealt with in turn.

The results of the consistency evaluation showed that the model predictions are generally not consistent with one another, although the inconsistencies may not be revealed by synthetic sounds with simple envelope shapes or brief natural sounds with simple onset structure. In contrast, long speech sounds, instrument tones, and even some percussion hits revealed the greatest differences between individual model predictions.

The prediction error evaluation results showed Pompino-Marschall's model yielded the most accurate RPC predictions on the measured corpus, which comprised speech and synthetic sounds exclusively. As suggested earlier,

care needs to be taken when interpreting this result due to the fairly limited nature of the corpus. In particular, the consistency evaluation showed that Pompino-Marschall's model tended to predict among the earliest RPCs for speech but the latest RPCs for some instrumental sounds. Inspection of the individual instrumental sounds suggested the model prediction for these sounds was not correct and was being distorted by features of the sound not often encountered in speech. Nevertheless, P-centres would need be measured for these sounds to confirm this.

Despite (or perhaps because of) their simplicity, the models of Rapp-Holmgren, Marcus, Vos and Rasch, and Gordon cannot be recommended for predicting P-centres based on the evaluation results in this study. Howell's model does not predict within the acceptable error range but its prediction error appears relatively constant with the measured corpus. It is possible that a linear correction could improve its predictions. Nevertheless, the very simple integration approach used has little relationship to the psychoacoustics of hearing and it may not be productive to pursue the model further. Scott's model yields reasonable RPC predictions for simple onset sounds but suffers with more complex sounds. Nevertheless it may be a suitable model with constrained sound sets. Finally, the results show that Harsin's model generally performs relatively well, though sounds with late energy can cause it problems.

So, which model should one use? The answer to this depends on the intended use. If the model is to be used to predict P-centres, then for sounds whose spectra do not have significant gaps and which have relatively simple envelopes and onset structures, Scott's model may be suitable. It is certainly straightforward to implement. For more complex speech sounds Pompino-Marschall's model seems to be better, but the results suggested that it may perform significantly less well with non-speech sounds such as instrumental tones. In summary, there is no single model which appears to predict the P-centres all sound types accurately and reliably.

If a researcher is wondering which model to use as a basis for refinement and further development, then the models of Scott, Harsin, and Pompino-Marschall would all be recommended. Various operational ambiguities and shortcomings were noted which could be rectified. The problem sounds or sound types identified in this study could also be analysed more closely so that specific model solutions could be synthesized. In particular it would be desirable to simplify the operation of Pompino-Marschalls's model and examine the integration scheme in detail, perhaps incorporating ideas from Scott's and Harsin's model. In short, there is room for model enhancement and there are also several obvious starting points for such enhancement.

Even though this study has shown that the existing models have some problems with isolated speech or non-speech events, the bigger open problem is the extension of P-centre models to continuous events. Marcus (1981) addressed the question of continuous speech but did not provide a concrete strategy for handling it. A simplistic approach would simply add an event segmentation stage prior to P-centre detection. Event segmentation, however, may be no easier than P-centre modelling (see for example Villing, Timoney & Ward 2006; Villing et al. 2004) and presupposes that there are event boundaries to be detected. It seems more likely that continuous event handling needs to be integrated into the model itself, a feature that will almost certainly require a compromise between the dichotomous approaches of onset models and global models that currently exist.

Finally, it was noted earlier that each P-centre model had originally been developed and tested with relatively sparse corpus. In fact the measured corpus used in this study was also relatively small for much the same reasons, specifically, that it is time consuming to make P-centre measurements. This does, however, present a problem for future research and model development. If each researcher must assemble their own corpus and make their own P-centre measurements, the task of modelling P-centres becomes unnecessarily arduous and will continue to yield results that are difficult to replicate. An alternative approach would be for each

researcher to develop and test P-centre models against a common corpus of P-centre labelled sounds, analogous to the prosodically and phonetically labelled corpora used in the domain of speech synthesis and recognition (for example Garofolo et al. 1993). Using this approach, researchers would be free to focus on the problems of modelling alone, thus lowering the barrier to entering the field, and ensuring that model prediction results could be easily replicated and compared.

# Chapter 6

# Concluding remarks

From the outset, it was apparent that research into the P-centre phenomenon had progressed quite slowly and intermittently. The P-centre term was coined more than thirty years ago (Morton, Marcus & Frankish 1976) and there was some directly relevant research which predated even that (for example Allen 1972a, 1972b; Rapp-Holmgren 1971). Nevertheless, there was little indication that the P-centre problem had been "solved" or even, perhaps, that a solution was close.

A practical solution to the P-centre problem would take the form of a model or algorithm that one could use to predict the P-centre of an event or, more usefully, those of a sequence of events. Such a model would find immediate applications in speech and music synthesis and research into event timing and rhythm. Despite the existence of several models, there was no indication that all the models had ever been compared (though certain subsets were), nor did the literature provide any help with answering the most fundamental question: which model should one use?

Theoretical progress on the P-centre problem had been frustrated by a number of factors: no comprehensive review of the literature existed; the empirical data were relatively sparse and divided into two research fields (speech and music) which had not been approached in a unified manner; and finally, behavioural measurement of P-centres had used a number of different methods such that it was unclear how the findings of the respective studies could be unified.

## 6.1 Summary of contributions

Having identified some issues with P-centre research reported to date, the work in this thesis focused on directly addressing a subset of these issues. Furthermore, it seemed both timely and necessary to establish a reliable foundation for subsequent research by critically integrating and evaluating developments to date. The alternative approach—exploring some empirical features of the P-centre and developing yet another model without derivation from those that had gone before—did not seem compelling.

In Chapter 2, the empirical data resulting from more than three decades of (acoustic) P-centre research was critically reviewed. The data consistently shows a strong effect of the onset segment and a weaker effect of post-onset segment. Generally the effects are stronger for speech (where the segments correspond to the syllable onset and rhyme) than for non-speech. It was hypothesised that the P-centre may be strongly influenced by change detection and the difference between speech and non-speech may prove to be due to greater degree of change that occurs in speech stimuli—the spectrum, amplitude, fundamental frequency, and harmonic to noise ratio may all change over a short time period. Results which require replication and significant unanswered empirical questions were also identified. Finally, the existing theoretical frameworks were reviewed and a modified theory, suggesting that the P-centre arises as a natural side effect of known psychoacoustic processing, was proposed.

As previously mentioned, there were several measurement problems associated with P-centre research: first, a number of methods had been used and it was not clear that these were compatible with each other; second, assumptions underlying the measurement methods had been insufficiently tested; and finally behavioural P-centre measurement is sufficiently time consuming that most research studies have been rather small (compared to other psychoacoustic studies such as for example the perception of pitch or loudness). In Chapter 3 the problems of measurement were investigated in detail. Past measurement methods were reviewed and

two, rhythm adjustment and tap asynchrony, were selected for further study alongside the new PCR method. Rhythm adjustment and the PCR method were shown to produce consistent P-centre estimates, indicating that they both measure the same percept. Although the PCR method permitted slightly more time-efficient measurement, either method would be recommended for future P-centre measurement. Despite its simplicity and attractive time-efficiency, the tap asynchrony method yielded P-centre measures which differed significantly from the other methods and therefore its use cannot really be recommended until this discrepancy has been investigated further. Additionally, the study investigated P-centre context independence, upon which all current measurement methods rely, and found no evidence of context dependence for the rhythm adjustment and PCR methods. Finally, the concept of P-centre clarity was introduced to describe the subjective precision with which an event's P-centre is perceived. Although it might naturally be expected that unclear P-centres would exhibit a greater dispersion of measurement observations than clear P-centres, the data in this study showed just one potentially reliable objective correlate, namely, the slope of the PCR function which indicates the strength of sensorimotor coupling.

Prior to this work, only behavioural P-centre measurement methods had been described in the literature. Whereas the tasks used by the tap asynchrony and PCR methods are largely unconscious and automatic, the tasks embodied by the rhythm adjustment and forced choice methods involve explicit subjective decision making. Neurophysiological measurement methods had never been directly applied to the measurement of P-centres, though they had been used in the related fields of rhythm and meter measurement (usually with the implicit and perhaps unrecognised assumption that the P-centres and onsets of the stimuli in use were approximately the same). Much as hearing thresholds can be measured using a variety of behavioural methods or by examining the auditory evoked potential in response to very brief stimuli, it seemed that a similar paradigm might work with P-centres. A neurophysiological correlate of the

P-centre would enable objective measurement, perhaps allowing the moment of perception itself (the elusive absolute P-centre) to be identified and certainly providing additional insight into the underlying psychophysiology of the P-centre itself. Chapter 4 described two experiments investigating neuroelectric correlates of the P-centre. It was shown that the phase of very low frequency oscillations in the evoked potential, specifically oscillations at the fundamental presentation rate, predicted the behaviourally measured P-centres. Oscillation at the presentation rate is a side effect of evoked potential components occurring at approximately the same latencies after each repeated stimulus presentation, thus forming a quasi-periodic waveform at that rate. As a consequence, the conclusion was that the low frequency phase is not directly the correlate of the P-centre but a side effect of altered timing in other, as yet unidentified, components of the evoked potential. This is an intriguing result which deserves further study.

The final study undertaken was an evaluation of the P-centre models that have been described in the literature. Implementing the models proved to be a substantial piece of work in itself. In many cases the model descriptions were either vague on certain points or missed them entirely. This problem is particularly prevalent with complex models published in journals where editorial concerns often seem to trade brevity and readability against the ability to replicate the model precisely. A simple resolution to this problem exists: the researcher's own code implementing the model should be published (with sufficient comments that it is readable in its own right). It is not sufficient to say that code is available on request; too many researchers have exited the field leaving no definitive model implementation behind. As an aid to future researchers, all models implemented in this thesis are documented in full in Appendix C. Furthermore details of assumptions that had to be made and alternative choices that could be made are described in Chapter 5. All models were applied to a large corpus of speech, instrumental, and synthetic sounds, the resulting P-centre predictions were compared, and the results showed that

the models are not consistent with one another. Subsequently the models were applied to the subset of the corpus for which behaviourally measured P-centres were available. In this case, the model predictions were compared with the measured P-centres and the most and least accurate models were identified. Even on this limited corpus, all of the models had some problems and thus, it appears that there is, as yet, no comprehensive and reliable P-centre model.

## 6.2 Future work

During this work various research avenues opened up that could not be pursued for one reason or another, although the most common reason was simply that a new research question only took form during the final analysis of a particular set of results. As always, a balance must be struck between opportunistic pursuit of new questions as they arise and the finite time that must ultimately be assigned to work such as this. In the end, this balance seemed appropriate.

Here, then, is a list of open problems that would seem to deserve further attention. Some of these are ongoing questions in P-centre research whereas others were formulated only during this work. The difficulty, scope, and eventual benefit of answering these questions varies greatly. To aid future researchers, problem groups are suggested and the anticipated impact of addressing these problem groups is indicated after the list.

1. Can a more reliable P-centre model be developed, possibly by extension and refinement of existing models, for well defined discrete events at least? This is perhaps the broadest and most significant open question to be addressed.

2. Before the P-centre of discrete events can be modelled reliably there would appear to be a number of open empirical questions relating the P-centre to various acoustic features that should be answered (see section 2.3.2, questions 2–7).

3. P-centre Models were evaluated using a large corpus of sounds for which P-centres had not been measured and a much smaller corpus for which they had. Unfortunately this work suffers, as has that of researchers before, from the time consuming nature of P-centre data collection and the consequent sparse data set. The task of modelling P-centres is, it seems, unnecessarily arduous. Each researcher must first collect their own data before modelling can begin. A better approach used in the domain of speech recognition and synthesis is to prepare a labelled corpus, either as one dedicated research project, or as an ongoing activity taking contributions from many researchers. A corpus of sounds, labelled with measured P-centres would allow researchers focus on modelling alone.

4. Previous P-centre research has used a variety of presentation configurations including speakers and headphones or earphones of various qualities, in a variety of acoustic environments. How robust is the P-centre in the face of such variation? Is the P-centre essentially unaffected? It would be easier to assemble a large P-centre corpus if it the listening environment did not play a significant role in the timing of P-centres.

5. P-centres are typically measured using isochronous rhythms with moderate rates of about 2 Hz or less. However natural speech and music generally features event rates higher than this (about 3–4 Hz for speech syllables and maybe 8–12 Hz for sixteenth notes in music). Is this discrepancy of any significance? It certainly seems to be true that the just noticeable difference of isochrony is a constant fraction of the inter-stimulus interval up to about 5 Hz and thus it would seem that sequences which may sound approximately isochronous at slower rates may be perceived as anisochronous at faster rates. On a related note, is there any rate

limit at which P-centre context independence starts to break down?

6. In Chapter 3 the measurement method conclusions were based on experiments with a relatively small set of speech sounds. Can similar results be obtained with non-speech sounds? Furthermore the PCR method used musically skilled participants. Can the results be replicated, except perhaps for slightly greater variability, with less musically skilled participants?

7. The experiments in Chapter 3 indicated that the tap asynchrony method produced relative P-centre estimates which appeared to be underestimated in comparison with the other methods (when all methods used a common reference sound). In particular there was a difference between the tap asynchrony results and those of the very similar homogeneous EOS sequence tap asynchrony. Several possible explanations for the difference were offered, but ultimately further investigation is required. Such investigation remains attractive because the tap asynchrony method seems to be easier for participants than the others tested and it would also appear to be the most time efficient method for researchers to use if its results could be trusted.

8. The concept of P-centre clarity was introduced but no direct measurement of this attribute of the P-centre was attempted. Can P-centre clarity be measured in reproducible manner? If so, it would be useful to include this attribute in the labelling of any P-centre corpus. In turn this would permit greater certainty in relating objective properties of the sound to this very subjective quality. Ultimately a comprehensive P-centre model could indicate not only the predicted P-centre but also its predicted clarity.

9. The results in Chapter 3 also showed that the strength of sensorimotor coupling, $\alpha$, (or alternatively, the confidence with

which a participant responds to a phase perturbation in an isochronous sequence) might vary over the course of a mixed event sequence. This variation in $\alpha$ might take place as a single or infrequent step change, as a continuous and gradual adaptation, or in a discrete manner, depending on the most recent event only. Distinguishing between these competing hypotheses requires a carefully designed experimental paradigm.

10. The results in Chapter 4 provided a tantalizing indication that there is a neuroelectric correlate of the P-centre and moreover that it is of sufficient magnitude to affect the overall phase of EEG oscillations at the presentation rate. Nevertheless the specific evoked components that are correlated with the P-centre could not be identified. The experiment deserves to be replicated, but perhaps with some methodological differences. Changes to consider include denser electrode placement, carefully parameterized stimuli, a slower presentation rate, diotic presentation, a task to control for vigilance, and, as is always desirable, more participants.

11. The magnitude of long latency evoked response components tends to decrease with increased presentation rate and this appeared to affect the results in Chapter 4. On a related note Snyder and Large (2004) also found that long latency response for tones essentially disappeared when they were repeated at relatively short random intervals (375–750 ms). Would the magnitude of these components be affected if the random intervals were relatively long? Is it possible that a neuroelectric correlate of the P-centre could be identified again using random intervals? The underlying question to be addressed is: does the P-centre emerge only as a side effect of meter, rhythm, and temporal prediction, or is the P-centre an innate property of each individual event, whether or not the event occurs in isolation or in a sequence.

12. All measurement methods explored to date measure relative P-centres or biased event-local P-centres (cf. Chapter 3). Is it possible to design a behavioural (or other) method which would enable absolute P-centres to be measured? Might that method be based on an EEG based measure of the P-centre?

13. All P-centre measurement methods are designed for discrete events, yet the most naturally produced stimuli are continuous and boundaries between events are not clear and unambiguous. How should P-centres be measured for such continuous stimuli, for example a short speech utterance?

14. As all measurement methods have focused on discrete events, so too have all P-centre models. How should a model for continuous events be developed? Both Marcus (1981) and Scott (1993) believed that their models could be applied to continuous events, though in reality neither model can be applied directly. Whether a model requires a distinct event offset or not, all models currently express their P-centre prediction with respect to the event onset. If this onset is not precise, then the P-centre prediction becomes corresponding imprecise. It seems that it may be possible to modify existing models by incorporating some event segmentation process before the model proper or by using some sort of moving window approach, but it is not clear that this is the right approach to take. A key question to consider is whether people first segment events and then perceive each event's P-centre or instead perceive a sequence of P-centres and afterwards (or at least independently) infer a set of event boundaries between those P-centres?

15. Related to the two previous points, what new P-centre phenomena might emerge as a consequence of studying continuous events? Questions 8 and 9 in section 2.3.2 may provide a useful starting point.

16. What is the nature of the P-centre? What specifically does it encode? Is it the moment at which an event is first perceived as an integrated entity, or the moment at which it is recognised or classified? Why does it appear to be distinct from the perceptual onset (in speech at least)? Is there any good reason why humans should have evolved to synchronise with a time point other than the perceptual onset of the event? These are deep questions that may be difficult to answer, but questions 10 and 11 in section 2.3.2 indicate some possible first steps towards exploring these issues.

17. In keeping with the general nature of the term proposed by Morton et al. (1976), the final and perhaps most far reaching suggestion for future work is to broaden P-centre research beyond the domain of acoustic stimuli. Short visual events should have P-centres. Physical movements have P-centres. What can be learned by exploring P-centre phenomena in these domains? Ultimately, how should all these phenomena be unified in a cross modal theory of P-centre perception?

Questions 1–3 are focused on the immediate problem of developing a constrained but reliable P-centre model. This is almost certainly the problem with the highest potential impact. Questions 4 and 5 are related but minor issues. The basic measurement method questions (6 and 7) that arose during this work deserve to be addressed for the sake of completeness and because of the potential to make empirical P-centre measurement more efficient.

P-centre clarity and adaptive sensorimotor coupling (questions 8 and 9) are novel concepts which certainly deserve further study but their potential impact on P-centre research is probably not well understood yet. The EEG study in this thesis opened up some very interesting questions (10–12) and pursuing these further appears attractive because of the potential for unique and novel insight into the P-centre phenomenon. The remaining questions (13–17) are more tentative and exploratory in nature. Although

the difficulty of addressing these problems is not well understood as yet, there are clearly questions with potentially high impact to be addressed, particularly relating to the P-centres of continuous events (questions 13–15).

## 6.3 Conclusions

In conclusion, the main contribution of this work has been to establish a more rigorous foundation upon which future research can build. This was accomplished by critical integration and review of the empirical data literature, a detailed investigation of the methods used to measure P-centres, the exploration of a novel EEG based approach to P-centre measurement, and finally, the implementation and evaluation of the existing P-centre models.

What the next three decades of research will bring remains to be seen, but it is to be hoped that our understanding of the P-centre, and event timing in general, will have advanced substantially by then. Perhaps it will finally be possible to do that which eludes us today—to measure and predict the perceived moment of an arbitrary event.

# Appendix A

# Experimental stimuli

All the sounds in this appendix were synthesized or recorded specifically for experiments conducted as part of this thesis. The sound waveform and spectrogram are presented for each sound giving some indication of the sound envelope and frequency content respectively.

## A.1 Digits

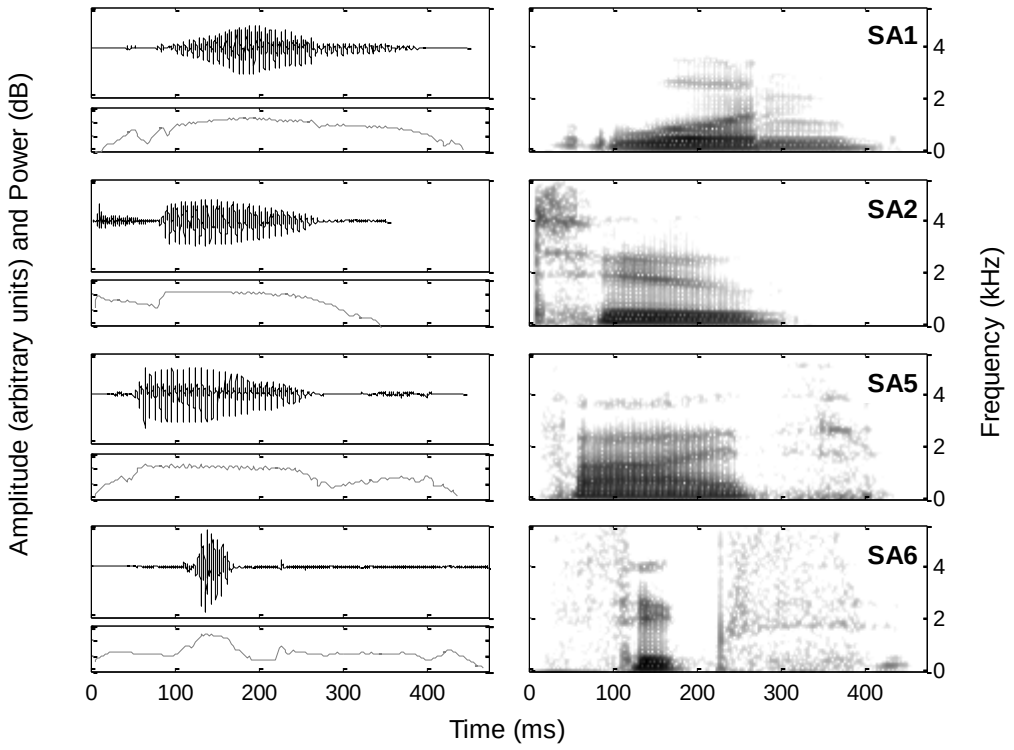


Figure A.1 Waveforms, short term power, and spectrograms for the digits "one", "two", "five" and "six" from speaker SA (female). Short term power is derived from the spectrogram analysis which uses a 10 ms window with an 80% overlap.

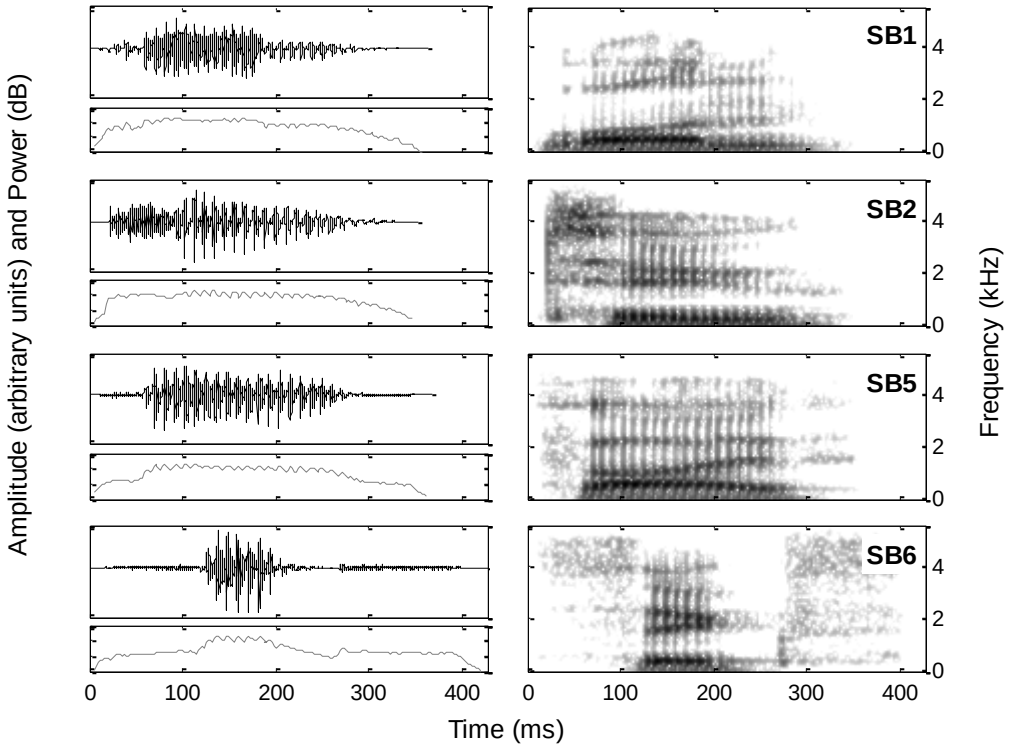


Figure A.2 Waveforms, short term power, and spectrograms for the digits "one", "two", "five" and "six" from speaker SB (male).

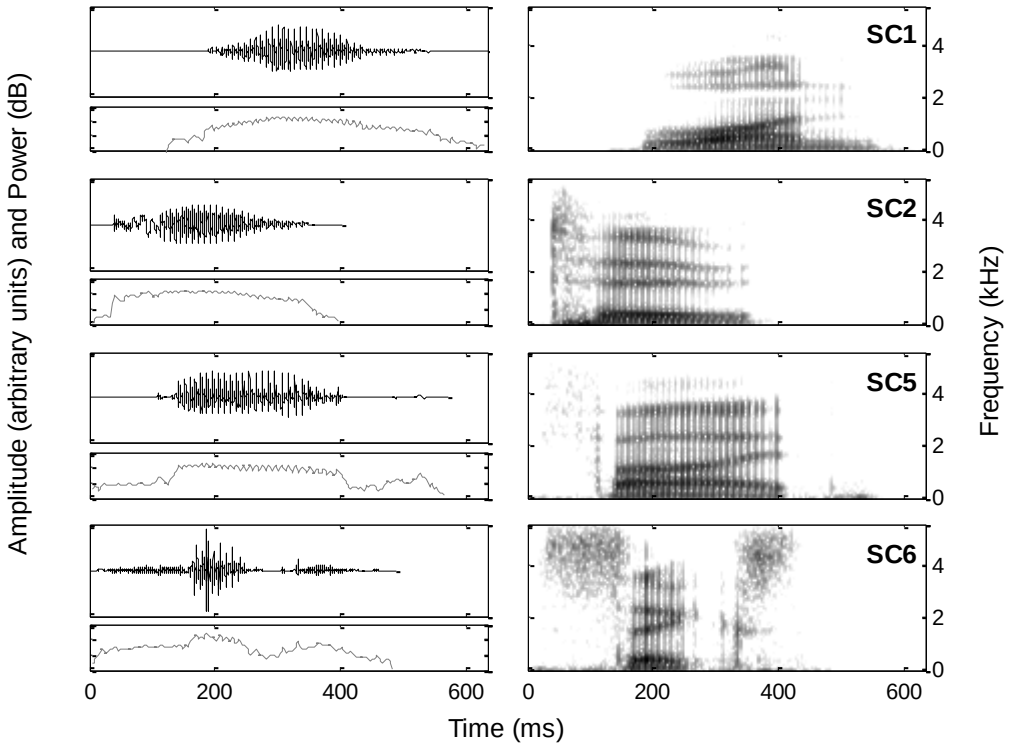


Figure A.3 Waveforms, short term power, and spectrograms for the digits "one", "two", "five" and "six" from speaker SC (male).

## A.2 Shaped Tones

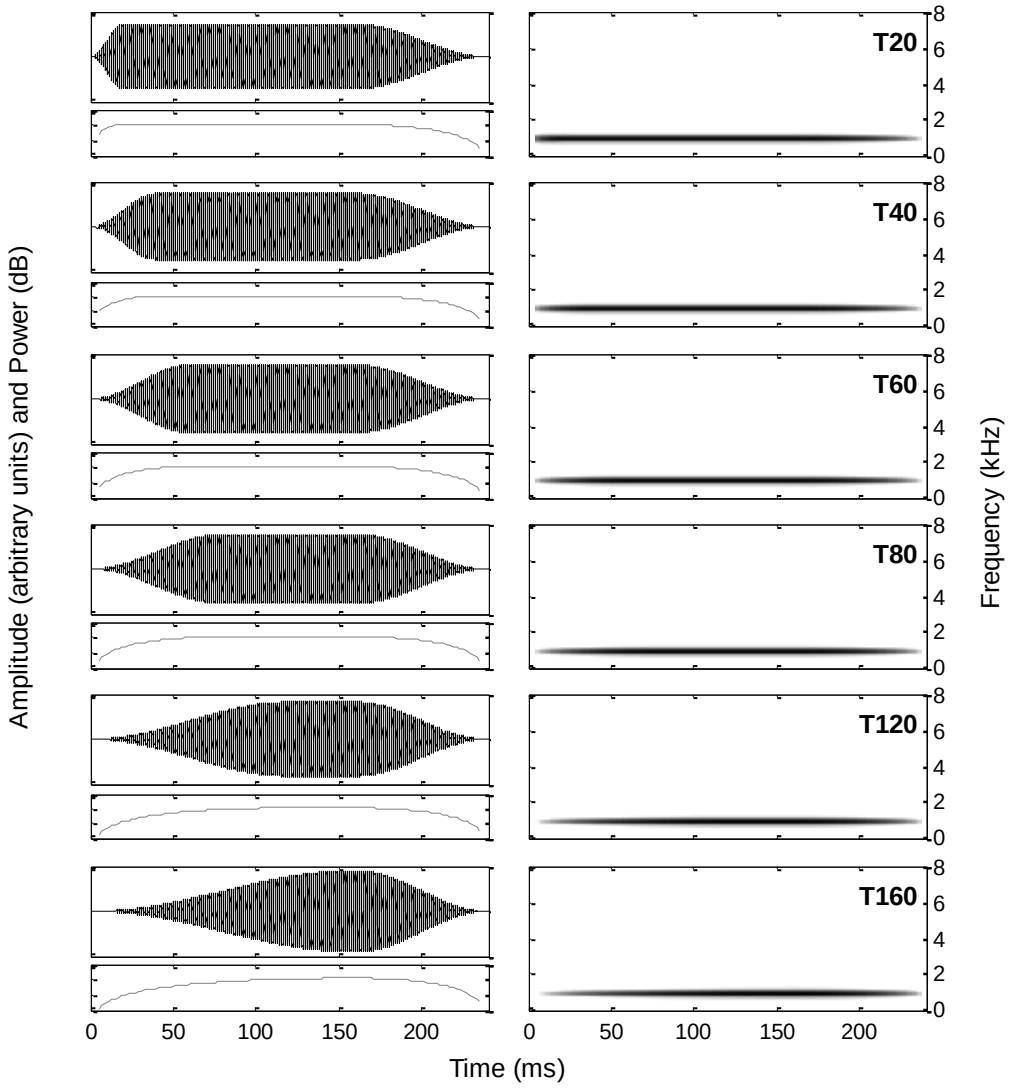


Figure A.4 Waveforms, short term power, and spectrograms for six tones. Each tone has a frequency 1000 Hz and is 240 ms long. Onsets and offsets are both shaped by raised cosines.

## A.3 Monosyllables

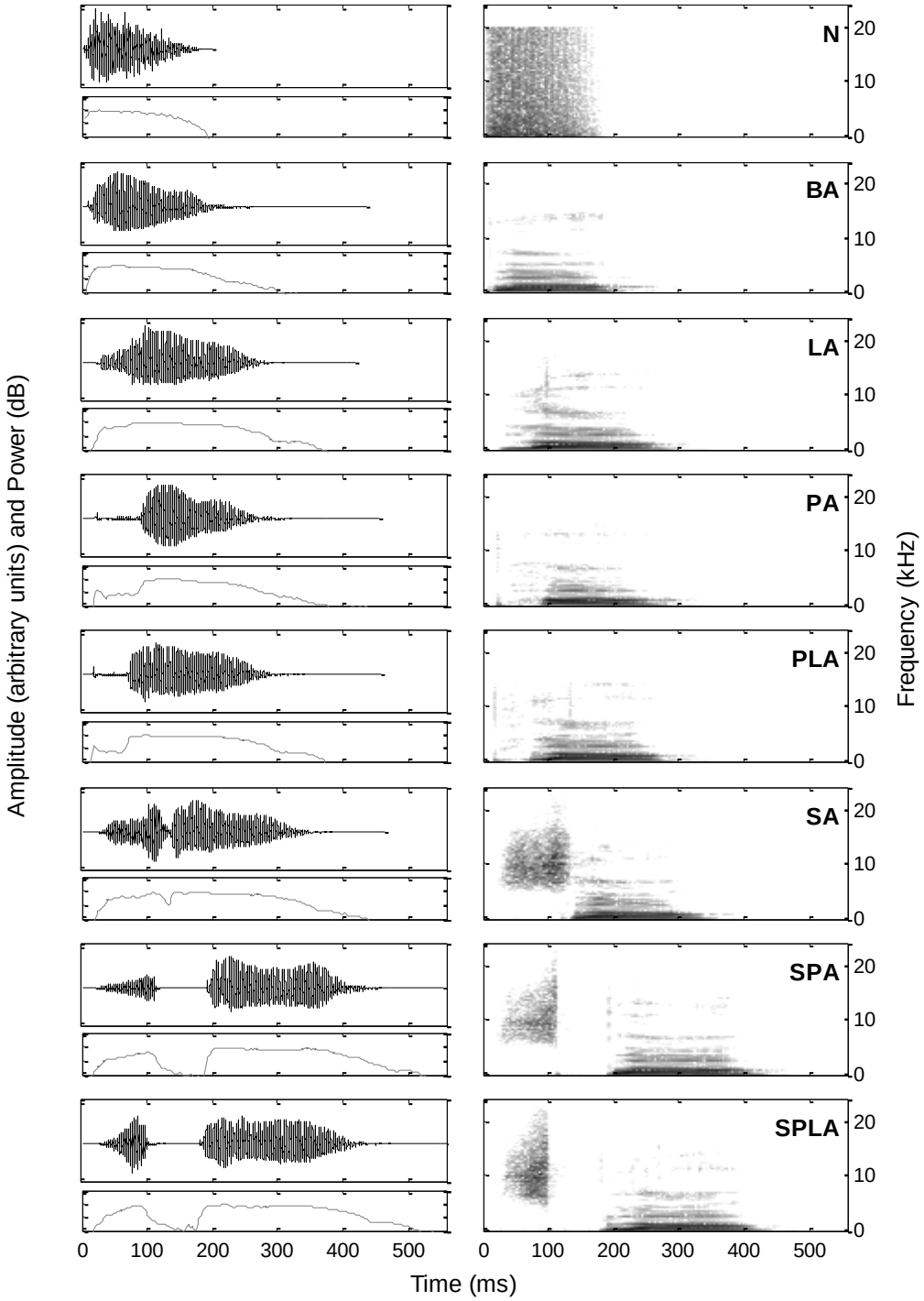


Figure A.5 Waveforms, short term power, and spectrograms for monosyllables and the reference noise. The monosyllables (BA, LA, PA, PLA, SA, SPA, SPLA) were produced by a female speaker and the reference noise (N) was a 1:1 mix of pink noise and pink harmonic spectrum.

# Appendix B

# Experimental equipment and software

This appendix briefly describes the main equipment and software developed for this thesis.

Software used to execute the PCR method of P-centre measurement was not developed as part of this thesis and is not detailed here as a consequence.

## B.1 Rhythm adjustment software

Software for executing the rhythm adjustment experiment was developed in Java and Jython (*Jython* 2006) so that it would be portable between operating systems, though in practice it was only ever used on Microsoft Windows XP based system.

High performance features such as the visual scroll wheel user interface element (see Figure B.1) and the real time sound adjustment and mixing subsystem were implemented in Java. The visual scroll wheel accurately modelled previous hardware based interfaces for the adjustment paradigm. Specifically it was textured so that movement could be clearly perceived but with a pattern that prevented participants associating its position with specific adjustment deviations. Adjustments could be made by "grabbing" the visual scroll wheel with the mouse, by rotating the physical scroll wheel on a scroll mouse, or by the arrow and page up/page down keys.

The main experiment flow and pre-processing of result data in preparation for analysis by statistical software were both implemented in Jython (a Java integrated variant of the scripting language Python). This approach, allowed changes in experiment design to be implemented more quickly and, it was hoped, would eventually allow the software to be usable by other researchers. Figure B.1 shows the appearance of the software in use .

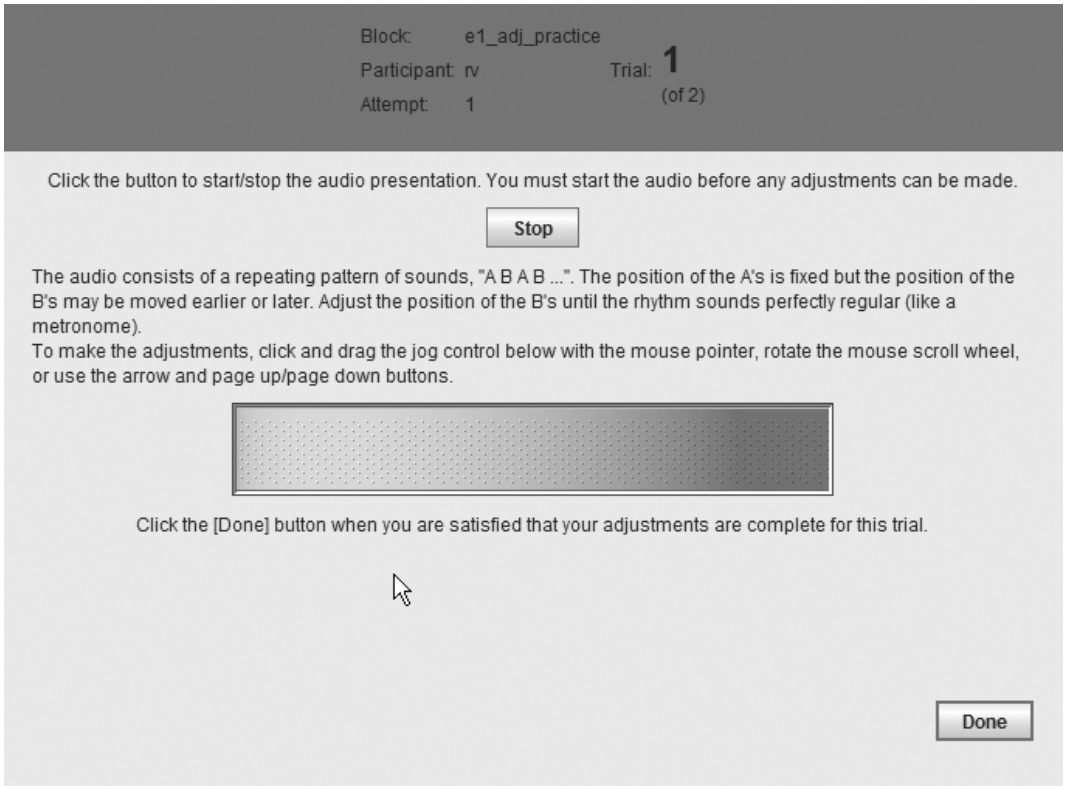


Figure B.1 The main screen of the adjustment software while a trial is running. As a sequence is currently being presented the start/stop button indicates that the next press will stop playback.

## B.2 Tap asynchrony equipment and software

The tap asynchrony measurement method relies on accurate synchronisation of the presented sound sequence to the elicited tap responses. Because of the scheduling algorithms adopted by most non-real-time operating systems (including Windows XP), and the necessity for buffers in the path between a software application and hardware such as a sound card, the delay between the software considering that it had started

sound presentation and that sound actually being audible could easily reach 50 ms or more. The tap response could undergo a similar delay on the input path to the computer.

To overcome these delays two sound cards were used, connected as shown in Figure B.2. (Theoretically a single full-duplex sound card could be used.) The experiment software generates a digital mono audio sample stream which is communicated over USB to the output sound card. This card performs the digital to analogue conversion and the analogue audio signal appears simultaneously on both the Line Out and Headphone Out ports. (Although the sound card outputs a 2 channel stereo, this is a diotic signal—the same mono signal is being presented at both ears.)

The second, input sound card receives two line level "audio" inputs, one from the output sound card and one from the tap detector. The input sound card digitises both channels in precise synchronization and then communicates the digital sample stream via USB to the computer where it is buffered and delayed before eventually being read by the software application and stored to disk as wav (audio) file. Because the tap signal and audio signal were digitised together, however, there can be no possibility that the tap signal is not synchronised with the audio presentation.

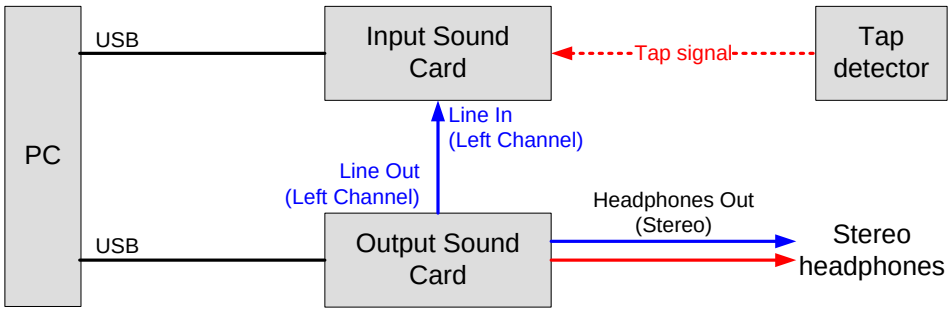


Figure B.2 Schematic illustration of equipment used to implement the tap asynchrony P-centre measurement method.

The software application for executing the tap asynchrony method was implemented in Java and Jython and during a running trial it appeared as in

Figure B.3. This software managed the main experiment flow for each block of trials and initial preprocessing of results.

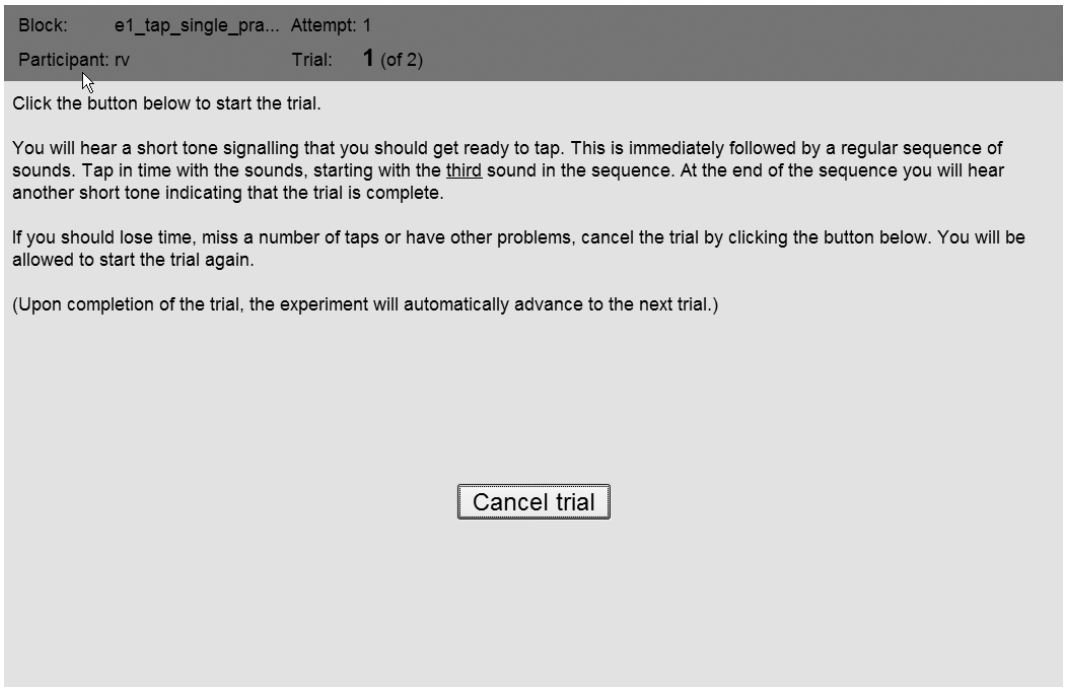


Figure B.3 Main screen of the tap asynchrony software during a running trial. If the participant cancels the trial, or before they start a trial the "cancel trial" button reads "start trial" instead.

Because synchronised audio and tap signals were recorded as audio files, a marker tone was presented at the start and end of each trial so that the audio signal timing could be established without reference to each individual stimulus presentation. Further signal processing was required to determine the moment of tap contact but this processing involved a number of heuristics which were specific to exact tap detector circuit used and are not included for that reason.

## B.3 Bootstrap resampling with replacement (PCR method)

The PCR method of P-centre measurement uses linear regression and transformation of the regression coefficients to solve for the P-centre estimate. Because of this it is not possible to estimate a within-participant standard deviation (or standard error) directly. The standard error of the regression estimate can be obtained but is not easily transformed in a standard error for the P-centre estimate. As a results a bootstrap resampling with replacement method was implemented to estimate this standard error. The following MATLAB code is the specific implementation used.

```
function bootstrap_data;
%% PART 1: processing the original data
% load the data

% naming scheme: i indicates an index. As a prefix, it suggests a
% vector or value that only takes on unique values in a range
% (e.g. when looping). As an underscore suffix, it is typically a
% long vector of "lookup" indexes into a shorter unique vector of
% values.

% d struct gathers all original (and trivially derived) data
% together
[d.subject d.block d.permBlk_i d.rep d.eos_base, ...
        d.sndA d.sndB d.permName d.comboName d.comboOrder, ...
        d.sndA_i d.sndB_i d.combo_i d.perm_i ...
        d.eos_onset d.pcr d.subject_i ] = ...
    textread('bruno4_regression_data.txt', ...
        '%s %n %n %n %n %s %s %s %s %n %n %n %n %n %n %n %n', ...
        'headerlines', 1);

d.onset = d.eos_onset - d.eos_base;

% u struct gathers all unique transformations of original
% (non-unique) data vectors together
[u.subject_i, iu] = unique(d.subject_i);
u.subject = d.subject(iu);

[u.permBlk_i, iu] = unique(d.permBlk_i);
u.permBlk_i = u.permBlk_i(1:end-1); % remove 'N_N'

u.pb.permBlk_i = d.permBlk_i(iu);
u.pb.block = d.block(iu);
u.pb.perm_i = d.perm_i(iu);
u.pb.permName = d.permName(iu);
u.pb.combo_i = d.combo_i(iu);
```

```
u.pb.comboOrder = d.comboOrder(iu);
u.pb.comboName = d.comboName(iu);

[u.combo_i, iu] = unique(d.combo_i);
u.combo_i = u.combo_i(1:end-1);
u.comboName = d.comboName(iu);

u.c.combo_i = d.combo_i(iu);
u.c.comboName = d.comboName(iu);
for i=1:length(iu)
    u.c.perm_i{i} = ...
        unique(u.pb.perm_i(u.pb.combo_i == u.c.combo_i(i)));
    u.c.permBlk_i{i} = ...
        u.pb.permBlk_i(u.pb.combo_i == u.c.combo_i(i));
end

u.comboOrderStr = {'Fwd','Rev'};

% eosbase = the unique base EOS levels and eos_i = the vector of
% indexes into the unique levels
[u.eos, junk, d.eos_i] = unique(d.eos_base);
u.eos6 = [-50 -30 -10 10 30 50];

% FOR TESTING ONLY
% u.subject_i = u.subject_i(2);
% u.combo_i = u.combo_i([2,6]);
% % u.combo_i = u.combo_i([1,6]);
% u.permBlk_i = u.pb.permBlk_i(ismember(u.pb.combo_i, u.combo_i));

u.n_subj = length(u.subject_i);

% reindex the data by subject, permuted pair, and session

[wsp,wsp_index] = summarize_wsp(d,u);
[wsc,wsc_index] = summarize_wsc(d,u,wsp);
[bs] = summarize_bs(d,u,wsc);

% bootstrapping
bootstrap_wsc_se(u,wsp,wsp_index,wsc,wsc_index);




%% SUMMARIZE_WSP ------------------------------------------------
% Summarize data within each subject and permutation of sounds
% (i.e. order/sound roles are important). The RPC6 value is based
% on using only 6 of the 11 EOS levels (see method comparison
% section of thesis for details)
function [wsp, wsp_index] = summarize_wsp(d,u)

fprintf('WITHIN SUBJECT+PERM:\n');
str = sprintf(['Subject | Block | ComboName | ComboPerm | '...
    'PcrX | PcrS | R | SEE | RPC | RPC6\n']);
% str = strrep(str, ' | ', '\t');
fprintf(str);
```

```
sp = 0;
for ius=1:length(u.subject_i)
    for ipb=1:length(u.permBlk_i)
        sp = sp+1;
        s = u.subject_i(ius);
        p = u.permBlk_i(ipb);

        k = find((d.subject_i == s) & (d.permBlk_i == p));
        k1 = k(1);

        subject_i = d.subject_i(k1);
        permBlk_i = d.permBlk_i(k1);
        combo_i = d.combo_i(k1);

        subject = d.subject{k1};
        block = d.block(k1);
        permName = d.permName{k1};
        comboName = d.comboName{k1};
        comboOrder = d.comboOrder(k1);
        onset = d.onset(k1);

        % there will be 11 unique EOS levels. The eos_i field
        % indexes into those unique levels, such that we can
        % recover the vector of EOS levels corresponding to the
        % vector of pcr values

        eos_i = d.eos_i(k);
        eos = u.eos(eos_i);
        pcr = d.pcr(k);

        [rpc, reg.b_const, reg.b_slope, reg.r, reg.seest] = ...
            calc_rpc(eos + onset, pcr);

        i6 = find(ismember(eos, u.eos6));
        rpc6 = calc_rpc(eos(i6) + onset, pcr(i6));

        % negate RPCs for reverse
        if (comboOrder == 2)
            rpc = -rpc;
            rpc6 = -rpc6;
        end

        str = sprintf(['%s | %d | %s | %s | ' ...
            '%.2f | %.2f | %.2f | %.2f | %.2f | %.2f\n'],...
            subject, block, comboName, ...
            u.comboOrderStr{comboOrder},...
            reg.b_const, reg.b_slope, reg.r, ...
            reg.seest, rpc, rpc6);
        % str = strrep(str, ' | ', '\t');
        fprintf(str);

        wsp_index(s,p) = sp;
        wsp(sp) = struct('subject_i',subject_i,...
            'permBlk_i',permBlk_i,...
            'combo_i',combo_i,...
            'subject',subject, ...
            'block',block,...
            'permName',permName,...
```

```
            'comboName',comboName,...
            'comboOrder',comboOrder,...
            'onset',onset,...
            'eos_i',eos_i,...
            'pcr',pcr,...
            'reg',reg,...
            'rpc',rpc,...
            'rpc6',rpc6);
    end
end


%% SUMMARIZE_WSC ------------------------------------------------
% summarize the data within each subject and combination of
% sounds (i.e. order/sound roles unimportant)
function [wsc, wsc_index] = summarize_wsc(d,u,wsp)

% calculate within-subject rpc and rpc6
fprintf('\n\nWITHIN SUBJECT+COMBO:\n');
fprintf('Subject | Combo | rpc | rpc6\n');
sc = 0;
for ius=1:length(u.subject_i)
    for iuc=1:length(u.combo_i)
        sc = sc+1;
        s = u.subject_i(ius);
        c = u.combo_i(iuc);

        sp = find([wsp.subject_i] == s & [wsp.combo_i] == c);
        sp1 = sp(1);

        wsc_index(s,c) = sc;
        wsc(sc).subject_i = s;
        wsc(sc).subject = wsp(sp1).subject;
        wsc(sc).combo_i = c;
        wsc(sc).comboName = wsp(sp1).comboName;
        wsc(sc).rpc = mean([wsp(sp).rpc]);
        wsc(sc).rpc6 = mean([wsp(sp).rpc6]);


        str = sprintf('%s | %s | %.2f | %.2f\n',...
            wsc(sc).subject, wsc(sc).comboName, ...
            wsc(sc).rpc, wsc(sc).rpc6);
        % str = strrep(str, ' | ', '\t');
        fprintf(str);
    end
end


%% SUMMARIZE_BS
% summarize data between subjects. The prefix wc_ means
% within-combination [of sounds].
function bs = summarize_bs(d,u,wsc)
% calculate between participant mean, SD, and SE of rpc and rpc6

fprintf('\n\nBETWEEN SUBJECT, WITHIN COMBO:\n');
fprintf(['Combo | rpc | rpc6 | sd_rpc | '...
    'sd_rpc6 | se_rpc | se_rpc6\n']);
i = 0;
```

```
for iuc=1:length(u.combo_i)
    i = i+1;
    c = u.combo_i(iuc);

    wc_index(c) = i;
    wc(i).comboName = u.comboName{iuc};

    sc = find([wsc.combo_i] == c);

    rpc = [wsc(sc).rpc];
    wc(i).rpc = mean(rpc);
    wc(i).sd_rpc = std(rpc);
    wc(i).se_rpc = wc(i).sd_rpc / sqrt(u.n_subj);

    rpc6 = [wsc(sc).rpc6];
    wc(i).rpc6 = mean(rpc6);
    wc(i).sd_rpc6 = std(rpc6);
    wc(i).se_rpc6 = wc(i).sd_rpc6 / sqrt(u.n_subj);

    str = sprintf(['%s |%.2f | %.2f | %.2f | '...
        '%.2f | %.2f | %.2f\n'],...
        wc(i).comboName, wc(i).rpc, wc(i).rpc6, ...
        wc(i).sd_rpc, wc(i).sd_rpc6, ...
        wc(i).se_rpc, wc(i).se_rpc6);
    % str = strrep(str, ' | ', '\t');
    fprintf(str);
end

bs.wc_index = wc_index;
bs.wc = wc;

% and finally the between participant values averaged across all
% sounds

bs.sd_rpc = mean([wc.sd_rpc]);
bs.se_rpc = bs.sd_rpc / sqrt(u.n_subj);
bs.sd_rpc6 = mean([wc.sd_rpc6]);
bs.se_rpc6 = bs.sd_rpc6 / sqrt(u.n_subj);

fprintf('\n\nBETWEEN SUBJECT AVERAGES\n');
fprintf('sd_rpc | sd_rpc6 | se_rpc | se_rpc6\n');
str = sprintf('%.2f | %.2f | %.2f | %.2f\n',...
    bs.sd_rpc, bs.sd_rpc6, bs.se_rpc, bs.se_rpc6);
% s = strrep(str, ' | ', '\t');
fprintf(str);




%% PART 2: bootstrapping
% bootstrap resampling with replacement to esimate the average
% within-subject+combo SE of RPC derived from PCR regression
%
% The distribution of RPC estimates must be calculated
% individually for each combo because RPCs of different combos
% come from different populations and should not be mixed.
%
```

```
% Although the distribution of RPC estimates for a given combo
% are drawn from the same population (of all possible subject
% estimates), this is the between-subject distribution and not
% the distribution we want to estimate. We are trying to estimate
% the within-subject+combo distribution, therefore we must calc
% the mean and SD of RPC within each subject and combo
% individually.
%
% We will do 2 stages of resampling:
%
% First, for each subject+combo, we resample B1 times and
% calculate the RPC. Each resample mimics the original
% experiment, that is, 2 permutations x 5 trials each x 11 levels
% of EOS. Each permutation (5 x 11 data points) is regressed
% separately, then combined, by taking the mean and sign
% correcting, into a single WSC RPC estimate. The SD of these
% estimates is taken to be a WSC estimate of the SE of RPC.
%
% Second, we take the WSC SE RPC values calculated in the first
% step and resample B2 times (thus resampling the relative
% contributions of individual subjects and combos). We calculate
% the average SE for each resample. The mean of these averages is
% taken to be the bootstrapped estimate of the average WSC SE RPC

function bootstrap_wsc_se(u,wsp,wsp_index,wsc,wsc_index)

n_eosLevels = length(u.eos);
n_levelReps = 5; % 5 reps of each level
n_regress = n_eosLevels * n_levelReps;
n_resample1 = 1000; % stage1
n_resample2 = 100; % stage2

% do linear regression for resampled pcrs in each subject
fout = fopen('bootstrap_output.txt','w');

fprintf('\n\nBOOTSTRAP RESAMPLING, WITHIN SUBJECT+COMBO:\n');
str = ['Subject | Combo | ' ...
    'rpc | rs_rpc | rpc6 | rs_rpc6 | '...
    'rs_sd_rpc | rs_sd_rpc6\n'];
fprintf(['\n' str]);
fprintf(fout, strrep(str, ' | ', '\t'));


rs_eos = repmat(u.eos(:), 1, n_levelReps);
rs_pcr = zeros(n_eosLevels, n_levelReps);

i6 = find(ismember(u.eos, u.eos6));

plot_rs = 0;

% stage 1 resampling, for each subject+combo
wsc_sd_rpc = [];
wsc_sd_rpc6 = [];
for ius=1:length(u.subject_i)
    for iuc=1:length(u.combo_i)
        s = u.subject_i(ius);
        c = u.combo_i(iuc);
        sc = wsc_index(s,c);
```

```
permBlk_i = u.c.permBlk_i{u.c.combo_i == c};
n_pb = length(permBlk_i);

if plot_rs
    figure;
end

% group the pcr values for each permBlk, EOS level
pcr = cell(length(permBlk_i), length(u.eos));
for ipb=1:n_pb
    sp = wsp_index(s, permBlk_i(ipb));
    for ieos=1:length(u.eos)
        pcr{ipb,ieos} = ...
            wsp(sp).pcr(wsp(sp).eos_i == ieos);
    end

    if plot_rs
        subplot(n_pb,1,ipb);
        scatter(u.eos(wsp(sp).eos_i) + ...
            wsp(sp).onset, wsp(sp).pcr, '.');
    end
end

% do the resampling - each resample should approximate a
% genuine experimental run
wsp_rpc = zeros(size(permBlk_i));
wsp_rpc6 = zeros(size(permBlk_i));
for irs=1:n_resample1
    % calculate ordered-RPC for each of the 2 pair
    % permutations
    for ipb=1:n_pb
        p = permBlk_i(ipb);
        sp = wsp_index(s,p);

        % resample with replacement 5 times from each of
        % the 11 EOS levels
        for ieos=1:length(u.eos)
            % last arg=true => sample with replacement
            rs_pcr(ieos,:) = ...
                resample_replace(pcr{ipb,ieos}, ...
                n_levelReps)';
        end

        % sign correct reversed RPCs
        if (wsp(sp).comboOrder == 1)
            signcorr = 1;
        else
            signcorr = -1;
        end

        % sometimes the RPC estimate is garbage because
        % the regression slope is nearly flat or the
        % goodness of fit is bad. We dump such estimates.
        [rpc,b_const,b_slope,r,seest] = ...
            calc_rpc(rs_eos(:) + ...
            wsp(sp).onset, rs_pcr(:));

        if (b_slope < 0.05) || (r < 0.01)
```

```
                    rpc = NaN;
                end

                if plot_rs
                    subplot(n_pb,1,ipb);
                    hold on;
                    scatter(rs_eos(:) + ...
                        wsp(sp).onset, rs_pcr(:));
                    eos_onset = u.eos + wsp(sp).onset;
                    plot(eos_onset, ...
                        polyval([b_slope,b_const],eos_onset), ...
                        'r-');
                    % rpc = -pcr_x_intercept
                    plot([-rpc -rpc], [-40 40], 'k-');
                    hold off;
                end

                wsp_rpc(ipb) = signcorr * rpc;

                [rpc6,b_const,b_slope,r,seest] = ...
                    calc_rpc(cv(rs_eos(i6,:)) + ...
                    wsp(sp).onset, cv(rs_pcr(i6,:)));

                if (b_slope < 0.05) || (r < 0.01)
                    rpc6 = NaN;
                end
                wsp_rpc6(ipb) = signcorr * rpc6;
            end

            % combine sign-corrected order-specific RPCs
            wsc_rpc(irs) = nanmean(wsp_rpc);
            wsc_rpc6(irs) = nanmean(wsp_rpc6);
        end
        wsc_sd_rpc(end+1) = std(getfinite(wsc_rpc));
        wsc_sd_rpc6(end+1) = std(getfinite(wsc_rpc6));

        str = sprintf(['%s | %s | %.2f | %.2f | '...
            '%.2f | %.2f | %.2f | %.2f\n'],...
            wsc(sc).subject, wsc(sc).comboName,...
            wsc(sc).rpc, mean(wsc_rpc), ...
            wsc(sc).rpc6, mean(wsc_rpc6),...
            wsc_sd_rpc(end), wsc_sd_rpc6(end));
        fprintf(str);
        fprintf(fout, strrep(str, ' | ', '\t'));

    end
end

fclose(fout);

% stage 2 resampling: random selection of subjects and combos for
% final averaging

sdrpc = wsc_sd_rpc(:);
sel = logical(zoutlier(sdrpc, 3.29) & (sdrpc < 200));
fprintf('\nSD RPC: stage 1 included = %d, excluded = %d\n', ...
    sum(sel), sum(~sel));
sdrpc = sdrpc(sel);
```

```
sdrpc6 = wsc_sd_rpc6(:);
sel = logical(zoutlier(sdrpc6, 3.29) & (sdrpc6 < 200));
fprintf('SD RPC6: stage 1 included = %d, excluded = %d\n', ...
    sum(sel), sum(~sel));
sdrpc6 = sdrpc6(sel);

for i=1:n_resample2
    % take the average of the resampled set of WSC SD's
    avg_wssd(i) = mean(resample_replace(sdrpc, n_resample1));
    avg_wssd6(i) = mean(resample_replace(sdrpc6, n_resample1));
end

fprintf('\nResampled sd_rpc=%.2f, sd_rpc6=%.2f\n\n',...
    mean(avg_wssd), mean(avg_wssd6));


% TBD: need to add in scatter plot tests which show original
% data, resampled data, all regression lines & RPC intercepts,
% and proves that resampling is working



%% REGRESS ------------------------------------------------------
function [b, r, seest] = regress(x,y)

x = x(:);
y = y(:);
n = length(x);

A = [ones(size(x)) x];
b = pinv(A) * y;

yhat = b(1) + b(2)*x;
SSR = sum((y - yhat).^2);
SST = sum((y - mean(y)).^2);

r = sqrt(1 - SSR/SST);
seest = sqrt(SSR / (n - 2));


%% CALC_RPC -----------------------------------------------------
function [rpc, b_const, b_slope, r, seest] = calc_rpc(x, y)

[b, r, seest] = regress(x,y);
b_const = b(1);
b_slope = b(2);

rpc = b_const / b_slope;



%% ZSCORE
function z = zscore(x)
z = (x - mean(x)) / std(x);
```

```
%% ZOUTLIER
function b = zoutlier(x, threshold)
if length(x) == 1
    b = 1;
else
    b = zscore(x) < threshold;
end

%% RV (make a row vector)
function v = rv(x)
v = x(:)';

%% CV (make a column vector)
function v = cv(x)
v = x(:);

%% NANMEAN
function m = nanmean(x, varargin)
x = x(isfinite(x));
if isempty(x)
    m = NaN;
else
    m = mean(x, varargin{:});
end

%% GETFINITE
function y = getfinite(x)
y = x(isfinite(x));
```

## B.4 EEG equipment configuration

The general equipment configuration used for neuroelectric measurements is illustrated in Figure B.4. A trigger signal (a short square pulse) was embedded on the left audio channel of a 2 channel (stereo) audio stream. The audio channel containing the trigger signal was routed to a trigger device, a simple threshold circuit which would then output a TTL signal pulse to either trigger recording with the Biopac MP100 system, or simply mark the start of an epoch with the BrainVision QuickAmp system.

For presentation to the participant only the right channel (which contained the audio signal) was routed to the earphone.

Audio files whose duration was equal to one epoch were created with trigger signal embedded. An EEG run consisted of many (typically 500) epochs and audio presentation for this was achieved simply by opening the

audio file in Goldwave (*Goldwave*) and playing it repeatedly in a loop. (Pilot tests had indicated that looping the sound file did not cause timing errors.)

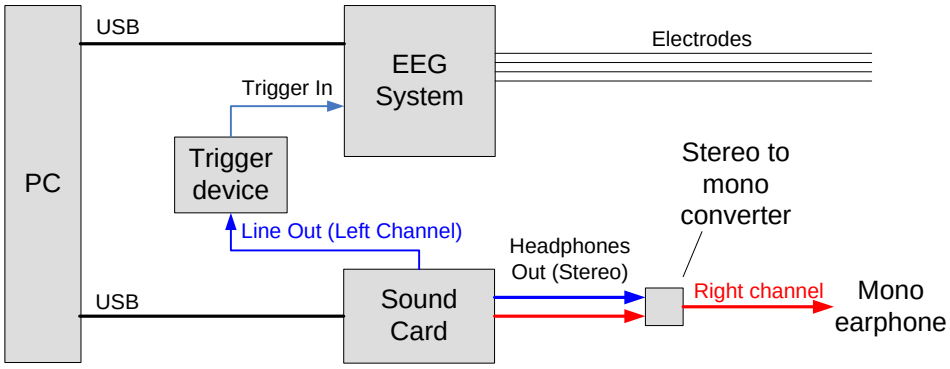


Figure B.4 Schematic layout of equipment used to measure EEG (AEP) signals. The layout shown was used with the Biopac MP100 system. With the BrainVision QuickAmp system the EEG system and sound card were connected to different PCs, but the layout was otherwise identical.

# Appendix C

# P-centre model code listings

## C.1 Marcus and Rapp-Holmgren

```
%%PC_MARCUS_RAPP
%
%  Calculate the single P-centre of an acoustic signal according
%  to the model proposed in Marcus, S.M. (1981)
%
%  [pc_ms, info] = pc_marcus_rapp(x,fs,options...)
%
%  pc_ms:      the calculated P-centre in ms
%  info:       optional internal values from the model
%  x:          the signal
%  fs:         sampling frequency of the signal
%  options:    optional name, value pairs
%
%    'threshold_db'
%               specifies the intensity threshold that indicates
%               physical onset/offset of the signal. The default
%               is -30dB relative to peak short term power.
%
%    'threshold_type'
%               specifies whether the intensity threshold is
%               'relative' [default] or 'absolute'.
%
%               For absolute threshold, the amplitude **MUST** be
%               calibrated to an RMS amplitude reference of 1. A
%               calibrated 0 dB sine wave will peak at approx.
%               1.414 and have an RMS value of 1. See
%               RECALIBRATE_AMPLITUDE.
%
%    'params'
%               A triplet specifying the values of alpha, beta and
%               k in the equation: pc = alpha * x + beta * y + k.
%               The values used for Marcus own model are [0.65,
%               0.25, 0] and those for Marcus's version of Rapp's
%               model are [0.5, 0, 0]. Alternatively this value
%               may be specified as a string where 'rapp' results
%               in the values for rapp's model being used.

% Created:  Rudi Villing Modified: Rudi Villing (30/09/2009),
% changed default sampling rate from
%           10 kHz to 20 kHz based on "GENERAL METHOD" section of
%           Marcus (1981)

function [pc_ms, info] = pc_marcus_rapp(x,fs,varargin);
```

```
o.threshold_db = -30;
o.threshold_type = 'relative';
o.params = [0.65,0.25,0];
o.fs = 20000;
o.Nfft = 512;
o.midband_f = [500 1500];
o.increment_calc = 'after_db'; % otherwise 'before_db'

o = getopt_name(varargin, o);


if ischar(o.params)
    if strcmpi(o.params, 'rapp')
        o.params = [0.5 0 0];
    else
        o.params = [0.65, 0.25, 0];
    end
end

if fs ~= o.fs
    x = resample(x, o.fs, fs);
    fs = o.fs;
end

% STEP 1: calculate spectra every 10ms
% each spectrum is the average over a 10 ms period
Ts_pwr_ms = 10;
Nwin = fix(Ts_pwr_ms * fs/1000);
Nfft = o.Nfft;

frames = buffer(x,Nwin,0,'nodelay'); % each frame is a column
pwr = pwr_fft(frames, boxcar(Nwin), Nfft); % each pwr is a column
total_pwr_db = pwr2db(sum(pwr,1));

f_bins = [0:Nfft/2] * fs/Nfft;
% centre of each frame
pwr_ms = Ts_pwr_ms * (0.5 + [0:length(total_pwr_db)-1]);


% STEP 2: identify onset and offset
% onset when total power first exceeds threshold
% offset when total power last exceeds threshold

switch lower(o.threshold_type)
    case 'absolute'
        threshold_pwr_db = o.threshold_db;
    otherwise
        threshold_pwr_db = max(total_pwr_db) + o.threshold_db;
end

i_valid = find(total_pwr_db > threshold_pwr_db);
if length(i_valid)<2
    error(['total power per 10ms period is less than ' ...
        'specified threshold/default threshold']);
end
i_valid = i_valid(1):i_valid(end);
```

```
onset_ms = pwr_ms(i_valid(1));
offset_ms = pwr_ms(i_valid(end));

% STEP 3: identify vowel onset
% Vowel is the peak dB increment in midband power.

% energy and energy increments in mid band
k_midband = ...
    find(o.midband_f(1) <= f_bins & f_bins < o.midband_f(2));
midband_pwr = sum(pwr(k_midband,:),1);
midband_pwr_db = pwr2db(midband_pwr); % dB power

if strcmpi('after_db',o.increment_calc)
    % increment in log power (max = max relative increment)
    midband_pwr_db_incr = [0, diff(midband_pwr_db)];

    % mid band power increments only within valid range
    [tmp, i_vowel] = max(midband_pwr_db_incr(i_valid));
    i_vowel = i_valid(1) + i_vowel - 1;
else % 'before_db'
    % increment in linear power (max = max abs increment)
    midband_pwr_incr = [0, diff(midband_pwr)];

    % mid band power increments only within valid range
    [tmp, i_vowel] = max(midband_pwr_incr(i_valid));
    i_vowel = i_valid(1) + i_vowel - 1;
end

vowel_ms = pwr_ms(i_vowel);


% model

xx = vowel_ms - onset_ms;
yy = offset_ms - vowel_ms;

alpha = o.params(1);
beta = o.params(2);
k = o.params(3);

pc_ms = onset_ms + alpha * xx + beta * yy + k;


if nargout==2
    info.params = o.params;
    info.increment_calc = o.increment_calc;

    info.pwr_db = pwr2db(pwr);
    info.pwr_ms = pwr_ms;
    info.pwr_f = f_bins;
    info.midband_f = o.midband_f;
    info.total_pwr_db = total_pwr_db;
    info.threshold_pwr_db = threshold_pwr_db;
    info.onset_ms = onset_ms;
    info.offset_ms = offset_ms;
    info.midband_pwr_db = midband_pwr_db;
```

```
    info.midband_pwr_db_incr = midband_pwr_db_incr;
    info.vowel_ms = vowel_ms;
end


%% PWRDB ------------------------------------------------------
function db = pwr2db(px)
db = 10*log10(max(px,1e-20));

%% PWR_FFT ----------------------------------------------------
function sspwr = pwr_fft(x,w,N);

Nw = length(w);
XW = fft(x.*repmat(w,1,size(x,2)), N);
% matches power in time domain xw
dspwr = 1/(N * Nw) * abs(XW).^2;

% correct for effect of window (gain and spectral leakage) to
% match power in time domain x (in the average sense)
G = mean(w);
B = mean(w.^2) / mean(w)^2;
dspwr = (1/B) * (1/G)^2 * dspwr;

% convert to single sided power
sspwr = [dspwr(1,:); 2*dspwr(2:N/2,:); dspwr(N/2+1,:)];
```

## C.2 Vos and Rasch

```
%% PC_VOSRASCH
%   calculate the P-Centre of a sound using the model described
%   in Vos, J., Rasch, R., "The perceptual onset of musical
%   tones", Perception and Psychophysics, 1981, vol 29, pp
%   323-335
%
%   [pc_ms, info] = pc_vosrasch(x, fs, options...);
%
%   pc_ms:     the calculated p-centre (in ms)
%
%   info:      optional internal values from model
%
%   x:         signal for which to calculate p-centre (assumes
%              signal contains just one p-centre).
%
%              The amplitude **MUST** be calibrated to an RMS
%              amplitude reference of 1. A calibrated 0 dB sine
%              wave will peak at approx. 1.414 and have an RMS
%              value of 1. See RECALIBRATE_AMPLITUDE.
%
%   fs:        sampling frequency
%
%   options:    optional name value pairs
%
%     'level_db'
%              the peak RMS level of the signal in dB. This level
%              (combined with the masker level) is used to choose
%              the appropriate relative threshold.
```

```
%
%               If the level is not specified it is estimated as
%               the peak level of the exponentially averaged
%               (default tau = 125 ms) RMS amplitude.
%
%
%     'masker_db'
%               the level of the masker in dB [default = 0]. This
%               value is combined with the level_db to choose the
%               appropriate relative threshold.
%
%     'tau_ms'
%               the exponential average time constant for initial
%               RMS level calculation


function [pc_ms, info] = pc_vosrasch(x, fs, varargin)

o.tau_ms = 125; % integrator time constant (secs)
o.level_db = NaN;
o.masker_db = 0;
o = getopt_name(varargin, o);

x = x(:)';
if isnan(o.level_db)
    max_rms = max(rms_integrator(x,fs,o.tau_ms));
    o.level_db = rms2db(max_rms);
end

% See Vos & Rasch, Figure 5 and consult numerical results for
% each experiment. The best fit to these data is used to
% calculate the relative threshold (see pc_vosrasch_threshold for
% raw data and regression).
level_above_threshold_db = (o.level_db - o.masker_db);
relative_thresh_db = -3.18 - 0.17 * level_above_threshold_db;

% low pass filter the channels to get the channel envelope
% 2nd order low pass filter at 25Hz
[b_env a_env] = butter(2,100/(fs/2));
env = filter(b_env, a_env, abs(x));

% the low pass filtered envelope roughly approximates an RMS
% value - actually it is usually somewhat less
env_db = rms2db(env);

% peak_db = o.level_db;
peak_db = max(env_db);
threshold_db = peak_db + relative_thresh_db;

i_thresh = find(env_db >= threshold_db);
i_thresh = i_thresh(1);

pc_ms = (i_thresh-1) * 1000/fs;

if nargout == 2
    info.tau_ms = o.tau_ms;
    info.masker_db = o.masker_db;
    info.env_db = env_db;
```

```
    info.peak_db = peak_db;
    info.threshold_db = threshold_db;
end



%% RMS2DB -----------------------------------------------------
function db = rms2db(rms);
ref = 1; % rms of 1 => 0 dB
db = 20 * log10(max(rms / ref, 1e-10));

%% RMS_INTEGRATOR ---------------------------------------------
% exponential moving average RMS value
function y = rms_integrator(x,fs,tau_ms);

tau = tau_ms / 1000;
Ts = 1/fs;
alpha = 1 - exp(-Ts/tau);

y = sqrt(filter(alpha, [1 -(1-alpha)], x.^2));
```

## C.3 Gordon

```
%% PC_GORDON
%  Determine P-centre according to models proposed by Gordon.
%
%  Gordon, J.W. (1987), "The perceptual attack time of musical
%    tones", J. Acoust. Soc. Am., 82:88-105
%  Gordon, J.W. (1984), "Perception of attack transients in
%    musical tones", PhD Thesis
%
%  [pc_ms, info] = pc_gordon(x,fs,options...)
%
%  pc_ms:    the calculated pcentre in milliseconds
%
%  info:     optional internal values from the model
%
%  x:        the signal for which p-centre will be calculated. It
%            is assumed that the signal is a sound perceived as
%            having exactly one p-centre.
%
%  fs:       the sampling frequency of the signal
%
%  options:  optional name, value pairs
%
%    'model'
%            the model to use for p-centre calculation. The
%            possible values are: 'time_of_max',
%            'absolute_threshold', 'percent_of_max', 'energy',
%            'normalized_slope', 'normalized_with_rise'. The
%            default value is 'normalized_with_rise', the best
%            performing model as reported by Gordon.
%
%    'env_calc'
%            the envelope type used as input to the p-centre
%            model. The possible values are: 'amplitude',
```

```
%             'power'. The default is 'power'. (A loudness
%             envelope based on the algorithms of Zwicker was also
%             used by Gordon for some models, but is not supported
%             by this implementation.)
%
%  NOTE 1: certain combinations of model and envelope type are
%  incompatible - see Gordon's paper for details.
%
%  NOTE 2: Gordon (1984) describes certain model modifications to
%  handle special cases: (1) when there is more than one slope
%  threshold crossing, the contribution of each crossing is
%  weighted; (2) linear interpolation is used to determine
%  p-centres to more than millisecond accuracy; (3) for impulsive
%  attacks the interpolated p-centre is earlier than the physical
%  onset and is delayed by up 4ms. None of these adjustments was
%  implemented.

function [pc_ms, info] = pc_gordon(x, fs, varargin)

o.model = 'normalized_with_rise';
o.env_calc = 'power';
o = getopt_name(varargin, o);

switch lower(o.env_calc)
    case 'amplitude', env_calc = 'amp';
    case 'power', env_calc = 'pwr';
    otherwise, error('invalid env_calc "%s"', o.env_calc);
end

switch lower(o.model)
    case 'time_of_max', model = 'max';
    case 'absolute_threshold', model = 'abs';
    case 'percent_of_max', model = 'pct';
    case 'energy', model = 'ene';
    case {'normalized_slope', 'normalised_slope'}
        model = 'ns';
    case {'normalized_with_rise', 'normalised_with_rise'}
        model = 'nwr';
    otherwise
        error('invalid model "%s"', o.model);
end

x = x(:)';

% STEP 1: get the envelope

fs_env = 1000;
ts_env_ms = 1000/fs_env; % sampling period in milliseconds
amp_env = envelope(x, fs, fs_env);

switch lower(env_calc)
    case 'amp', env = amp_env;
    case 'pwr', env = amp_env .^ 2;
end

% STEP 2: apply the appropriate model

switch model
```

```
    case 'max'
        imax = find(env == max(env));
        pc_ms = imax(1) * ts_env_ms;

    case 'abs'
        if ~strcmp(env_calc, 'amp')
            error('env_calc "%s" not valid for model "%s"', ...
                o.env_calc, o.model);
        end
        % Gordon states presentation level as 90dBA, we assume
        % 90dBSPL for convenience
        threshold = 10^(90/20) * 0.041;
        ithresh = find(env >= threshold);
        pc_ms = ithresh(1) * ts_env_ms;

    case 'pct'
        if ~strcmp(env_calc, 'amp')
            error('env_calc "%s" not valid for model "%s"', ...
                o.env_calc, o.model);
        end
        relative_threshold = 0.0582;
        threshold = max(env) * relative_threshold;
        ithresh = find(env >= threshold);
        pc_ms = ithresh(1) * ts_env_ms;

    case 'ene'
        switch env_calc
            case 'amp', threshold = 1.2;
            case 'pwr', threshold = 0.03;
            otherwise
                error('env_calc "%s" not valid for model "%s"',
...
                    o.env_calc, o.model);
        end

        energy = cumsum(env);
        ithresh = find(energy >= threshold);
        pc_ms = ithresh(1) * ts_env_ms;

    case 'ns'
        switch env_calc
            case 'pwr'
                threshold = 0.0104; % Gordon 1984, p106
            otherwise
                error('env_calc "%s" not valid for model "%s"',
...
                    o.env_calc, o.model);
        end
        % although Gordon (1987) suggests that normalisation was
        % performed on the envelope prior to calculating slope,
        % no threshold parameters are given for this. The
        % threshold in Gordon (1984) can be applied relative to
        % the maximum slope (this is what Gordon himself appears
        % to do) or as an absolute threshold for the slope
        % normalized to its maximum - do the latter here.
        env_slope = smooth_slope(env);
        % normalize by max slope
        env_slope = env_slope ./ max(env_slope);
        ithresh = find(env_slope >= threshold);
```

```
        pc_ms = (ithresh(1) - 1) * ts_env_ms;

    case 'nwr'
        switch env_calc
            case 'pwr'
                % See Gordon 1987, Figure 15, p104; Gordon 1984,
                % p111
                threshold = 0.36e-3;
                b_risetime = 0.08;
            otherwise
                error(['env_calc "%s" ' ...
                    'not valid for model "%s"]', ...
                    o.env_calc, o.model);
        end
        % normalize envelope (by max envelope) then calculate
        % slope
        env_slope = smooth_slope(env ./ max(env));
        ithresh = find(env_slope >= threshold);

        % rise time is time between start and end of contiguous
        % env_slope >= threshold
        ibegin = ithresh(1);
        iend = find(env_slope < threshold);
        iend = iend(iend > ibegin); % end must be after start
        % just want the first point below threshold
        iend = iend(1);
        t_rise_ms = (iend - ibegin) * ts_env_ms;

        pc_ms = ((ibegin-1) * ts_env_ms) ...
            + (b_risetime * t_rise_ms);
end


if nargout == 2
    info.model = model;
    info.env_calc = env_calc;
    info.env = env;
    info.fs_env = fs_env;

    switch model
        case 'max'; % nothing required here
        case 'abs', info.threshold = threshold;
        case 'pct', info.threshold = threshold;
        case 'ene'
            info.threshold = threshold;
            info.energy = energy;
        case 'ns'
            info.threshold = threshold;
            info.env_slope = env_slope;
        case 'nwr';
            info.threshold = threshold;
            info.env_slope = env_slope;
            info.rise_begin_ms = (ibegin-1) * ts_env_ms;
            info.rise_end_ms = (iend-1) * ts_env_ms;
    end
end

%% SMOOTH_SLOPE ------------------------------------------------
function dx = smooth_slope(x);
```

```
Nwin = 19;
k = (0:Nwin-1)';

% ASSUMPTION WARNING (not specified by Gordon (1987)):
% measure slope at temporal centre of sample points
frames = buffer([zeros(1,9) x zeros(1,9)], ...
    Nwin, Nwin-1, 'nodelay');

dx = zeros(size(x));
for i=1:size(frames,2)
    p = polyfit(k, frames(:,i), 1); % best linear fit
    dx(i) = p(1); % slope in p(1)
end

%% ENVELOPE ------------------------------------------------------
function env = envelope(x,fs,fs_env)

fc = 100;
[b,a] = butter(2,fc/(fs/2));
env1 = filter(b,a,abs(x));
env = resample(env1,fs_env,fs);
env = max(env,0);
```

## C.4 Howell

```
%% PC_HOWELL
%  An implementation inspired by the model architecture proposed
%  by Howell:
%
%  Howell, P 1984, 'An Acoustic Determinant of Perceived and
%    Produced Anisochrony', paper presented to 10th International
%    Congress of Phonetic Sciences, Dordrecht.
%
%  Howell, P 1988, 'Prediction of P-center location from the
%    distribution of energy in the amplitude envelope: I',
%    Perception & Psychophysics, vol. 43, pp. 90-3.
%
%  Howell did not make the model or its parameterization
%  explicit, so the implementation is based on the Appendix, p227
%  of
%
%  Scott, SK 1993, 'P-Centres in speech: an acoustic analysis',
%    Unpublished PhD thesis, University College London.
%
%  pc_ms = pc_howell(x, fs, options...)
%
%  pc:      the calculated p-centre (in secs)
%
%  x:       signal for which to calculate p-centre (assumes
%           signal contains just one p-centre)
%
%  fs:      sampling frequency
%
%  options: optional name value pairs
%
```

```
%    'mass_calc'
%           Specify how to esimate the "mass" that will be used
%           in the centre of mass calculation. Valid choices are
%           'amp' [default] the absolute amplitude or 'energy'
%           the energy signal (x^2).
%
%    'pc_calc'
%           specify the method used to calculate the p-centre.
%           Choices are: 'cofg' [default] the normal centre of
%           gravity calculation, or 'half' where half the total
%           weight is exceed. (The latter was Scott's
%           implementation but it is not generally the same as
%           the centre of gravity).
%
%    'threshold_db'
%           specify the threshold at which the signal onset and
%           offset are identified. The default value is -30 dB
%           relative to signal maximum (similar to Marcus's
%           model). Signal energy is evaluated only between the
%           onset and offset.
%
%    'threshold_type'
%           Specify the type of threshold, either 'relative'
%           [default] or 'absolute'. This threshold is applied to
%           the instantaneous envelope of the signal.
%
%           For absolute threshold, the amplitude **MUST** be
%           calibrated to an RMS amplitude reference of 1. A
%           calibrated 0 dB sine wave will peak at approx. 1.414
%           and have an RMS value of 1. See
%           RECALIBRATE_AMPLITUDE.

function [pc_ms, info] = pc_howell(x, fs, varargin)

o.threshold_db = -30;
o.threshold_type = 'relative';
o.mass_calc = 'amp'; % or 'energy'
o.pc_calc = 'cofg'; % or 'half'
o.lowpass = true;
o = getopt_name(varargin, o);

x = x(:)';

% estimate onset and offset location

[b_env,a_env] = butter(2,50/(fs/2));

env = abs(x); % instantaneous envelope
if o.lowpass
    env = max(filter(b_env, a_env, env), 0);
end
env_db = amp2db(env);

switch lower(o.threshold_type)
    case 'absolute'
        % correct from RMS level to peak level
        threshold_db = o.threshold_db + amp2db(sqrt(2));
    otherwise
        threshold_db = max(env_db) + o.threshold_db;
```

```
end


sel = find(env_db >= threshold_db);
if length(sel)<2
    error(['envelope does not exceed threshold for ' ...
        'long enough - check signal scaling']);
end
% join non-contiguous segments that are above threshold
sel = sel(1):sel(end);

onset_ms = (sel(1)-1) * 1000/fs;
offset_ms = (sel(end)-1) * 1000/fs;


% calculate approximate centre of signal integral

switch lower(o.mass_calc)
    case 'energy'
        xm = abs(x).^2;
        threshold = db2amp(threshold_db) ^ 2;
    otherwise
        xm = abs(x);
        threshold = db2amp(threshold_db);
end

if o.lowpass
    xm = max(filter(b_env,a_env,xm), 0);
end

switch lower(o.pc_calc)
    case {'scott','half'}
        % Scott's original implementation finds the half "mass"
        % point. This is not the same as the centre of mass,
        % because the positions of the two half masses will
        % usually not be equal. Consider for example half the
        % mass spread over positions 1 to 10 and the other half
        % spread over positions 12 to 100. Clearly the true
        % centre of mass will be somewhere within the range 12 to
        % 100 and not at position 11, the half mass point.
        ihalfint = find(cumsum(x2(sel)) >= (sum(x2(sel)) / 2));
        ipc = sel(1) - 1 + ihalfint(1);

    otherwise
        % normal centre of mass calculation
        ipc = sum(sel .* xm(sel)) / sum(xm(sel));
end

pc_ms = ipc * 1000/fs;

if nargout == 2
    info.xm = xm;
    info.threshold = threshold;
    info.onset_ms = onset_ms;
    info.offset_ms = offset_ms;
end

end
```

```
%% DB2AMP ---------------------------------------------------
function amp = db2amp(db)
ref = 1;
amp = ref * 10.^(db/20);
end

%% AMP2DB ---------------------------------------------------
function db = amp2db(amp)
ref = 1;
db = 20 * log10(max(amp/ref,eps));
end
```

## C.5 Pompino-Marschall

### C.5.1 Main code

```
%% PC_PMARSCHALL
%   An implementation of the P-Centre model described in
%
%     [PM89] Pompino-Marschall, B. (1989), "On the
%       psychoactoustic nature of the P-center phenomenon",
%       Journal of Phonetics, 17, 175-192.
%
%   and clarified with the aid of
%
%     [PM07] Pompino-Marschall, B. personal communication, May
%          2007
%     [PM90] Pompino-Marschall, B. (1990), "Die Silbenprosodie.
%          Ein elementarer Aspekt der Wahrnehmung von
%          Sprachrhythmus und Sprechtempo", Tubingen: Niemeyer.
%     [PM91] Pompino-Marschall, B. (1991), "The syllable as a
%          prosodic unit and the so-called P-centre effect",
%          FIPKM 29, 65-123
%     [ZW99] Zwicker, E & Fastl, H 1999, "Psychoacoustics: facts
%          and models", Second updated edn, Springer series in
%          information sciences, Springer, Berlin; New York.
%
%
%   [pc_ms, info] = pc_pmarschall(x,fs,options...)
%
%   pc_ms:      the calculated P-centre (ms)
%   info:       [OPTIONAL] internal intermediate stage data from
%               the model
%   x:          the signal for which a single p-centre will be
%               calculated.
%
%               The signal **MUST** be calibrated to an RMS
%               amplitude reference of 1. A calibrated 0 dB sine
%               wave will peak at approx. 1.414 and have an RMS
%               value of 1. See RECALIBRATE_AMPLITUDE.
%
%   fs:         sampling frequency
```

```
%   options:   optional name, value pairs
%
%     'pc_calc'
%               Which P-centre calculation to use. It can be
%               'onset' [default], or 'cofg'.

% NOTE: there is ambiguity in the specification of what
% constitutes a partial event. For example in one figure, partial
% events on the rising flank may begin from minima or increases
% in rising edge slope. However in a subsequent figure (and
% corresponding fortran code) it appears that subsequent partial
% events on the rising flank nominally start at the previous
% maxima/inflection point and not at any local minimum.

function [pc_ms, info] = pc_pmarschall(x,fs, varargin)

o.pc_calc = 'onset';
o.W_fall_calc = 'PM90'; % otherwise 'PM89', see below
% normally CofG weight is simply the sum of all the weights. The
% calculation in PM90 is different and is the default here.
o.cofgw_calc = 'PM90'; % otherwise 'sum'

o = getopt_name(varargin, o);

Ts_frame_ms = 15; % PM89
tau_ms = 50; % 50ms [PM89, p183]

% Values obtained from [ZW99]
bark.f_lower = [0, 100, 200, 300, 400, 510, 630, 770, ...
    920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, ...
    3700, 4400, 5300, 6400, 7700, 9500, 12000];
bark.f_upper = [100, 200, 300, 400, 510, 630, 770, 920, ...
    1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, ...
    4400, 5300, 6400, 7700, 9500, 12000, 15500];
bark.f_centre = [50, 150, 250, 350, 450, 570, 700, 840, ...
    1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, ...
    4000, 4800, 5800, 7000, 8500, 10500, 13500];

n_cb = 19; % See [PM89, fig. 5]


% STEP 0: convert to fs of 20000 [PM07]
if fs ~= 20000  % [PM07]
    x = resample(x,20000,fs);
    fs = 20000;
end

% IMPLEMENTATION NOTE: Later processing to find partial onsets
% benefits from a smooth beginning to signals, so prepend some
% silence. Compensate by subtracting duration later when
% calculating times.
delay_ms = 60;
x = [zeros(1,delay_ms * fs/1000), x(:)'];

% Step 1: every 15ms an fft is calculated on the signal using
% multiple windows.

t_win = [60, 30, 15] ./ 1000;
```

```
f_band = [ 16, 500, 1500, 5267 ];

Nfft = round(t_win(1) * fs);
pad = zeros(1,Nfft);
f_bin = [0:Nfft/2-1] * fs/Nfft;

for i=1:length(t_win)
    Nw = round(t_win(i) * fs);
    Nadv = round(Ts_frame_ms/1000 * fs);
    Nlap = Nw - Nadv;

    w = hanning(Nw,'periodic'); % [PM07]

    % the windows are aligned at their temporal starts (i.e. each
    % window starts on the first sample) - stated in [PM07],
    % confirmed by code in [PM90]. The result is that lower
    % frequencies, which use longer windows, are **time
    % advanced** relative to high frequencies in the resulting
    % spectrum.

    % Nw x Nframes
    xmat = buffer([x pad(1:Nlap)],Nw,Nlap,'nodelay');

    x1{i} = mag_fft(xmat, w, Nfft);
end

% each row is for one frequency, each column is for one time
% frame Populate with highest temporal resolution/lowest freq
% resolution values by default
x2 = x1{end};
for i=1:length(t_win)
    selected_bins = ...
        find(f_band(i) < f_bin & f_bin <= f_band(i+1));
    x2(selected_bins,:) = x1{i}(selected_bins,:);
end

% STEP 2: pool FFT bins into bark scale critical bands

x3 = avg_per_bark(x2, f_bin, bark, n_cb);
n_frames = size(x3,2);

% STEP 3: linear lowpass filter each critical band

x4 = filter([1 0.0067], 1, x3, [], 2);

% STEP 4: log-linear low pass filter each critical band

x5 = zeros(size(x4));
for cb = 1:n_cb
    x5(cb,1) = 0.15 * x4(cb,1);
    for n = 2:n_frames
        if x4(cb,n) >= x4(cb,n-1)
            x5(cb,n) = 0.15 * x4(cb,n) + 0.85 * x4(cb,n-1);
        else
            x5(cb,n) = x5(cb,n-1) ...
                * exp(0.21 * log(x4(cb,n) / x5(cb,n-1)));
        end
```

```
    end
end

% STEP 5: scale to dB

SILENCE = 1e-5; % Amplitude corresponding to -100dB
x6_db = 20 * log10(max(x5, SILENCE));

% STEP 6: calculate specific loudness in each critical band
% according to Paulus & Zwicker (1972)

zwicker.pres_type = 'free';
% sample the specific loudness every 0.2 bark
zwicker.Z_inc = 0.2;
n_per_bark = 5; % so 5 samples per bark

% the upward spreading of excitation in the loudness function
% allows us to change n_cb here
n_cb = 24;
x7 = zeros(n_cb, n_frames);
for frame=1:n_frames
   [sones, internals] = ...
       loudness_zwicker1972(x6_db(:,frame), zwicker);

   for bark=1:24
       Ndash_cb = ...
           internals.Ndash(1:n_per_bark + (bark-1)*n_per_bark);
       Ndash(bark) = mean(Ndash_cb);
   end
   x7(:,frame) = Ndash';
end

% STEP 7: smooth the specific loudness using simple moving
% average

% compensate for delay so that x8 is centred in averager [PM90,
% p214, Code marked "GLAETTUNG"]
x8 = filter([1 1 1]/3, 1, [x7 zeros(n_cb,1)]);
x8(:,1) = [];

% STEP 8: calculate partial onset and offset events and integrate
% within channels

n_frames = size(x8,2);

% create an array of structures
tmp = cell(1,n_cb);
ch = struct('xmm',tmp,'imm',tmp,'Tmm',tmp,...
    'dLmm',tmp,'Wmm',tmp,...
    'Wr',tmp,'Tr',tmp,...
    'Wf',tmp,'Tf',tmp,...
    'Wp',tmp,'Tp',tmp,...
    'Wra',tmp,'Tra',tmp,...
    'Wpa',tmp,'Tpa',tmp);


for cb=1:n_cb
```

```
    chn = ch(cb);
    chn = find_partial_events(x8(cb,:), Ts_frame_ms, chn);
    chn = integrate_events(chn, tau_ms, o);
    ch(cb) = chn;
end;

% STEP 9: calculate overall syllable onset and the syllable
% centre of gravity

[Wso,Tso] = cofg([ch.Wra], [ch.Tra], o);
[Wscg,Tscg] = cofg([ch.Wpa], [ch.Tpa], o);

switch lower(o.pc_calc)
    case 'onset', pc_ms = Tso;% - delay_ms;
    otherwise, pc_ms = Tscg;% - delay_ms;
end

if nargout==2
    info.pc_calc = o.pc_calc;
    info.W_fall_calc = o.W_fall_calc;
    info.cofgw_calc = o.cofgw_calc;

    info.x = x;
    info.fs = fs;
    info.delay_ms = delay_ms;
    info.Ts_frame_ms = Ts_frame_ms;
    info.n_cb = n_cb;
    info.n_frames = n_frames;
    info.spec_f = f_bin;
    info.fmax = f_band(end);
    info.x2_db = 20*log10(x2);
    info.x3_db = 20*log10(x3);
    info.x4_db = 20*log10(x4);
    info.x5_db = 20*log10(x5);
    info.x6_db = x6_db;
    info.x7 = x7;
    info.x8 = x8;
    info.ch = ch;
    info.Wso = Wso;
    info.Tso = Tso;
    info.Wscg = Wscg;
    info.Tscg = Tscg;
end


%% FIND_PARTIAL_EVENTS ----------------------------------------
% xmm is the start/end values of x for rising/falling edges ((mm)
% notation comes from min/max). imm is the index of the start of
% a rising/falling edge
function ch = find_partial_events(x, Ts_ms, ch)

x=x(:);

% values to be used if no suitable events found
ch.xmm = zeros(0,2);
ch.imm = zeros(0,2);
ch.Tmm = zeros(1,0);
ch.dLmm = zeros(1,0);
```

```
% short circuit if no events possible due to zero loudness range
% in band
if length(x) < 3 || (max(x) - min(x)) == 0; return; end

Lmax = max(x);
min_dL = Lmax * 0.12;
dL_fract = 0.4;

% Sound must start with an onset. If the signal is initially
% increasing then the onset starts immediately, otherwise find
% the first local minimum

dx = [diff(x); 0];
i1 = find(dx > 0);
if isempty(i1); return; end
i1 = i1(1);

% Seek along signal.
% If delta L (dL) between current value and previous extremum
% exceeds min_dL then we have found a partial onset and its end
% will be the next local max. Stay in onset mode until a partial
% offset detected. If dL < -min_dL then we have found a partial
% offset and its end will be the next local min. Stay in offset
% mode until a partial onset detected.

prev_exi = i1;
i = prev_exi;

ievt = 0;
xmm = [];
imm = [];

for i=1:length(x)
    dL = x(i) - x(prev_exi);

    if dL >= min_dL % partial onset detected
        % are we at local max (dx = x(i+1)-x(i) = -ve at max)
        if dx(i) < 0
            ievt = ievt+1;
            xmm(ievt,:) = [x(prev_exi), x(i)];
            imm(ievt,:) = [prev_exi, i];
            prev_exi = i;
        end
    elseif dL <= -min_dL % partial offset detected
        % partial offset ends at local min
        % local min when dx +ve [ dx = x(i+1)-x(i) ]
        if dx(i) > 0
            ievt = ievt+1;
            xmm(ievt,:) = [x(prev_exi), x(i)];
            imm(ievt,:) = [prev_exi, i];
            prev_exi = i;
        end
    elseif x(i) == Lmax % signal max = onset end & offset start
        ievt = ievt+1;
        xmm(ievt,:) = [x(prev_exi), x(i)];
        imm(ievt,:) = [prev_exi, i];
        prev_exi = i;
```

```
    end
end

% tag on final onset/offset if necessary
if (prev_exi ~= length(x)) && (x(prev_exi) ~= x(i))
    ievt = ievt+1;
    xmm(ievt,:) = [x(prev_exi), x(i)];
    imm(ievt,:) = [prev_exi, i];
end

% linear interpolate to find the time and loudness for each
% partial event delimited by xmm
for ievt=1:size(xmm,1)
    dL = xmm(ievt,2) - xmm(ievt,1);
    dLmm(ievt) = dL * dL_fract; % +ve rising, -ve falling

    Tmm(ievt) = ...
        interp_event_time(x, imm(ievt,1), dLmm(ievt), Ts_ms);
end

ch.xmm = xmm;
ch.imm = imm;
ch.Tmm = Tmm;
ch.dLmm = dLmm;


% INTEGRATE_EVENTS ------------------------------------------
% rise, fall, peak (rise followed by fall)
function ch = integrate_events(ch, tau_ms, o)

% integrate partial events separately on rising and falling
% flanks of individual peaks within channel. (Subscript mm =
% min/max, so one per loudness increment, r = integrated in
% rising flank, p = integrated for peak, i.e. rising + falling
% flank).
ch.Wmm = [];
ch.Wr = []; ch.Tr = [];
ch.Wp = []; ch.Tp = [];
ch.Wf = []; ch.Tf = [];
imax = length(ch.Tmm);
i = 1;
while i<=imax
    % find a range of contiguous onsets (the rising flank)
    % followed by a sequence of contiguous offsets (the falling
    % flank). Together this set of events defines a single peak
    % event.
    irise = [];
    while i<=imax && ch.dLmm(i)>0
        irise(end+1) = i;
        i = i+1;
    end
    ifall = [];
    while i<=imax && ch.dLmm(i)<0
        ifall(end+1) = i;
        i = i+1;
    end

    % scale W for onsets by distance from current onset to final
```

```
    % onset on rising flank
    Lrise = ch.dLmm(irise); Trise = ch.Tmm(irise);
    ch.Wmm(irise) = Lrise .* exp(-(Trise(end) - Trise) ./ tau_ms);

    Lfall = ch.dLmm(ifall); Tfall = ch.Tmm(ifall);
    if strcmpi('PM90',o.W_fall_calc)
        % scale W for offsets by distance from first offset to
        % current offset - this is almost exactly the opposite of
        % PM89 Fig. 6
        ch.Wmm(ifall) = ...
            0.5.*Lfall .* exp(-(Tfall - Tfall(1)) ./ tau_ms);
    else % 'PM89'
        % scale W for offsets by distance from current offset to
        % final offset on falling flank
        ch.Wmm(ifall) = ...
            0.5.*Lfall .* exp(-(Tfall(end) - Tfall) ./ tau_ms);
    end

    % integrate all partial events on each rising flank and
    % falling flank to form a single rising and falling event
    [Wr,Tr] = cofg(ch.Wmm(irise), Trise, o);
    [Wf,Tf] = cofg(abs(ch.Wmm(ifall)), Tfall, o);

    % [PM90, p218]
    if strcmpi('PM90',o.W_fall_calc)
        Wf = Wf*exp(-(Tf - Tr) ./ tau_ms);
    end

    % integrate the rising and falling events to form a single
    % peak event
    [Wp,Tp] = cofg([Wr,Wf], [Tr,Tf], o);

    ch.Wr(end+1) = Wr; ch.Tr(end+1) = Tr;
    ch.Wf(end+1) = Wf; ch.Tf(end+1) = Tf;
    ch.Wp(end+1) = Wp; ch.Tp(end+1) = Tp;
end

% integrate the rising events and peak events
% channel integrated rising event
[ch.Wra, ch.Tra] = cofg(ch.Wr, ch.Tr, o);
% channel integrated peak events
[ch.Wpa, ch.Tpa] = cofg(ch.Wp, ch.Tp, o);


%% COFG ------------------------------------------------------
function [cgw,cgt] = cofg(w,t,o)

sumw = sum(w);
sumt = sum(t);
if sumw == 0 || sumt == 0
    cgw = 0; cgt = 0; return;
else
    sumwt = sum(w.*t);
    cgt = sumwt / sumw;

    if strcmpi('PM90',o.cofgw_calc)
        cgw = sumwt / sumt;
    else % normal cofg weight calculation
```

```
        cgw = sumw; % different than PM90 implementation
    end
end

%% INTERP_EVENT_TIME ------------------------------------------
function Ti = interp_event_time(x, i1, dL, Ts_ms)

Nx = length(x);
i = i1;
thr = x(i1) + dL;
if dL > 0
    thcheck = x - thr; % thcheck > 0 iff x > thr
else
    thcheck = -(x - thr); % thcheck > 0 iff x < thr
end

% find threshold crossing
while (i < Nx) && (thcheck(i) <= 0)
    i = i+1;
end

if x(i) == thr % interpolation not required
    Ti = Ts_ms * (i-1);
else % interpolation required
    if i > 1
        ii = (i-1) + (thr - x(i-1)) / (x(i) - x(i-1));
    else
        ii = 1 + (thr / x(i));
    end
    Ti = Ts_ms * (ii - 1);
end


%% MAG_FFT ----------------------------------------------------
% simple magnitude FFT without window compensation (see in code)
function mag = mag_fft(frames,w,N)

Nw = length(w);
for i=1:size(frames,2)
    frames(:,i) = frames(:,i) .* w; % window it
end

% perform FFT, operates on columns by default
XW = fft(frames, N);

% simple approach without compensation for window [PM90]
dspwr = (abs(XW) / N) .^ 2;
pwr = [dspwr(1,:); 2*dspwr(2:N/2,:); dspwr(N/2+1,:)];
mag = sqrt(pwr);


%% AVG_PER_BARK -----------------------------------------------
% frames has one frame per column, each row is a frequency bin
function avg = avg_per_bark(frames, f, bark, n_cb)

avg = zeros(n_cb, size(frames,2));
for i=1:n_cb
   % determine the range in vector fHz which corresponds to a
```

```
    % bark band
    f_in_cb = (bark.f_lower(i) < f & f <= bark.f_upper(i));

    % take the mean intensity in the band
    avg(i,:) = mean(frames(f_in_cb,:),1);
end
```

### C.5.2 Loudness model

```
%% LOUDNESS_ZWICKER1972
% A matlab implementation of the fortran program II published in
% Acustica vol 27, 1972, pp 253-266 by E. Paulus and E. Zwicker.
% This program calculates the specific loudness in bark spaced
% bark bandwidth critical bands.
%
% [sones, info]=SPECIFIC_LOUDNESS_ZWICKER(L_G, fieldType)
%
%   L_G:        level (dB SPL) of each bark band (indicated by
%               subscript G)
%   opt:        [OPTIONAL] configuration options structure, with
%               possible fields as follows:
%
%     .pres_type
%               Type of presentation field. May be 'diffuse'
%               [default] or 'free'
%     .Z_inc
%               Specifies the delta critical band rate sampling
%               to use (in Bark) which also determined the min
%               critical band rate to evaluate

function [sones, info]=loudness_zwicker1972(L_G, opt);

L_G = L_G(:)';

% if not all bark bands have been specified, set unspecified ones
% to -100dB which is effectively silence
if length(L_G) < 24
    L_G((length(L_G)+1):24) = -100;
end

% default options
defaults.pres_type = 'diffuse'; % alternative is free
defaults.Z_inc = 0.2;

if nargin == 2
    opt = getoptstruct(opt, defaults);
else
    opt = defaults;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% L_EHS (in dB)
L_EHS = [42 18.5 11.5 8.3 6.7 5.5 4.8 4.3 repmat(4.0, 1,16)];

% attenuation of the ear, a_0 (in dB)
```

```
a_0 = [repmat(0, 1,10) -0.2 -0.5 -1.2 -2.1 -3.2 -4.6 ...
    -5.5 -5.6 -4.3 -2.5 -0.1 2.8 6.4 20.0];

% delta L_ED (in dB), correction for diffuse sound field
DL_ED = [ 0.0 0.1 0.6 1.0 1.4 2.0 2.5 2.8 3.0 2.3 1.3 ...
    0.0 -0.9 -1.6 -2.0 -1.9 -1.6 -1.0 0.2 2.0 3.4 4.1 4.2 3.5];

% specific loudness lower bound for edge steepness data
% GRENZE means limit, boundary
GRENZE = [23.5 19.0 15.1 11.9 9.0 6.6 4.6 3.2 2.13 ...
    1.36 0.82 0.43 0.21 0.08 0.03 0.0];

% specific loudness spread: slope of upper edge vs. level (rows)
TANG = [13.0 8.2 5.7 repmat(5.0, 1,5)
   9.0 7.5 6.0 5.1 repmat(4.5, 1,4)
   7.8 6.7 5.6 4.9 4.4 repmat(3.9, 1,3)
   6.4 5.5 4.7 4.1 3.6 repmat(3.2, 1,3)
   5.6 5.0 4.5 4.3 3.5 repmat(2.9, 1,3)
   4.2 3.9 3.7 3.3 2.9 repmat(2.42, 1,3)
   3.2 2.8 2.5 2.3 repmat(2.2, 1,3) 2.02
   2.8 2.1 1.9 1.8 1.7 1.6 1.6 1.41
   1.6 1.5 1.4 1.3 1.2 1.1 1.1 1.02
   1.5 1.2 0.94 repmat(0.77, 1,5)
   0.72 0.66 0.61 repmat(0.54, 1,5)
   0.44 0.41 0.40 repmat(0.39, 1,5)
   0.29 0.25 repmat(0.22, 1,6)
   0.15 repmat(0.13, 1,7)
   0.06 repmat(0.05, 1,7)
   repmat(0.04, 1,8)];

nBands = 24;
bands = [1:nBands];

% main loudness calculation before spreading/masking
%
% we want to implement equation:
%
%   KERN = NS_0 * (1/s * E_HS/E_0)^k * [ (1 + s * E/E_HS)^k - 1 ]
%
% where
%
%   s = 0.25;
%   k = 0.25;
%   NS_0 = 0.064; % (sone/Bark)
%
% to work with level in dB we need to remember
%   L_E = 10*log10 (E / E_0)
% and
%   L_EHS = 10*log10 (E_HS / E_0)

L_E_free = L_G - a_0; % excitation level less attenuation of ear

if strcmpi(opt.pres_type, 'diffuse')
   L_E = L_E_free + DL_ED;
else
   L_E = L_E_free;
end;
```

```
%
% loudness scale to sones and set floor to EHS
%
% NS = 0.064 * (10 ^ (0.025 * L_EHS))
%      * [ ((1 + 0.25 * 10 ^ (0.1 * (L_E-L_EHS))) ^ 0.25) - 1 ]
%

HSF = 0.064 .* (10 .^ (0.025 .* L_EHS));
KERN = HSF .* ...
    ( ((1 + 0.25 .* 10 .^ ((L_E - L_EHS) ./ 10)) .^ 0.25) - 1 );
KERN(L_E <= L_EHS) = 0;


%   label 3: Start Values
N = 0;
Z = opt.Z_inc;

f_bark = opt.Z_inc:opt.Z_inc:nBands;



Z1 = 0;
N1 = 0;
j = 16;
IZ = 1;

NS = zeros(size(f_bark));

for i=bands
    ZG = i;
    IG = i-1;
    if IG > 8; IG = 8; end;

    while Z1 < ZG   % from label 12 block: IF (Z1.LT.ZG) GO TO 4

        % label 4
        if N1 < KERN(i)
            % label 5
            for j=1:16
                if GRENZE(j) < KERN(i)
                    break;
                end;
            end;

            % label 7
            [Z2,N2,N] = label_7(ZG, KERN, i, N, Z1);

            % label 8
            [Z,IZ,NS] = label_8(Z,Z2,IZ,NS,N2,opt.Z_inc);

            % label 12
            [N1, Z1, j] = label_12(N2, Z2, GRENZE, j);

        elseif N1 == KERN(i)
            % label 7
            [Z2,N2,N] = label_7(ZG, KERN, i, N, Z1);
```

```
            % label 8
            [Z,IZ,NS] = label_8(Z,Z2,IZ,NS,N2,opt.Z_inc);

            % label 12
            [N1, Z1, j] = label_12(N2, Z2, GRENZE, j);

        else % N1 > KERN(i), i.e. excitation might be masked
            % label 9
            % if not masked use KERN(i)
            N2 = max(GRENZE(j), KERN(i));

            DZ = (N1-N2) / TANG(j,IG);
            Z2 = Z1 + DZ;
            if Z2 >= ZG
                Z2 = ZG;
                DZ = Z2 - Z1;
                N2 = N1 - DZ * TANG(j,IG);
            end;

            % label 10
            N = N + ((N1 + N2)/2) * DZ;

            % label 11
            while Z <= Z2
                NS(IZ) = N1 - (Z-Z1) * TANG(j,IG);
                IZ = IZ + 1;
                Z = Z + opt.Z_inc;
            end;

            % label 12
            [N1, Z1, j] = label_12(N2, Z2, GRENZE, j);

        end;
    end;
end;

sones = N;
specific_loudness = NS;


if nargout == 2
    info.f_bark = f_bark;
    info.Ndash = NS;
    info.a_0 = a_0;
    info.DL_ED = DL_ED;
    info.L_EHS = L_EHS;
    info.L_E_free = L_E_free;
    info.L_E = L_E;
    info.KERN = KERN;
end;


%% ------------------------------------------------------------
function [Z2,N2,N] = label_7(ZG, KERN, i, N, Z1)
% label 7
Z2 = ZG;
N2 = KERN(i);
N = N + N2 * (Z2 - Z1); % loudness integration
```

```
%% ------------------------------------------------------------
function [Z,IZ,NS] = label_8(Z,Z2,IZ,NS,N2,Z_inc)
while Z <= Z2
   NS(IZ) = N2;
   IZ = IZ + 1;
   Z = Z + Z_inc;
end;

%% ------------------------------------------------------------
function [N1, Z1, j] = label_12(N2, Z2, GRENZE, j)

if N2 == GRENZE(j); j=j+1; end;
if j>16; j=16; end;
N1=N2;
Z1=Z2;

%% ------------------------------------------------------------
function opt = getoptstruct(override, defvals)

opt = defvals;

names = fieldnames(override);
for i=1:length(names)
   opt.(names{i}) = override.(names{i});
end;
```

# C.6 Scott

## C.6.1 Main code

```
%% pc_scott
%   calculate the P-Centre of a sound using the model described
%   in Scott, S. "P-centers - an acoustic analysis", PhD Thesis,
%   University College London 1993.
%
%   pc = pc_scott(x, fs, options...);
%
%   pc:      the calculated p-centre (in secs)
%   x:       signal for which to calculate p-centre (assumes
%            signal contains just one p-centre)
%   fs:      sampling frequency
%
%   options: optional name value pairs
%
%     'threshold_db'
%            specify the threshold at which the signal onset is
%            identified. This threshold is applied to the (low
%            passed) envelope of the total signal (not just one
%            band). The default value is -30 dB relative to
%            envelope maximum (similar to Marcus's model). The
%            time of of 50% max amplitude is expressed relative
%            to the onset time.
```

```
%
%     'threshold_type'
%            Specify the type of threshold, either 'relative'
%            (the default) or 'absolute'.
%
%            If an absolute threshold is specified, then the
%            amplitude **MUST** be calibrated to an RMS amplitude
%            reference of 1. A calibrated 0 dB sine wave will
%            peak at approx. 1.414 and have an RMS value of 1.
%            See RECALIBRATE_AMPLITUDE.
%
%   Depends on Malcolm Slaney's auditory toolbox and a modified
%   version of the MakeERBFilters function which can accept a
%   bandwidth specification.

function [pc_ms, info] = pc_scott(x, fs, varargin)

o.threshold_db = -30;
o.threshold_type = 'relative';
o = getopt_name(varargin, o);

x = x(:)';

% prepare the envelope low pass filter
[b_env a_env] = butter(2,25/(fs/2));

% using a threshold, identify the signal onset which may be some
% way into the sampled data. To the fullband signal, apply the
% same envelope processing that will later be used on the subband
% signal (i.e. full wave rectify, then low pass filter - together
% these approximate the RMS)
env = filter(b_env, a_env, abs(x));
env_db = rms2db(env);

switch lower(o.threshold_type)
    case 'absolute'
        threshold_db = o.threshold_db;
    otherwise % relative threshold
        threshold_db = max(env_db) + o.threshold_db;
end

ionset = find(env_db >= threshold_db);
ionset = ionset(1);
onset_ms = ionset * 1000/fs;

% Create filterbank, then filter to create subbands. Each subband
% has a bandwidth of 4 ERBs (See Scott 1993, figure 12.2 and
% section 12.5.2). The following subband centre frequencies were
% specified:
%
%   108 299 578 997 1638 2651 4342
%
% However Scott only used subband 3 (578Hz) for further
% processing. (Scott counts down from the highest subband and
% calls this channel 5 in her thesis.)

bw_erb = 4;
f_subband = [578];
```

```
fcoefs = mod_MakeERBFilters(fs,f_subband,bw_erb);

subband = ERBFilterBank(x, fcoefs);

sub_env = filter(b_env, a_env, abs(subband));
sub_env_db = rms2db(sub_env);

% now find max amplitude
half_max_amp = 0.5 * max(sub_env);
half_max_amp_db = rms2db(half_max_amp);

% now find first location of 50% max amplitude
i50 = find(sub_env_db >= half_max_amp_db);
i50 = i50(1);
half_max_amp_ms = i50(1) * 1000/fs;

% now calculate p-center (assuming t and pc in millisecs)
pc_ms = onset_ms ...
    + (-11.2 + 0.407 * (half_max_amp_ms - onset_ms));

if nargout == 2
    info.env_db = env_db;
    info.threshold_db = threshold_db;
    info.subband = subband;
    info.sub_env_db = sub_env_db;
    info.half_max_amp_db = half_max_amp_db;
    info.onset_ms = onset_ms;
    info.half_max_amp_ms = half_max_amp_ms;
end

end


%% RMS2DB -----------------------------------------------------
function db = rms2db(amp)
ref = 1;
db = 20 * log10(max(amp/ref,1e-5));
end
```

## C.6.2 Code for non-standard Gammatone filter bandwidth

```
function fcoefs=mod_MakeERBFilters(fs,channels,bwERB)
% function [fcoefs]=MakeERBFilters(fs,channels,bwERB) This
% function computes the filter coefficients for a bank of
% Gammatone filters.  These filters were defined by Patterson and
% Holdworth for simulating the cochlea.
%
% The result is returned as an array of filter coefficients.
% Each row of the filter arrays contains the coefficients for
% four second order filters.  The transfer function for these
% four filters share the same denominator (poles) but have
% different numerators (zeros).  All of these coefficients are
% assembled into one vector that the ERBFilterBank can take apart
% to implement the filter.
%
% Channels input argument is a vector, then the values of this
```

```
% vector are taken to be the center frequency of each desired
% filter.  (The lowFreq argument is ignored in this case.)
%
% MODIFIED: 29 Jan 2003, Rudi Villing: added parameter bwERB to
%           allow bandwidth of each filter to be different than a
%           single ERB

T = 1/fs;
    cf = channels(1:end);
    if size(cf,2) > size(cf,1)
        cf = cf';
    end


% Change the following three parameters if you wish to use a
% different ERB scale.  Must change in ERBSpace too.
EarQ = 9.26449;             %  Glasberg and Moore Parameters
minBW = 24.7;
order = 1;

ERB = ((cf/EarQ).^order + minBW^order).^(1/order);
B=1.019*2*pi*ERB*bwERB;

A0 = T;
A2 = 0;
B0 = 1;
B1 = -2*cos(2*cf*pi*T)./exp(B*T);
B2 = exp(-2*B*T);

A11 = -(2*T*cos(2*cf*pi*T)./exp(B*T) ...
    + 2*sqrt(3+2^1.5)*T*sin(2*cf*pi*T)./ exp(B*T))/2;
A12 = -(2*T*cos(2*cf*pi*T)./exp(B*T) ...
    - 2*sqrt(3+2^1.5)*T*sin(2*cf*pi*T)./ exp(B*T))/2;
A13 = -(2*T*cos(2*cf*pi*T)./exp(B*T) ...
    + 2*sqrt(3-2^1.5)*T*sin(2*cf*pi*T)./ exp(B*T))/2;
A14 = -(2*T*cos(2*cf*pi*T)./exp(B*T) ...
    - 2*sqrt(3-2^1.5)*T*sin(2*cf*pi*T)./ exp(B*T))/2;

gain = abs((-2*exp(4*i*cf*pi*T)*T + ...
    2*exp(-(B*T) + 2*i*cf*pi*T).*T.* ...
    (cos(2*cf*pi*T) - sqrt(3 - 2^(3/2))* ...
    sin(2*cf*pi*T))) .* ...
    (-2*exp(4*i*cf*pi*T)*T + ...
    2*exp(-(B*T) + 2*i*cf*pi*T).*T.* ...
    (cos(2*cf*pi*T) + sqrt(3 - 2^(3/2)) * ...
    sin(2*cf*pi*T))).* ...
    (-2*exp(4*i*cf*pi*T)*T + ...
    2*exp(-(B*T) + 2*i*cf*pi*T).*T.* ...
    (cos(2*cf*pi*T) - ...
    sqrt(3 + 2^(3/2))*sin(2*cf*pi*T))) .* ...
    (-2*exp(4*i*cf*pi*T)*T + 2*exp(-(B*T) + 2*i*cf*pi*T).*T.* ...
    (cos(2*cf*pi*T) + sqrt(3 + 2^(3/2))*sin(2*cf*pi*T))) ./ ...
    (-2 ./ exp(2*B*T) - 2*exp(4*i*cf*pi*T) +  ...
    2*(1 + exp(4*i*cf*pi*T))./exp(B*T)).^4);

allfilts = ones(length(cf),1);
fcoefs = [A0*allfilts A11 A12 A13 A14 ...
    A2*allfilts B0*allfilts B1 B2 gain];
```

```
if (0)                          % Test Code
    A0  = fcoefs(:,1);
    A11 = fcoefs(:,2);
    A12 = fcoefs(:,3);
    A13 = fcoefs(:,4);
    A14 = fcoefs(:,5);
    A2  = fcoefs(:,6);
    B0  = fcoefs(:,7);
    B1  = fcoefs(:,8);
    B2  = fcoefs(:,9);
    gain= fcoefs(:,10);
    chan=1;
    x = [1 zeros(1, 511)];
    y1=filter([A0(chan)/gain(chan) A11(chan)/gain(chan) ...
        A2(chan)/gain(chan)],[B0(chan) B1(chan) B2(chan)], x);
    y2=filter([A0(chan) A12(chan) A2(chan)], ...
            [B0(chan) B1(chan) B2(chan)], y1);
    y3=filter([A0(chan) A13(chan) A2(chan)], ...
            [B0(chan) B1(chan) B2(chan)], y2);
    y4=filter([A0(chan) A14(chan) A2(chan)], ...
            [B0(chan) B1(chan) B2(chan)], y3);
    semilogx((0:(length(x)-1))*(fs/length(x)),...
        20*log10(abs(fft(y4))));
end
```

## C.7 Harsin

```
%% PC_HARSIN
%   An implementation of the Harsin's P-Centre model.
%
%   Harsin, C.A. (1997), "perceptual-center modeling is affected
%     by including acoustic rate-of-change modulations",
%     Perception and Psychophysics, vol 59, pp 243-251
%
%   Harsin, C.A. (1993), 'Perceptual Centers and the Relation of
%     Acoustic Energy Modulation to Speech Timing', Unpublished
%     PhD thesis, University of New Orleans.
%
%   pc_ms = pc_harsin(x, fs)
%
%   pc_ms:   the calculated pcentre in milliseconds
%   info:    optional internal data/values from the model
%   x:       the signal for which p-centre will be calculated. It
%            is assumed that the signal is a sound perceived as
%            having exactly one p-centre.
%   fs:      the sampling frequency of the signal

function [pc_ms, info] = pc_harsin(x,fs, varargin)

% 'harsin' or 'stevens' of Stevens's power law fame
o.loud_calc = 'harsin';
o.band_calc = 'power'; % otherwise 'magnitude'
% M_calc other values: 'pos' (positive only), or 'signed'
% (negative will subtract)
o.M_calc = 'abs';
```

```
o = getopt_name(varargin, o);

if strcmpi('power', o.band_calc)
    combine_pwr = true;
else
    combine_pwr = false;
end


x = x(:);

% STEP 1: convert to Fs of 10 kHz to match Harsin's paper
% ____________________________________________________________________

if fs ~= 10000
    x = resample(x,10000,fs);
    fs = 10000;
end

% STEP 2: bandpass into channels
% ____________________________________________________________________

% upper and lower 3dB cutoffs (2 critical band filters)
% See Harsin 1993, Table 4, p66 for details
fc_chan = [366, 659;
   1073, 1293;
   1635, 1928;
   2172, 2586;
   2904, 3514;
   3956, 4758];
Nchannels = size(fc_chan,1);

for ch=1:size(fc_chan,1)
    [b,a] = butter(2, fc_chan(ch,:)./(fs/2));
    ch_sig(:,ch) = filter(b, a, x);
end

% STEP 3: extract envelope and downsample
% ____________________________________________________________________

% envelope = absolute value of signal in each channel (full wave
% rectification)
ch_env1 = abs(ch_sig);

% lowpass filter with 3rd order butterworth at 100Hz
[b, a] = butter(3, 100/(fs/2));
ch_env2 = filter(b, a, ch_env1);

% decimate 25:1 to get sample rate down to 400Hz
fs_env = 400;
% downsample without filtering
ch_env3 = downsample(ch_env2, fix(fs/fs_env));

% lowpass filter with 3rd order butterworth again to remove
% discontinuities remove negative values (side effect of IIR
% filtering)
```

```
[b, a] = butter(3, 100/(fs_env/2));
ch_env = filter(b, a, ch_env3);
ch_env = max(ch_env, 0);

% STEP 4: loudness scale by raising to 0.3 power
% ____________________________________________________________

% NOTE: Stevens's power law for loudness scaling is
% pressure ^0.6 (or intensity ^ 0.3) and not pressure ^ 0.3 (i.e.
% not amplitude ^ 0.3).

if strcmpi('harsin', o.loud_calc)
    ch_loud = ch_env .^ 0.3;
else
    ch_loud = ch_env .^ 0.6;
end

% STEP 5: convert to psychoacoustic envelope for each channel
% ____________________________________________________________

% for each channel...
%  prepend with 512 zeroes
%  calc 512 point FFT for a rectangular window, then advance 10ms
%  (4 env samples)
%  take magnitude power spectrum of each FFT
%  scale power spectrum bins according to modulation weightings
%  sum all bins to give "perceptual envelope"

Nfft = 512;
Nadv = fix(10 * fs_env/1000);
Noverlap = Nfft - Nadv;
fs_penv = fs_env/4; % because we advance 4 points for each FFT


% perceptual weight, lower band freq, upper band freq, for each
% modulation band
w_modband = [1 3.1 5.5;
    0.8 6.25 11.75;
    0.45 12.5 23.5;
    0.2 24 47];

f_fft = fft_freq(Nfft, fs_env);
w_fft = zeros(1,Nfft);
for i=1:size(w_modband,1)
    band_bins(i,:) = logical((w_modband(i,2) <= abs(f_fft)) ...
        & (abs(f_fft) < w_modband(i,3)));
    w_fft(band_bins(i,:)) = w_modband(i,1);
end

% modulationWeights = zeros(1,Nfft);
% modulationWeights(5:8) = 1;
% modulationWeights(9:16) = 0.8;
% modulationWeights(17:31) = 0.45;
% modulationWeights(32:61) = 0.2;

for ch=1:Nchannels
    % each buffered frame is a column
    loud = buffer([zeros(Nfft,1); ch_loud(:,ch)], ...
```

```
        Nfft, Noverlap, 'nodelay');

    % the double sided magnitude spectrum of each frame is also a
    % column
    %
    % NOTE 1: Neither ref specifies spectrum scaling (e.g.
    % normalization by N, single sided vs. double sided spectrum,
    % etc) but weighted sums calculated later are indpendent of
    % any constant scaling factor so it doesn't matter.
    %
    % NOTE 2: Harsin 1997 refers to a "power" spectrum. Harsin
    % 1993, p41, also calls it a "power spectrum" but the
    % calculation described yields a magnitude spectrum. Thus a
    % magnitude spectrum is what is calculated here.

    % size = [Nfft, nFrame]
    pwr = (abs(fft(loud,Nfft)) .^ 2) / Nfft;
    mag = sqrt(pwr);

    % NOTE: for both the channel modulations and psychoacoustic
    % envelope, Harsin only says that values should be combined.
    % In Harsin 1993, Figures 15, 16, and 17 are all in volts or
    % arbitrary amplitude units. This seems to suggest that he
    % was always working with the magnitude and not the power
    % values.
    %
    % However: sum(mag) ~= sqrt(sum(pwr)) Magnitudes would not
    % generally be summed directly

    % channel modulations (i.e. the power in specific sub-bands
    % of the "loudness" envelope) before perceptual scaling.
    for bnd=1:size(band_bins,1)
        if combine_pwr
            % sum down each column
            band_pwr = sum(pwr(band_bins(bnd,:),:),1);
            ch_mod(:,ch,bnd) = sqrt(band_pwr)';
        else
            % sum down each column
            band_mag = sum(mag(band_bins(bnd,:),:),1);
            ch_mod(:,ch,bnd) = band_mag';
        end
    end

    % scale the channel modulations to yield the psychoacoustic
    % envelope
    % mag [Nfft,nFrame], w_fft [1,Nfft]
    % => mag' * w_fft' = [nFrame,Nfft] x [Nfft,1] = [nFrame, 1]
    % => each col = one channel psych envelope

    if combine_pwr
        ch_penv(:,ch) = sqrt(pwr' * (w_fft.^2)');
    else
        ch_penv(:,ch) = (mag' * w_fft');
    end
end;


% STEP 6: find envelope velocity peaks, times and envelope
```

```
%         magnitude differences
% ______________________________________________________________

% prepend zero to account for data loss by diff
ch_vel = [zeros(1,Nchannels); diff(ch_penv,1,1)];

% V: the magnitude of a peaks in the velocity (first derivative)
%    of the perceptual envelopes for channel (channel_penv)
% T: The time at which the peak occurs (in seconds)
% M: magnitude difference between perceptual envelope at time
%    T(i) relative to magnitude at T(i-1)

ch_V = cell(Nchannels,1);
ch_T_ms = cell(Nchannels,1);
ch_M_raw = cell(Nchannels,1);
ch_M = cell(Nchannels,1);
for ch=1:Nchannels
    % find the velocity peaks
    pk = find_maxima(ch_vel(:,ch));

    ch_V{ch} = ch_vel(pk,ch);

    % Harsin does not specify the time units - so assume
    % millisecs
    ch_T_ms{ch} = (pk-1) .* (1000/fs_penv);

    dpeakmag = [ch_penv(pk(1),ch); diff(ch_penv(pk,ch))];
    ch_M_raw{ch} = dpeakmag;

    % Harsin refers to magnitude increments (suggesting
    % positivity) but never explicitly specifies that decrements
    % should be excluded. Furthermore, Harsin (1997, p249),
    % states that each increment "is the amount of change [...]
    % since the last velocity peak" which is interpreted here as
    % indicating absolute value rather than a signed value for
    % the change.

    switch lower(o.M_calc)
        case 'signed', ch_M{ch} = dpeakmag;
        case 'abs', ch_M{ch} = abs(dpeakmag);
        case 'pos'
            % only an onset whose magnitude is greater than that
            % of the previous retained onset will be non-zero
            % (and thus retained)
            prev_env = ch_penv(pk(1),ch);
            pm = zeros(length(pk), 1);
            pm(1) = prev_env;
            for i=2:size(pk)
                if ch_penv(pk(i),ch) > prev_env
                    pm(i) == ch_penv(pk(i),ch) - prev_env;
                    prev_env = ch_penv(pk(i),ch);
                end
            end
            ch_M{ch} = pm;
        otherwise
            error(['unrecognised calculation for ' ...
                'magnitude increment M: %s'], o.M_calc);
    end
```

```
end

% STEP 7: calc per band magnitude weighted velocity model (BMVM)
% ______________________________________________________________

V = cat(1,ch_V{:});
T_ms = cat(1,ch_T_ms{:});
M = cat(1,ch_M{:});

BMVM = sum(M .* V .* T_ms) / sum(M .* V);

% STEP 8: p-center from regression equation
% ______________________________________________________________

pc_ms = 9.3 + (1.12 * BMVM);

% outputs
if nargout == 2
    info.Nchannels = Nchannels;
    info.ch_sig = ch_sig;
    info.fs = fs;
    info.ch_env = ch_env;
    info.fs_env = fs_env;
    info.ch_loud = ch_loud;
    info.ch_mod = ch_mod;
    info.fs_penv = fs_penv;
    info.ch_penv = ch_penv;
    info.ch_M_raw = ch_M_raw;
    info.ch_M = ch_M;
    info.ch_T_ms = ch_T_ms;
    info.ch_V = ch_V;
    info.ch_vel = ch_vel;
end


%% FIND_MAXIMA ------------------------------------------------
function imax = find_maxima(x);

sdx = sign(diff(x,1)); % 1=rising, 0=flat, -1=falling

% if a flat segment occurs before a rising segment, consider it
% part of the rising segment. If it occurs before a falling
% segment, consider it part of the falling segment.

for i=length(sdx):-1:2
    if (sdx(i-1) == 0)
        sdx(i-1) = sdx(i);
    end
end

% Maxima are located where sdx transitions from rising (1) to
% flat (0) or fall (-1). So diff sdx would be -1 or -2.

% add 1 to correct for diff being shorter vector than x
imax = find(diff(sdx) < 0) + 1;
imax = imax(x(imax) > 0);
```

```
%% FFT_FREQ ------------------------------------------------------
function f = fft_freq(Nfft, fs)

f = (0:Nfft-1) * fs/Nfft;
neg = f > fs/2;
f(neg) = -fs + f(neg);
```

## C.8 Additional support code required

### C.8.1 Recalibrate Amplitude

```
%% RECALIBRATE_AMPLITUDE
%  Recalibrate the amplitude of a signal to a particular
%  reference level.
%
%  y = calibrate_amplitude(x, old_ref, old_type, ...
%                          new_ref, new_type)
%
%  y:         the recalibrated signal
%  x:         the original signal
%  old_ref:   the original reference level, often 1
%  old_type:  the original reference type, often 'peak'. Can also
%             be 'rms'
%  new_ref:   the new reference level
%  new_type:  the new reference type, either 'peak' or 'rms'
%
%  EXAMPLE:
%
%  % create a normal full scale sine wave which peaks at 1
%  x = sin((1:100) * 2*pi*10/1000);
%
%  % scale to 60 dB above the reference, i.e. 1000 times higher
%  x = x .* 1000;
%
%  % recalibrate to the SPL RMS reference
%  spl_ref = 2e-5;
%  y = recalibrate_amplitude(x,1,'peak',spl_ref,'rms');
%
%  % verify that the RMS of the resulting signal is 60 dB
%  % above the SPL reference
%  rms_y = sqrt(mean(y.^2));
%  20 * log10(rms_y / spl_ref)
%
%  ans =
%
%     60.0000
%

function y = recalibrate_amplitude(x,old_ref,old_ref_type,...
    new_ref,new_ref_type)
```

```
if strcmpi('rms',old_ref_type)
    % convert to equivalent (sine wave) peak ref
    old_ref = old_ref * sqrt(2);
end
if strcmpi('rms',new_ref_type)
    % convert to equivalent (sine wave) peak ref
    new_ref = new_ref * sqrt(2);
end

y = x .* (new_ref / old_ref);
```

### C.8.2 Exponential averaged loudness based on BS.1770

```
% loudness_bs1770_integrator
%   An implementation of recommendation ITU-R BS.1770
%   "Algorithms to measure audio programme loudness and true-peak
%   audio level" - modified to use leaky integration of RMS with
%   specific time constant
%
%   BS1770 is essentially an RMS model of loudness applied to a
%   filtered version of the signal being measured.
%
%   [loudness_db, internals] = loudness_bs1770(x,win,n_adv)
%
%   loudness_db = loudness level (dB) per frame
%   internals (optiona) = internal data from algorithm for
%                         debugging/insight
%
%   x =  signal to be measured (NOTE: sampling frequency must be
%        48000 or pre filters will be incorrect)
%   fs:  sampling frequency
%   tau: integrator time constant (seconds)

function [loudness_db, internals] = ...
    loudness_bs1770_integrator(x,fs,tau)

if min(size(x)) > 1
    error('x must be a vector');
end
if fs ~= 48000
    x = resample(x,48000,fs);
    fs = 48000;
end

% simulation of head (HRTF) as a rigid sphere, gives a 4dB step
% up in gain between 1 and 3 kHz (assuming 48kHz sampling rate)
head_b = ...
    [ 1.53512485958697, -2.69169618940638, 1.19839281085285 ];
head_a = [ 1 -1.69065929318241, 0.73248077421585 ];

% Revised Low frequency B-weighting filter (fs=48kHz)
rlb_b = [ 1 -2 1 ];
rlb_a = [ 1 -1.99004745483398, 0.99007225036621 ];

x_head = filter(head_b, head_a, x);
```

```
x_rlb = filter(rlb_b, rlb_a, x_head);

% now do the exponentially averaged RMS part

Ts = 1/fs;
alpha = 1 - exp(-Ts/tau);

z = filter(alpha, [1 -(1-alpha)], x_rlb.^2);

loudness_db = -0.691 + 10 * log10(max(z,eps));


if nargout == 2
    internals.head_b = head_b;
    internals.head_a = head_a;
    internals.rlb_b = rlb_b;
    internals.rlb_a = rlb_a;
    internals.x_head = x_head;
    internals.x_rlb = x_rlb;
    internals.z = z;
end
```

### C.8.3 Getopt name

```
%%GETOPT_NAME
%
%  Helper for functions which can take named optional arguments.
%  Where an optional argument is not supplied the default value
%  is set instead.
%
%  opt = getopt_name(args, default_opt, mode)
%
%  opt:      returned structure of values
%  args:     the supplied arguments. Any combination of
%            structures, cell arrays with name-value pairs, and
%            name-value pairs of arguments are supported.
%  default_opt: the default value for any optional arguments not
%            supplied
%  mode:     [OPTIONAL] qualifies the operation of getopt_name.
%            Possible values are
%
%    'merge_extra' {default}, unrecognized options are merged
%            into output struct
%    'split_extra' unrecognized options are collected in a single
%            field called 'unrecognized'
%    'reject_extra' unrecognized options are rejected
%
%  NOTE 1: options must be case-insensitive unique
%  NOTE 2: Abbreviated option names can be passed in, but they
%          must match uniquely, or match a short field name
%          exactly
%
%  Example use:
%
%    function y = foo(x, varargin)
%    opt.a = -1;
%    opt.b = 'empty';
```

```
%    opt.c = [];
%    opt.d = 'd';
%    opt.e = 'e';
%    opt.f = -1;
%    opt = getopt_name(varargin, opt);
%    ...
%
%  Then
%
%    args.a = 12;
%    args.f = 24;
%    foo(x, args, {'b', 'hello'}, 'c', [1,2,3,4])
%
%  will result in the values
%
%    opt.a = 12
%    opt.b = 'hello'
%    opt.c = [1,2,3,4]
%    opt.d = 'd';
%    opt.e = 'e';
%    opt.f = 24;
%
% Author: Rudi Villing

function opt = getopt_name(in_opt, default_opt, mode);

opt = default_opt;

if length(in_opt)==0
   return;
end

if nargin < 3
    mode = 'merge_extra';
end


% process options passed to getopt's caller - Any combination of
% structures, cell arrays of name/value pairs and name/value
% pairs is allowed
names = {};
values = {};
i = 1;
while i<=length(in_opt)
    if isstruct(in_opt{i})
        names = [ names, fieldnames(in_opt{i}) ];
        values = [ values, struct2cell(in_opt{i}) ];
        i = i+1;
    elseif iscell(in_opt{i})
        if mod(length(in_opt{i}),2) ~= 0
            error(['Cell array options at position '...
                '%d must consist of matching '...
                'name/value pairs'],i);
        end
        names = [ names, in_opt{i}(1:2:end) ];
        values = [ values, in_opt{i}(2:2:end) ];
        i = i+1;
    elseif ischar(in_opt{i});
        if (i+1) > length(in_opt)
```

```
            error('missing matching value at end of options');
        end
        names{end+1} = in_opt{i};
        values{end+1} = in_opt{i+1};
        i = i+2;
    else
        error(['Invalid option type %s at '...
            'position %d'], class(opt_in{i}), i);
    end
end

% now fill in values
% use strmatch to try and find matches based on abbreviated field
% names as long as they are unique.
% any name not in default_opt is treated as invalid
valid_names = fieldnames(opt);
lower_valid_names = lower(valid_names);

% check for unique field names in default options
unames=unique(lower_valid_names);
if length(unames) ~= length(valid_names)
    error(['default_opt names which differ only in case '...
        'are not supported:\n',...
        sprintf('  ''%s''\n', ...
        valid_names{[1:length(valid_names)]})]);
end

% check for unique names in input options
lower_names = lower(names);
unames=unique(lower_names);
if length(unames) ~= length(names)
    error(['in_opt names which differ only in case '...
        'are not supported:\n',...
        sprintf('  ''%s''\n', names{[1:length(names)]})]);
end

% OK, unique names used, so match them up
for i=1:length(lower_names)
    iname = strmatch(lower_names{i}, lower_valid_names);
    if length(iname)==1 % is it a unique match?
         % set the unique match
        opt.(valid_names{iname}) = values{i};
    elseif length(iname) > 1 % or is it non-unique?
        % are any of the matches exact?
        iexact = strcmpi(lower_names{i}, ...
            lower_valid_names(iname));
        if sum(iexact)==1
            iname = iname(iexact);
            % set the exact match
            opt.(valid_names{iname}) = values{i};
        else
            error(['cannot have in_opt abbreviation which '...
                'partly matches multiple field names:\n',...
                sprintf('  ''%s''\n', valid_names{iname})]);
        end
    else % or was there no match => an unrecognised option?
        switch lower(mode)
            case {'merge','merge_extra'}
                opt.(names{i}) = values{i};
```

```
            case {'split','split_extra'}
                opt.unrecognized.(names{i}) = values{i};
            otherwise
                error('unrecognised option name ''%s''', ...
                    names{i});
        end
    end
end
```

# Appendix D

# International Phonetic Alphabet (IPA)

**Table D.1 IPA for English consonants**

| Pan-English | Phones | Examples |
|---|---|---|
| p | $p^h$, p | **p**en, s**p**in, ti**p** |
| b | b | **b**ut, we**b** |
| t | $t^h$, t, ɾ, ʔ | **t**wo, s**t**ing, be**t** |
| d | d, ɾ | **d**o, o**dd** |
| tʃ | $tʃ^h$, tʃ | **ch**air, na**t**ure, tea**ch** |
| dʒ | dʒ | **g**in, **j**oy, e**dge** |
| k | $k^h$, k | **c**at, **k**ill, s**k**in, **q**ueen, uni**que**, thi**ck** |
| g | g | **g**o, **g**et, be**g** |
| f | f | **f**ool, enou**gh**, lea**f**, o**ff**, **ph**oto |
| v | v | **v**oice, ha**ve**, o**f** |
| θ | θ | **th**ing, tee**th** |
| ð | ð | **th**is, brea**the**, fa**th**er |
| s | s | **s**ee, **c**ity, pa**ss** |
| z | z | **z**oo, ro**s**e |
| ʃ | ʃ | **sh**e, **s**ure, emo**ti**on, lea**sh** |
| ʒ | ʒ | plea**s**ure, bei**ge**, sei**z**ure |
| x (k) | x | lo**ch** (Scottish) |
| h | h, ɦ | **h**am |
| m | m | **m**an, ha**m** |
| n | n | **n**o, ti**n** |

| Pan-English | Phones | Examples |
|---|---|---|
| ŋ | ŋ | ri**ng**er, si**ng**, fi**n**ger, dri**n**k |
| l | l, ɬ | **l**eft, be**ll** |
| r | ɹʷ, ɹ, ɾ | **r**un, ve**r**y |
| w | w | **w**e, q**u**een |
| j | j | **y**es |
| hw (w) | hw | **wh**at |

*Note*—Table reproduced (with minor reformatting) from (Wikipedia Contributors 2009)

**Table D.2 IPA for English marginal sounds and reduced vowels**

| Pan English | Phones | Examples |
|---|---|---|
| ʔ | ʔ | uh-(ʔ)oh |
| ə | Reduced /ʌ, æ, ɑː, ɒ/ | |
| ɨ (ə) | Reduced /ɪ, iː, ɛ, eɪ, aɪ/ | |
| ʉ (ə) | Reduced /ʊ, uː/ | |
| ɵ (ə) | Reduced /oʊ/ | |
| ɚ (ə) | Reduced /ɜː/ (ɜr) | |

*Note*—Table reproduced from (Wikipedia Contributors 2009)

**Table D.3 IPA for English vowels**

| Pan-English | GA | IrE | RP | Lexical set | Examples |
|---|---|---|---|---|---|
| æ | æ, eə | ɑ/æ | æ | TRAP | lad, bad, cat |
| ɑː | ɑ | ɑː | ɑː | PALM | father |
| ɒ | | ɑ | ɒ | LOT | not, wasp |
| ɔː | ɔ | ɔː | ɔː | THOUGHT | law, caught, all, halt, talk |
| ə | ə | | ə | COMMA | about |
| ɨ | ɨ | | ɪ | | spotted |
| ɪ | ɪ | ɪ | ɪ | KIT | sit |
| i | i | | i | HAPPY | city |
| iː | | iː | iː | FLEECE | see |

| Pan-English | GA | IrE | RP | Lexical set | Examples |
|---|---|---|---|---|---|
| | | | | | meat |
| eɪ | eɪ | eː | eɪ | FACE | date |
| | | | | | day, pain, whey, rein |
| ɛ | ɛ | ɛ | ɛ | DRESS | bed |
| | | ʌɾ | | | burn |
| ɜr | ɜˠ/ɹ̩ | ɛɾ | ɜː(ɹ) | NURSE | herd, earth |
| | | ɪɾ | | | bird |
| ər | ɚˠ/ɹ̩ | | ə(ɹ) | LETTER | winner |
| ʌ | ʌ | ɔ, ʊ | ʌ | STRUT | run, won, flood |
| ʊ | ʊ | | ʊ | FOOT | put |
| | | | | | hood |
| uː | u | uː | uː | GOOSE | through, you |
| | | | | | threw, yew |
| juː | (j)u | juː | juː | | cute, dew, ewe |
| aɪ | aɪ, ʌi | ɔɪ | aɪ | PRICE | my, wise, high |
| ɔɪ | ɔɪ | | ɔɪ | CHOICE | boy, hoist |
| oʊ | oʊ | oː | əʊ | GOAT | no, toe, soap |
| | | | | | tow, soul, roll, cold, folk |
| aʊ | aʊ | | aʊ | MOUTH | now, trout |
| ɑr | ɑɹ | | ɑː(ɹ) | START | arm, car |
| ɪər | ɪɹ | | ɪə(ɹ) | NEAR | deer, here |
| ɛər | ɛɹ | | eə(ɹ) | SQUARE | mare, there, bear |
| ɔr | ɔɹ | ɑɾ | ɔː(ɹ) | NORTH | sort, warm |
| ɔər | oɹ, ɔɹ | oːɾ | | FORCE | tore, boar, port |
| ʊər | ʊɹ | | ʊə(ɹ) | CURE | tour, moor |
| jʊər | jʊɹ, jɜˠ | | jʊə(ɹ), jɔː(ɹ) | CURE | pure, Europe |

*Note*—Table reproduced from (Wikipedia Contributors 2009) and edited to remove dialects of English unnecessary to this thesis

# Appendix E

# Glossary

*Alpha (band)*

EEG oscillations in the frequency band 8–13 Hz

*Anisochrony*

Occurring at different intervals (typically used to refer to deviation from isochrony)

*Beta (band)*

EEG oscillations in the frequency band 12–30 Hz

*Complex tone*

A periodic waveform consisting of multiple partials which might or might not be related to one another harmonically

*Delta (band)*

EEG oscillations in the frequency band 1–4 Hz

*Diotic*

Same (mono) signal presented to both ears

*Diphthong*

A gliding vowel that changes quality during pronunciation

*Distal (source)*

The far away (original) source [of an event]

*Disyllable*

A word consisting of exactly two syllables

*Event*

Any brief occurrence including short speech sounds, musical notes, brief flashes, or gestures

*Event Related Potentials*

Low amplitude changes in neuroelectric activity that are time-locked to sensory, motor, or cognitive events. This definition incorporates evoked potentials as a subset of event related potentials.

*Eevoked power*

Power in the EEG components that are phase locked to event onset

*Gamma (band)*

EEG oscillations in the frequency band 20–60 Hz

*Harmonic tone*

A complex tone in which all the partials are related to one another harmonically (i.e. all integer multiples of the fundamental frequency)

*Induced power*

Amplitude modulation that is time-locked to the event onset though the underlying EEG oscillations are not

*Inharmonic tone*

A complex tone in which at least some partials are not related to one another harmonically

*Inter-trial phase coherence*

The coherence across many trials of phase angles measured at corresponding time-frequency points

*Isochrony*

The state of being isochronous, that is, occurring at identical intervals.

*Just noticeable difference*

The smallest detectable difference between a starting and secondary level of a particular sensory stimulus, also known as the difference limen.

*Meter*

The temporal framework in which rhythm exists

*Monophthong*

A pure vowel pronounced with the articulators kept rather still

*Monosyllable*

A word consisting of just one syllable

*Objective onset*

The objectively measurable onset of an event usually determined by a threshold.

*Perceptual centre (P-centre)*

The specific moment at which a brief event is perceived to occur

*Perceptual onset*

The moment at which the initial sensations associated with an event are first detected. This may precede its P-centre. (In a speech syllable like "sat", for example, the beginning of the initial /s/ may be perceived distinct from and clearly preceding the P-centre of the syllable.)

*Physical onset*

See objective onset.

*Point of objective isochrony*

The point at which consecutive inter-onset intervals between events are identical

*Point of perceptual isochrony*

See point of subjective isochrony

*Point of perceptual synchrony*

See point of subjective synchrony

*Point of subjective isochrony*

The point at which consecutive inter-P-centre intervals between events are identical. The events are perceived to occur at identical intervals.

*Point of subjective synchrony*

The relative timing at which two events are perceived to occur in synchrony with one another

*Pulse*

The basic periodic beat in music

*Pulse group*

A group of pulses with a particular stress pattern

*Pure tone*

A sinuisoidal waveform consisting of a single frequency component

*Rhotic [vowel]*

An r-coloured vowel whose distinctive feature is a low third formant

*Rhythm*

A temporal pattern with some element of regularity and predictability

*Stimulus origin*

Used in this work to refer to the time of the first data sample for digitally stored waveforms. The onset of acoustic energy may occur some delay after the origin.

*Syllable nucleus*

The vowel or vowel-like sound that is required for all syllables.

*Syllable onset*

The initial consonant or consonants preceding the nuclear vowel in a syllable. In some syllables there may be none.

*Synchrony*

The state of being synchronous, that is, occurring at the same time.

*Tempo*

The rate at which music (or a temporal pattern) is presented.

*Theta (band)*

EEG oscillations in the frequency band 4–8 Hz

# References


Addison, P. S. 2002, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*, Institute of Physics Publishing, Bristol, UK.

Allen, G. D. 1972a, 'The location of rhythmic stress beats in English: an experimental study I', *Language and Speech*, vol. 15, pp. 72-100.

Allen, G. D. 1972b, 'The location of rhythmic stress beats in English: an experimental study II', *Language and Speech*, vol. 15, pp. 179-95.

Aristotle 1993, *De anima: books II and III with passages from, Book I*, 2nd edn, (DW Hamlyn, Trans.), Clarendon Aristotle series, Clarendon Press, (350 BC).

Aschersleben, G. 2002, 'Temporal control of movements in sensorimotor synchronization', *Brain and Cognition*, vol. 48, no. 1, pp. 66-79.

Aschersleben, G., Gehrke, J. & Prinz, W. 2004, 'A psychophysical approach to action timing', in C Kaernbach, E Schröger & H Müller (eds), *Psychophysics beyond sensation: Laws and Invariants of Human Cognition*, Erlbaum, Hillsdale, NJ, pp. 117-36.

Bakeman, R. 2005, 'Recommended effect size statistics for repeated measures designs', *Behavior Research Methods*, vol. 37, no. 3, pp. 379-84.

Barry, R. J. 2009, 'Evoked activity and EEG phase resetting in the genesis of auditory Go/NoGo ERPs', *Biological Psychology*, vol. 80, p. 292–9.

Bell, A. & Biasca, D. H. 1994, *Perceptual centers are affected by stress location in English disyllables*, Austin, Texas, December 2, Poster.

Bell, A. & Morishima, Y. 1994, *Perceptual centers in Japanese disyllables*, Austin, Texas, December 2, Poster.

Bengtsson, S. L., Ullén, F., Ehrsson, H. H., Hashimoto, T., Kito, T., Naito, E., Forssberg, H. & Sadato, N. 2009, 'Listening to rhythms activates motor and premotor cortices', *Cortex*, vol. 45, no. 1, pp. 62-71.

Benguerel, A.-P. & D'Arcy, J. 1986, 'Time-warping and the perception of rhythm in speech', *Journal of Phonetics*, vol. 14, no. 2, pp. 231-46.

Bregman, A. S. 1999, *Auditory scene analysis: the perceptual organization of sound*, 2nd edn, MIT Press, Cambridge, MA (1990).

Buhusi, C. V. & Meck, W. H. 2005, 'What makes us tick? Functional and neural mechanisms of interval timing', *Nature Reviews. Neuroscience*, vol. 6, no. 10, pp. 755-65.

Burger, M., Hoppe, U., Lohscheller, J., Eysholdt, U. & Döllinger, M. 2009, 'Speech-Evoked Potentials Revealed by Approximations of Tone-Evoked Waveforms', *Ear and Hearing*, vol. 30, no. 1, pp. 16-22.

Buus, S., Florentine, M. & Poulsen, T. 1997, 'Temporal integration of loudness, loudness discrimination, and the form of the loudness function', *Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 669-80.

*Cambridge Dictionaries Online*, 2009, Cambridge University Press, viewed 20 October 2009, <http://dictionary.cambridge.org/>.

Collins, N. 2006, 'Investigating computational models of perceptual attack time', paper presented to 9th International Conference on Music Perception and Cognition (ICMPC9).

Cooper, A. M., Whalen, D. H. & Fowler, C. A. 1986, 'P-centers are unaffected by phonetic categorization', *Perception and Psychophysics*, vol. 39, no. 3, pp. 187-96.

Cooper, A. M., Whalen, D. H. & Fowler, C. A. 1988, 'The syllable's rhyme affects its P-center as a unit', *Journal of Phonetics*, vol. 16, pp. 231-41.

Davis, M. A. 1939, 'Effects of acoustic stimulation on the waking human brain', *Journal of Neurophysiology*, vol. 2, pp. 494-9.

de Jong, K. J. 1994, 'The correlation of P-center adjustments with articulatory and acoustic events', *Perception & Psychophysics*, vol. 56, pp. 447-60.

Delorme, A. & Makeig, S. 2004, 'EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent

component analysis', *Journal of Neuroscience Methods*, vol. 134, pp. 9-21.

Digeser, F. M., Wohlberedt, T. & Hoppe, U. 2009, 'Contribution of Spectrotemporal Features on Auditory Event-Related Potentials Elicited by Consonant-Vowel Syllables', *Ear and Hearing*, vol. 30, no. 6, p. Epub.

Efron, R. 1970a, 'Effect of stimulus duration on perceptual onset and offset latencies', *Perception & Psychophysics*, vol. 8, no. 4, pp. 231-4.

Efron, R. 1970b, 'The minimum duration of a perception', *Neuropsychologia*, vol. 8, pp. 57-63.

Efron, R. 1970c, 'The relationship between the duration of a stimulus and the duration of a perception', *Neuropsychologia*, vol. 8, pp. 37-55.

Eggermont, J. J. 2001, 'Between sound and perception: reviewing the search for a neural code', *Hearing Research*, vol. 157, pp. 1-42.

Eggermont, J. J. & Ponton, C. W. 2002, 'The Neurophysiology of Auditory Perception: From Single Units to Evoked Potentials', *Audiology and Neuro-Otology*, vol. 7, pp. 71-99.

Elberling, C. & Don, M. 1984, 'Quality estimation of averaged auditory brainstem responses', *Scandinavian Audiology*, vol. 13, no. 3, pp. 187-97.

Eling, P. A., Marshall, J. C. & van Galen, G. P. 1980, 'Perceptual centres for Dutch digits', *Acta Psychologica*, vol. 46, no. 2, pp. 95-102.

Epstein, M., Florentine, M. & Buus, S. 2001, 'Measuring loudness of long and short tones using magnitude estimation', paper presented to Fechner Day, Leipzig.

Fisch, B. J. 1999, *Fisch and Spehlmann's EEG Primer: Basic Principles of Digital and Analog EEG*, Third revised and enlarged edition edn, Elsevier Science, Amsterdam, The Netherlands.

Florentine, M., Epstein, M. & Buus, S. 2001, 'Loudness functions for long and short tones', paper presented to Fechner day, Leipzig.

Fowler, C. A. 1979, ''Perceptual centers' in speech production and perception', *Perception and Psychophysics*, vol. 25, no. 5, pp. 375-88.

Fowler, C. A. 1983, 'Converging sources of evidence on spoken and perceived rhythms of speech: cyclic production of vowels in monosyllabic stress feet', *Journal of experimental psychology. General.*, vol. 112, no. 3, pp. 386-412.

Fowler, C. A. 1996, 'Listeners do hear sounds, not tongues', *Journal of the Acoustical Society of America*, vol. 99, no. 3, pp. 1730-41.

Fowler, C. A., Smith, M. R. & Tassinary, L. G. 1986, 'Perception of syllable timing by prebabbling infants', *Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 814-25.

Fowler, C. A. & Tassinary, L. G. 1981, 'Natural measurement criteria for speech: the anisochrony illusion', in J Long & A Baddeley (eds), *Attention and Performance IX*, Erlbaum, Hillsdale, NJ., vol. 9, pp. 521-35.

Fowler, C. A., Whalen, D. H. & Cooper, A. M. 1988, 'Perceived timing is produced timing: a reply to Howell', *Perception & Psychophysics*, vol. 43, pp. 94-8.

Fox, R. A. & Lehiste, I. 1987a, 'Effect of unstressed affixes on stress-beat location in speech production and perception', *Perceptual and Motor Skills*, vol. 65, pp. 35-44.

Fox, R. A. & Lehiste, I. 1987b, 'The effect of vowel quality variations on the stress-beat location', *Journal of Phonetics*, vol. 15, pp. 1-13.

Fraisse, P. 1984, 'Perception and estimation of time', *Annual Review of Psychology*, vol. 35, pp. 1-36.

Friberg, A. & Sundberg, J. 1995, 'Time discrimination in a monotonic, isochronous sequence', *Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2524-31.

Fuentemilla, L., Marco-Pallares, J. & Grau, C. 2006, 'Modulation of spectral power and of phase resetting of EEG contributes differentially to the generation of auditory event-related potentials', *Neuroimage*, vol. 30, pp. 909-16.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. & Zue, V. 1993, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, University of Pennsylvania.

Gelfand, S. A. 1998, *Hearing: an introduction to psychological and physiological acoustics*, Third Edition edn, Marcel Dekker, Inc., New York.

Glasberg, B. R. & Moore, B. C. J. 2002, 'A model of loudness applicable to time-varying sounds', *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331-42.

Goebl, W. & Palmer, C. 2009, 'Synchronization of Timing and Motion Among Performing Musicians', *Music Perception*, vol. 26, no. 5, pp. 427-38.

Goldstein, R. & Aldrich, W. M. 1999, *Evoked Potential Audiometry: Fundamentals and Applications*, Allyn & Bacon, Boston.

*Goldwave* (version 5.52), Goldwave Inc., St. Johns, NL, Canada, <http://www.goldwave.com/>.

Gordon, J. W. 1987, 'The perceptual attack time of musical tones', *Journal of the Acoustical Society of America*, vol. 82, no. 1, pp. 88-104.

Grassi, M. & Darwin, C. J. 2001, 'Perception of the duration of ramped and damped sounds with raised cosine ramps', paper presented to Sixteenth Annual Meeting of the International Society for Psychophysics (Fechner Day 2001), Leipzig, Germany.

Gregory, A. 1978, 'Perception of clicks in music', *Perception & Psychophysics*, vol. 24, no. 2, pp. 171-4.

Grondin, S. 2001, 'From Physical Time to the First and Second Moments of Psychological Time', *Psychological Bulletin*, vol. 127, no. 1, pp. 22-44.

Gruber, W. R., Klimesch, W., Sauseng, P. & Doppelmayr, M. 2005, 'Alpha phase synchronization predicts P1 and N1 latency and amplitude size', *Cerebral Cortex*, vol. 15, no. 4, pp. 371-7.

Gurtubay, I. G., Alegre, M., Valencia, M. & Artieda, J. 2006, 'Cortical gamma activity during auditory tone omission provides evidence for the involvement of oscillatory activity in top-down processing', *Experimental Brain Research*, vol. 175, no. 3, pp. 463-70.

Harsin, C. A. 1993, 'Perceptual Centers and the Relation of Acoustic Energy Modulation to Speech Timing', Unpublished PhD Thesis thesis, University of New Orleans.

Harsin, C. A. 1997, 'Perceptual-centre modeling is affected by including acoustic rate-of-change modulations', *Perception and Psychophysics*, vol. 59, pp. 243-51.

Heil, P. 1997, 'Auditory Cortical Onset Responses Revisited. I. First-Spike Timing', *Journal of Neurophysiology*, vol. 77, pp. 2616-41.

Heil, P. & Neubauer, H. 2001, 'Temporal integration of Sound Pressure Determines Thresholds of Auditory-Nerve Fibers', *The Journal of Neuroscience*, vol. 21, no. 18, pp. 7404-15.

Hoequist, C. E., Jr. 1983, 'The perceptual center and rhythm categories', *Language and Speech*, vol. 26, no. 4, pp. 367-76.

Holmes, J. & Holmes, W. 2001, *Speech synthesis and recognition*, second edn, Taylor & Francis, London.

Hove, M. J., Keller, P. E. & Krumhansl, C. L. 2007, 'Sensorimotor synchronization with chords containing tone-onset asynchronies', *Perception and Psychophysics*, vol. 69, no. 5, pp. 699-708.

Howell, P. 1984, 'An Acoustic Determinant of Perceived and Produced Anisochrony', paper presented to 10th International Congress of Phonetic Sciences, Dordrecht.

Howell, P. 1988, 'Prediction of P-center location from the distribution of energy in the amplitude envelope: I', *Perception & Psychophysics*, vol. 43, pp. 90-3.

ISO/TC43 2003, *ISO226:2003, Acoustics – Normal equal-loudness-level contours*, International Organization for Standardization, Geneva, Switzerland.

ITU-R 2006, *Recommendation ITU-R BS.1770*, International Telecommunication Union, Geneva, Switzerland.

ITU-T 2001, *Recommendation P.862*, International Telecommunication Union, Geneva, Switzerland.

Janker, P. M. 1996a, 'Evidence for the p-center syllable-nucleus-onset correspondence hypothesis', *ZAS Papers in Linguistics (ZASPIL)*, vol. 7, pp. 94-124.

Janker, P. M. 1996b, 'The range of subjective simultaneousness in tapping experiments with speech stimuli', paper presented to ESCA Workshop on the Auditory Basis of Speech Perception, Keele University (UK), July.

Janker, P. M. & Pompino-Marschall, B. 1991, 'Is the P-Center position influenced by 'tone'?' paper presented to International Congress on Phonetic Sciences, Aix-en-Provence.

*Jython* 2006, (version 2.5.0), <http://www.jython.org/>.

Kim, W. S. & Han, S. K. 2006, 'Phase analysis of single-trial EEGs: Phase resetting of alpha and theta rhythms', *Neurocomputing*, vol. 69, p. 1337–40.

Klimesch, W., Hanslmayr, S., Sauseng, P. & Gruber, W. R. 2006, 'Distinguishing the evoked response from phase reset: A comment to Ma¨kinen et al.' *Neuroimage*, vol. 29, pp. 808-11.

Klimesch, W., Sauseng, P., Hanslmayr, S., Gruber, W. R. & Freunberger, R. 2007, 'Event-related phase reorganization may explain evoked neural dynamics', *Neuroscience and Biobehavioral Reviews*, vol. 31, pp. 1003-16.

Knief, A., Schulte, M., Bertrand, O. & Pantev, C. 2000, 'The perception of coherent and non-coherent auditory objects: a signature in gamma frequency band', *Hearing Research*, vol. 145, no. 1-2, pp. 161-8.

Kurby, C. A. & Zacks, J. M. 2008, 'Segmentation in the perception and memory of events', *Trends in Cognitive Sciences*, vol. 12, no. 2, pp. 72-9.

Lehiste, I. 1973, 'Rhythmic units and syntactic units in production and perception', *Journal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1228-34.

Low, Y. F. & Strauss, D. J. 2009, 'EEG phase reset due to auditory attention: an inverse time-scale approach', *Physiological Measurement*, vol. 30, no. 8, pp. 821-32.

Luo, H. & Poeppel, D. 2007, 'Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex', *Neuron*, vol. 54, pp. 1001-10.

Madison, G. S. & Merker, B. H. 2002, 'On the limits of anisochrony in pulse attribution', *Psychological Research*, vol. 66, pp. 201-7.

Madison, G. S. & Merker, B. H. 2004, 'Human sensorimotor tracking of continuous subliminal deviations from isochrony', *Neuroscience Letters*, vol. 370, no. 1, pp. 69-73.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E. & Sejnowski, T. J. 2002, 'Dynamic Brain Sources of Visual Evoked Responses', *Science*, vol. 295, no. 5555, pp. 690-4.

Mäkinen, V., Tiitinen, H. & May, P. 2005, 'Auditory event-related responses are generated independently of ongoing brain activity', *Neuroimage*, vol. 24, pp. 961-8.

Marcus, S. M. 1976, 'Perceptual Centres', unpublished PhD thesis, King's College.

Marcus, S. M. 1981, 'Acoustic determinants of Perceptual-center (P-Center) location', *Perception and Psychophysics*, vol. 30, pp. 247-56.

Martin, B. A., Tremblay, K. L. & Korczak, P. 2008, 'Speech Evoked Potentials: From the Laboratory to the Clinic', *Ear and Hearing*, vol. 29, pp. 285-313.

Mason, S. M. 2004, 'Evoked potentials and their clinical application', *Current Anaesthesia & Critical Care*, vol. 15, pp. 392-9.

Merker, B. H. 2000, 'Synchronous Chorusing and Human Origins', in NL Wallin, B Merker & S Brown (eds), *The Origins of Music*, MIT Press, Cambridge, MA, pp. 315-27.

Merker, B. H., Madison, G. S. & Eckerdal, P. 2009, 'On the role and origin of isochrony in human rhythmic entrainment', *Cortex*, vol. 45, no. 1, pp. 4-17.

Morton, J., Marcus, S. & Frankish, C. 1976, 'Perceptual Centers (P-centers)', *Psychological Review*, vol. 83, no. 5, pp. 405-8.

Murata, K., Nakadai, K., Yoshii, K., Takeda, R., Torii, T., Okuno, H. G., Hasegawa, Y. & Tsujino, H. 2008, 'A robot uses its own microphone to synchronize its steps to musical beats while scatting and singing', paper presented to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008). Nice, France, September.

Olejnik, S. & Algina, J. 2003, 'Generalized eta and omega squared statistics: measures of effect size for some common research designs', *Psychological Methods*, vol. 8, no. 4, pp. 434-47.

Palva, S., Palva, J. M., Shtyrov, Y., Kujala, T., Ilmoniemi, R. J., Kaila, K. & Naatanen, R. 2002, 'Distinct gamma-band evoked responses to speech and non-speech sounds in humans', *Journal of Neuroscience*, vol. 22, no. 4, pp. RC211:1-5.

Pastor, M. A., Artieda, J., Arbizu, J., Marti-Climent, J. M., Peñuelas, I. & Masdeu, J. C. 2002, 'Activation of human cerebral and cerebellar cortex by auditory stimulation at 40 Hz', *Journal of Neuroscience*, vol. 22, no. 23, pp. 10501-6.

Pastor, M. A., Vidaurre, C., Fernández-Seara, M. A., Villanueva, A. & Friston, K. J. 2008, 'Frequency-specific coupling in the cortico-cerebellar auditory system', *Journal of Neurophysiology*, vol. 100, no. 4, pp. 1699-705.

Patel, A. D., Lofqvist, A. & Naito, W. 1999, 'The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-Centre problem', paper presented to 14th International Congress of Phonetic Sciences, San Francisco.

Paulus, E. & Zwicker, E. 1972, 'Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Grequenzgruppenpegeln', *Acustica*, vol. 27, no. 253-266.

Perez, P. E. 1997, 'Consonant duration and stress effects on the P-centers of English disyllables', PhD thesis thesis, University of Arizona.

Pompino-Marschall, B. 1987, 'Segments, syllables, and the perception of speech rate and rhythm', paper presented to European Conference on Speech Technology, Edinburgh.

Pompino-Marschall, B. 1989, 'On the psychoacoustic nature of the P-center phenomenon', *Journal of Phonetics*, vol. 17, pp. 175-92.

Pompino-Marschall, B. 1990, *Die Silbenprosodie: Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo*, Max Niemeyer Verlag, Tübingen, Germany.

Pompino-Marschall, B. 1991, 'The syllable as a prosodic unit and the so-called P-centre effect', *Forschungsbericthe des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, vol. 29, pp. 65-123.

Pompino-Marschall, B. 2007, Re: Questions about your 1989 P-centre model to R Villing, 30 April.

Pöppel, E. 1997, 'A hierarchical model of temporal perception', *Trends in Cognitive Sciences*, vol. 1, no. 2, pp. 56-61.

Rapp-Holmgren, K. 1971, 'A study of syllable timing', *STL-QPSR*, vol. 12, no. 1, pp. 14-9.

Rasch, R. A. 1979, 'Synchronization in Performed Ensemble Music', *Acustica*, vol. 43, pp. 121-31.

Repp, B. H. 1995, 'Detectability of duration and intensity increments in melody tones: a partial connection between music perception and performance', *Perception and Psychophysics*, vol. 57, no. 8, pp. 1217-32.

Repp, B. H. 2002, 'Automaticity and voluntary control of phase correction following event onset shifts in sensorimotor synchronization', *Journal of Experimental Psychology: Human Perception and Performance*, vol. 28, no. 2, pp. 410-30.

Repp, B. H. 2005, 'Sensorimotor synchronization: a review of the tapping literature', *Psychon Bull Rev*, vol. 12, no. 6, pp. 969-92.

Roach, B. J. & Mathalon, D. H. 2008, 'Event-Related EEG Time-Frequency Analysis: An Overview of Measures and An Analysis of Early Gamma

Band Phase Locking in Schizophrenia', *Schizophrenia Bulletin*, vol. 34, no. 5, pp. 1-20.

Rodriguez, E., George, N., Lachaux, J. P., Martinerie, J., Renault, B. & Varela, F. J. 1999, 'Perception's shadow: long-distance synchronization of human brain activity', *Nature*, vol. 397, no. 6718, pp. 430-3.

Sanders, L. D., Newport, E. L. & Neville, H. J. 2002, 'Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech', *Nature Neuroscience*, vol. 5, no. 7, pp. 700-3.

Sauseng, P., Klimesch, W., Gruber, W. R., Hanslmayr, S., Freunberger, R. & Doppelmayr, M. 2007, 'Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion', *Neuroscience*, vol. 146, no. 4, pp. 1435-44.

Schlauch, R. S., Ries, D. T. & DiGiovanni, J. J. 2001, 'Duration discrimination and subjective duration for ramped and damped sounds', *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2880-7.

Schroeder, M. 1970, 'Synthesis of low-peak factor signals and binary sequences with low autocorrelation.' *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 85-9.

Schulze, H.-H. & Vorberg, D. 2002, 'Linear phase correction models for synchronization: parameter identification and estimation of parameters', *Brain and Cognition*, vol. 48, no. 1, pp. 80-97.

Schütte, H. 1978, 'Ein Funktionsschema für die Wahrnehmung eines gleichmässigen Rhythmus in Schallimpulsfolgen', *Biological Cybernetics*, vol. 29, pp. 49-55.

Scott, S. K. 1993, 'P-Centres in speech: an acoustic analysis', Unpublished PhD thesis thesis, University College London.

Scott, S. K. 1998, 'The point of P-centres', *Psychological Research*, vol. 61, pp. 4-11.

Seton, J. C. 1989, 'A psychophysical investigation of auditory rhythmic beat perception', PhD thesis, University of York.

Slaney, M. 1998, *Auditory Toolbox Version 2*, 1998-010, Interval Research Corporation.

Snyder, J. S. & Large, E. W. 2002, 'Neurophysiological correlates of meter perception: Evoked and induced gamma-band (20-60 Hz) activity', paper presented to Seventh International Conference on Music Perception and Cognition, Adelaide, Australia.

Snyder, J. S. & Large, E. W. 2004, 'Tempo Dependence of Middle- and Long-Latency Auditory Responses: Power and Phase Modulation of the EEG at Multiple Time-Scales', *Clinical Neurophysiology*, vol. 115, pp. 1885-95.

Snyder, J. S. & Large, E. W. 2005, 'Gamma-band activity reflects the metric structure of rhythmic tone sequences', *Cognitive Brain Research*, vol. 24, no. 1, pp. 117-26.

Stürzebecher, E., Cebulla, M. & Wernecke, K. D. 2001, 'Objective detection of transiently evoked otoacoustic emissions', *Scandinavian Audiology*, vol. 30, no. 2, pp. 78-88.

Tallon-Baudry, C. & Bertrand, O. 1999, 'Oscillatory gamma activity in humans and its role in object representation', *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 151-62.

Tallon-Baudry, C., Bertrand, O., Delpuech, C. & Pernier, J. 1996, 'Stimulus Specificity of Phase-Locked and Non-Phase-Locked 40 Hz Visual Responses in Human', *Journal of Neuroscience*, vol. 16, no. 13, p. 4240–9.

ten Hoopen, G., Nakajima, Y., Hartsuiker, R., Sasaki, T., Tanaka, M. & Tsumura, T. 1995, 'Auditory isochrony: time shrinking and temporal patterns', *Perception*, vol. 24, pp. 577-93.

Torrence, C. & Compo, G. 1998, *Wavelet Software* Boulder, CO., <http://atoc.colorado.edu/research/wavelets/>.

Tremblay, K. L., Friesen, L., Martin, B. A. & Wright, R. 2003, 'Test-Retest Reliability of Cortical Evoked Potentials Using Naturally Produced Speech Sounds', *Ear and Hearing*, vol. 24, pp. 225-32.

Tuller, B. & Fowler, C. A. 1980, 'Some articulatory correlates of perceptual isochrony', *Perception and Psychophysics*, vol. 27, no. 4, pp. 277-83.

Tuller, B. & Fowler, C. A. 1981, *The contribution of amplitude to the perception of isochrony*, SR-65, Haskins Laboratories, New Haven, CT.

van Noorden, L. P. 1975, 'Temporal coherence in the perception of tone sequences', Doctoral Dissertation thesis, Technisch Hogeschool Eindhoven.

Vidal, J., Bonnet-Brilhault, F., Roux, S. & Bruneau, N. 2005, 'Auditory evoked potentials to tones and syllables in adults: evidence of specific influence on N250 wave', *Neuroscience Letters*, vol. 378, no. 3, pp. 145-9.

Villing, R., Timoney, J. & Ward, T. 2006, 'Performance Limits for Envelope based Automatic Syllable Segmentation', paper presented to IET Irish Signals and Systems Conference, Dublin, Ireland, June 28-30.

Villing, R., Timoney, J., Ward, T. & Costello, J. 2004, 'Automatic Blind Syllable Segmentation for Continuous Speech', paper presented to Irish Signals and Systems Conference 2004, Belfast, June 30 - July 2.

Villing, R., Ward, T. & Timoney, J. 2003, 'P-Centre Extraction from Speech: the need for a more reliable measure', paper presented to Irish Signals and Systems Conference, Limerick.

Vos, J. & Rasch, R. A. 1981, 'The perceptual onset of musical tones', *Perception and Psychophysics*, vol. 29, no. 4, pp. 323-35.

Vos, P. G., Mates, J. & van Kruysbergen, N. W. 1995, 'The perceptual centre of a stimulus as the cue for synchronization to a metronome: evidence from asynchronies', *Quarterly Journal of Experimental Psychology*, vol. 48A, no. 4, pp. 1024-40.

Whalen, D. H., Cooper, A. M. & Fowler, C. A. 1989, 'P-center judgments are generally insensitive to the instructions given', *Phonetica*, vol. 46, no. 4, pp. 197-203.

Wikipedia Contributors 2009, *IPA chart for English dialects*, Wikipedia, The Free Encyclopedia, viewed 28 October 2009, <http://en.wikipedia.org/w/index.php?title=IPA_chart_for_English_dialects&oldid=322233491>.

Wong, P. K. H. & Bickford, R. G. 1980, 'Brain stem auditory evoked potentials: the use of a noise estimate', *Electroencephalography and Clinical Neurophysiology*, vol. 50, pp. 25-34.

Wright, M. 2008, 'The Shape of an Instant: Measuring and Modeling Perceptual Attack Time with Probability Density Functions', PhD Dissertation thesis, Stanford University.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. 2007, 'Event perception: a mind-brain perspective', *Psychological Bulletin*, vol. 133, no. 2, pp. 273-93.

Zacks, J. M. & Tversky, B. 2001, 'Event structure in perception and conception', *Psychological Bulletin*, vol. 127, no. 1, pp. 3-21.

Zanto, T. P., Large, E. W., Fuchs, A. & Kelso, J. A. S. 2005, 'Gamma-Band Responses to Perturbed Auditory Sequences: Evidence for Synchronization of Perceptual Processes', *Music Perception*, vol. 22, no. 3, p. 535–52.

Zanto, T. P., Snyder, J. S. & Large, E. W. 2006, 'Neural correlates of rhythmic expectancy', *Advances in Cognitive Psychology*, vol. 2, pp. 221-31.

Zimmer, K., Luce, R. D. & Ellermeier, W. 2001, 'Testing an new theory of psychophysical scaling: temporal loudness integration', paper presented to Seventeenth Annual Meeting of the International Society for Psychophysics (Fechner Day 2001), Leipzig, Germany.

Zwicker, E. & Fastl, H. 1999, *Psychoacoustics: facts and models*, Second updated edn, Springer series in information sciences, Springer, Berlin; New York.