P-Centre Extraction from Speech: the need for a more reliable measure

Rudi Villing[ϕ], Tomas Ward[c] and Joseph Timoney*

[ϕc]*Department of Electronic Engineering, NUI Maynooth, IRELAND*
E-mail: [ϕ] *rudi.villing@eeng.may.ie*
[c] *tomas.ward@eeng.may.ie*

* *Department of Computer Science, NUI Maynooth, IRELAND*
E-mail: jtimoney@cs.may.ie

Abstract – **P-Centres represent the perceptual moments of occurrence of acoustic signals and are an important parameter for duration modelling in speech synthesis applications. While a number of algorithms have been proposed previously to achieve P-Centre extraction it is shown in this paper that none yield reliable measures or are robust in implementation. A software only solution is presented which enables the rhythm setting experiment to be carried out in a reliable and flexible manner.**

Keywords – **P-Centre, Speech rhythm, Psychoacoustics.**

I INTRODUCTION

A current area of intensive research for concatenative speech synthesis systems is the creation of accurate and reliable models of speech prosody, meaning pitch, intensity and timing information. Prosody is essential for naturalness and is also indispensable to the listener attempting to form an adequate interpretation of the linguistic information being conveyed. This paper is focused on the timing aspects of prosodic modelling as still very little is known about the underlying processes that are present in natural speech [1].

A key problem in studying timing patterns is deciding how to measure speech unit duration in a way that reflects the perception of timing by listeners and the timing strategies of speakers [2],[3]. Measuring intervals between syllable onsets is unsatisfactory because abundant evidence shows that speakers and listeners, when asked to attend to the timing of syllables, are not concerned with the timing of onsets [4]. Over twenty years ago speech and psychology researchers posited that speakers and listeners gauge temporal intervals in speech on another basis labelled the perceptual center or "P-Centre" [5]. The P-Centre corresponds to a particular point within the syllable that *perceptually* feels like the syllable's "moment of occurrence" [5],[6]. It is thought that sequences of P-Centres underlie the perception and production of rhythm in perceptually regular speech sequences [6].

A significant difficulty with the P-Centre hypothesis is that a physical correlate has yet to be firmly established [6]. A number of algorithms have been proposed but their performance and robustness has not been verified independently. A robust algorithm could ultimately prove to be a key tool in duration modelling for speech synthesis technologies. This paper presents the results of ongoing work in developing a robust P-Centre algorithm. A series of experiments were performed to gather data. Two of the more recently-developed algorithms were tested to determine independently how well they performed the task of P-Centre prediction. Any observed shortfalls in their performance were analysed and avenues for further investigation suggested. The details of this work are given in the following sections.

II EXPERIMENTAL PROCEDURE

The experimental measurement of P-Centres suffers from a methodological problem: how can one measure where in time a participant hears a sound happen? Direct measures correspond to synchronising a finger tap or a reference sound with the perceived beat of a stimulus. Various motor function encoding and perceptual effects limit the accuracy of direct measures. The alternative is to use indirect measures such as the relative P-Centre [5]. This is the relative difference between the P-Centre of two sounds. To obtain relative P-Centres experimental participants are instructed to adjust the

time differences between a pair of sounds until it is perceived that they exhibit a regular rhythmic pattern. At this point it is assumed that the sounds have been aligned based on the perceptual `beats' (or P-Centres) of the stimuli. The intervals between the physical onsets of the stimuli can be readily measured. If the stimuli have different P-Centres, then the physical onset intervals will be anisochronous. This experimental procedure is called the *dynamic rhythm setting task*. Figure 1 illustrates how the subject alters the physical intervals between two repeating sounds, until perceptual isochrony is achieved. The final intervals chosen are recorded, and can be used to determine a `P-Centre fit' for each sound used in the experiment [5].

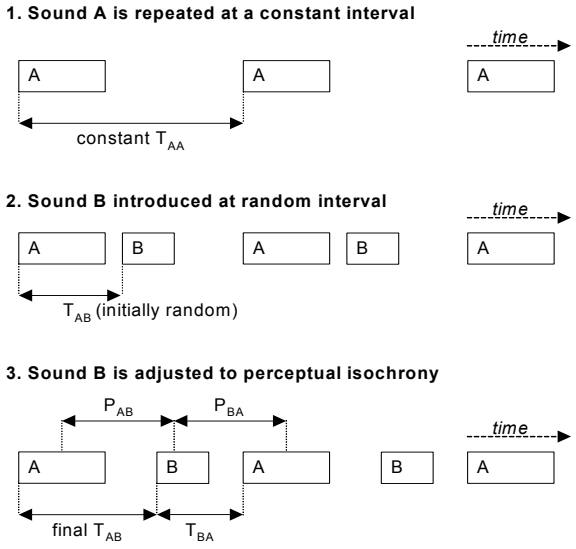

Figure 1: Outline of Indirect P-Centre Measurement.

The time interval between the repetitions of sound A is known and is constant. Sound B is inserted at a random point between A - A, so that the sequence is A - B - A - B - A and so on. The A - B interval is altered by the subject, until the A - B - A - B sequence is perceptually isochronous, that is, that the intervals between the sounds are perceived as equal. The time interval from the P-Centre of sound A to that of B, denoted P_{AB} should be equal to the interval from the P-Centre of sound B to that of A denoted, P_{BA}. Once the trial is halted, the A - B and B - A intervals are noted. The time interval from the onset of the waveform of sound A to that of sound B is labelled T_{AB}, and the time from the onset of sound B to sound A is T_{BA}. The times from physical onset of sounds A and B to their respective P-Centres are denoted P_A and P_B respectively. In the absence of measurement noise, the following idealized equations can be written:

$$P_{AB} = P_{BA} = \frac{T_{AA}}{2} \quad (1)$$

$$T_{AB} - P_{AB} = P_A - P_B \quad (2)$$

$$T_{BA} - P_{BA} = P_B - P_A \quad (3)$$

Let

$$D_{BA} = \frac{T_{BA} - T_{AB}}{2} = P_B - P_A \quad (4)$$

D_{BA} can be interpreted as the P-Centre of stimulus B relative to the P-Centre of stimulus A. In general the quantity D_{BA} is derived from two noisy measurements and so a least squares multiple linear regression is used to minimize the error when estimating it [6].

In most previous experiments the trial required the subject to adjust the relative distance between the sounds by turning a potentiometer knob that was finely calibrated [5],[6] although the use of keyboard buttons calibrated for coarse increments has also been used [7]. To modernise the experimental procedure and remove the need for a finely calibrated potentiometer and supporting A/D conversion, the P-Centre measurement program was written as a standalone software package in the Java language. An innovation devised to overcome the difficulties posed by the potentiometer unit was that the software allowed the user to adjust the sound interval by rolling the scroll wheel of a standard PC wheel mouse backwards or forwards to decrease or increase the A – B interval respectively. The mouse wheel was calibrated for a minimum adjustment of 2ms. Buttons were presented on the user interface for both fine (2ms) and coarse (50ms) adjustments. Finally the functions executed by the user interface buttons could also be accessed via keyboard hot keys. The software allows the user to start or stop the playback of the A – B stimuli at any time. While playing the stimuli are presented in a continuous loop separated by the currently adjusted intervals. As the user adjusts the intervals using the mouse wheel, user interface buttons or keyboard hot keys the intervals are dynamically updated for the next repetition of the loop. Figure 2 shows a screen shot of the user interface of the software.

Early use of this software indicated that the concentration levels required for the rhythm setting task can quickly lead to listener fatigue. For this reason the software was designed to allow subjects to complete a long trial over multiple sessions. The software records the session and cumulative results each time the user exits so that the user can return to where they might have finished previously. The results were automatically saved in a comma separated value (CSV) spreadsheet to ensure ease of access to the results by the experimenters.

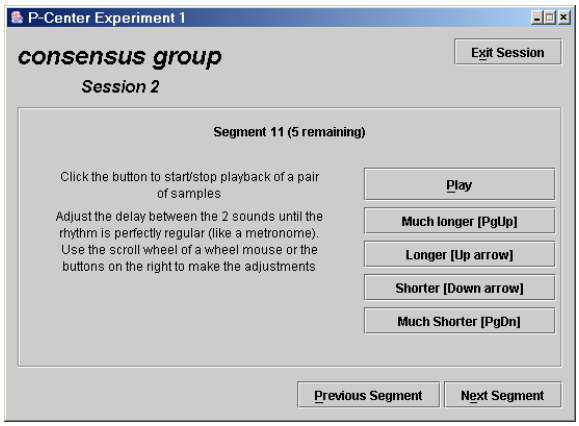

Figure 2 Screenshot of P-Centre Measurement Software

The software configuration used for all trials specified the constant A – A interval as 1200ms. The A – B interval was constrained to the range 300ms to 900ms. The software automatically chose an initial random interval in this range for each pair of stimuli to be presented.

Twenty seven sound stimuli were used in the trials. Four speakers, two male (rv, tw) and two female (fm, om) were recorded producing the stimulus tokens at a speaking rate corresponding to a "marching pace" (approximately two beats per second). This helped prevent the production of stimuli whose duration was too long to be used in the experiment. The tokens were recorded at a sampling rate of 11025 Hz with 16 bit resolution. A reference stimulus consisting of a 50ms white noise signal amplitude modulated with a 10ms linear on ramp, 30ms constant amplitude and 10ms linear off ramp envelope was generated synthetically. The stimuli were grouped into three sets.

Set 1 consisted of the reference stimulus and the token "one" produced by each speaker. Each stimulus was presented with every other stimulus in both the A and B position.

Set 2 consisted of the tokens "one", "two", "five" and "six" produced by each speaker. The token "you'll" produced by speakers fm and rv was also included. In this set the stimuli "two", "five", "six" and "you'll" were each presented with the stimulus "one" produced by the same speaker in both the A and B position.

Set 3 consisted of the tokens "one" and "two" produced by each speaker. A bandstop filter centered on 578Hz with bandwidth of 4 ERB was used to create bandstop filtered versions of the tokens "one" and "two" which were each presented with the unfiltered stimulus "one" produced by the same speaker in both the A and B position.

All stimuli were presented over Harmon/Kardon HK206 audio speakers in a quiet room. Initial trials were completed by subjects working independently. Each subject completed one trial of each of the three stimulus sets. The task was reported as difficult and standard deviation between subjects was deemed to be high at up to 45ms for certain stimuli. The experimental procedure was modified to allow subjects return to completed trials after a break and adjust the intervals if they felt it was necessary. This often resulted in minor adjustments which reduced the variance in results.

A final variant of the procedure utilized a number of subjects working together to align all stimuli in a trial. This variant known as the consensus approach was found to minimize the variance in results between the presentations of a pair of stimuli in the both the A and B positions. All results reported in this paper were produced by a consensus group of 3 native English speakers with varying musical training who were not naïve to the aims of the experiment (the authors).

A least squares multiple linear regression was performed on the results of stimulus set 1 to obtain a best estimate for each speaker's "one" relative to the reference stimulus. Equation (4) was used for the results of stimulus sets 2 and 3 to determine the P-Centre of each stimulus relative to the corresponding speaker's "one". Substituting the measurement of each speaker's "one" relative to the reference stimulus allowed all P-Centres to be expressed relative to the reference stimulus and thereby compared.

III AUTOMATIC P-CENTRE DETECTION

The performance of two recently proposed algorithms for P-Centre detection was examined. The first algorithm under investigation was that of Scott [5]. In her approach each speech sound was filtered using a seven-band GammaTone filter bank (4.0 ERB channel spacing and 4.0 ERB channel bandwidth). Each channel output was fully rectified and then smoothed using a 25Hz Butterworth filter. The time at which the channel amplitude reached 50% of its maximum value was denoted $50\%_{max\_amp}$. Scott used previously collected P-Centre data for 8 speakers producing the tokens "one" and "two" for modelling. A linear regression of experimentally measured P-Centre values found that the predictor $50\%_{max\_amp}$ of the channel centred at 578Hz was most significant. The final regression equation was:

$$y = -11.2 + 0.407(50\%_{max\_amp}) \quad (5)$$

The performance of this model against a number of stimuli was then evaluated by Scott and found to be a significant predictor of P-Centres generally. It should be noted that Scott reused the modelling data in her performance evaluation and this may have lead to a biased evaluation.

The second algorithm under scrutiny was that of Harsin [7]. His approach was based on the observed relationship between acoustic modulation and speech perception. Each speech sound was passed through a seven channel filter bank. The channel outputs were low pass filtered and

downsampled before raising to the 0.3 power to create a loudness envelope reflecting the power law relationship between signal intensity and loudness. The loudness envelope in each channel was converted to a psychoacoustic envelope by weighting modulation components in the range 3 to 47 Hz according to a perceived modulation magnitude function, effectively band pass filtering the loudness envelope. Harsin investigated the performance of a model he labelled the *per band magnitude-weighted velocity model* (BMVM). This model was essentially a temporal center of gravity model which incorporated peaks in the psychoacoustic envelope first derivative weighted by psychoacoustic envelope magnitude increment between peaks. He determined it was a significant predictor of P-Centres. The regression equation used was:

$$y = 9.3 + 1.12\ (BMVM) \qquad (6)$$

Harsin is unclear on a number of items which makes replication of his method difficult. Though he specifies a filter bank of six channels and makes reference to the work of Scott [5] he does not rigorously specify either the channel bandwidth or center frequencies. For this reason seven channels with center frequencies as specified in Scott's model were used. The bandwidth of each channel was equal to its center frequency. The filtering of the loudness envelope to produce a psychoacoustic envelope causes portions of the psychoacoustic envelope to be negative. He is unclear on whether he performs subsequent post processing of this envelope. In this paper the negative psychoacoustic envelope was clamped to zero. A second issue is that the negative portion of the envelope first derivative contains local maxima (peaks). These occur during sound offsets. Harsin does not make it clear whether these peaks and the negative envelope first derivative were ignored. In this paper the negative envelope first derivative was clamped to zero.

Both algorithms were implemented and tested in Matlab. The GammaTone filter bank used was taken from the Auditory toolbox of Malcom Slaney [10].

IV RESULTS

The P-Centres measured experimentally were all expressed relative to the P-Centre of the reference stimulus as described above. These relative P-Centres can be transformed to absolute P-Centres through the addition of a constant equivalent to the absolute P-Centre of the reference stimulus. Therefore a good P-Centre model should predict the measured P-Centres to within a constant offset. Stimulus set 1 was used exclusively to measure the relative P-Centres between the "one" stimuli and the reference stimulus. This stimulus set was not used for model prediction.

a) Scott's Model

Figure 3 compares the P-Centres obtained from Scott's model with the measured P-Centres for all stimuli in stimulus set 2. It can be seen that though the relationship appears linear Scott's model consistently under predicts the P-Centre. It is suspected that this is largely due to the 0.407 scaling factor in her regression equation.

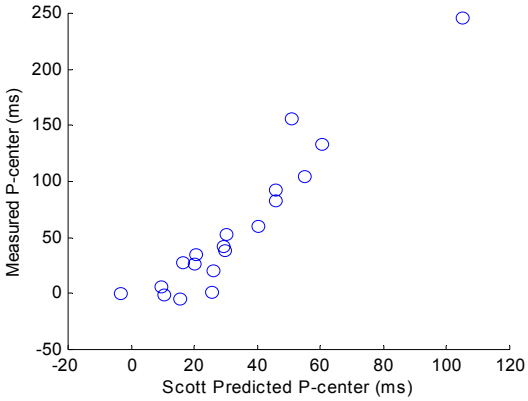

Figure 3: Comparison of P-Centres predicted by Scott's model with measured P-Centres in stimulus set 2

Based on the observation that the relationship of the data was approximately linear and in keeping with Scott's modelling approach it was decided to perform a new linear regression of the $50\%_{max\_amp}$ feature against the measured P-Centres for the "one" and "two" stimuli. The new regression equation was:

$$y = -65.6 + 1.07\ (50\%_{max\_amp}) \qquad (7)$$

This revised model was then evaluated for all stimuli other than those used for modelling. The results are plotted in Figure 4. The relationship between predicted and measured values appears linear though there is one outlier corresponding to the token "six" produced by the speaker tw who produced noticeably more drawn out stimuli than the other speakers. The error between predicted P-Centres and measured P-Centres is plotted in Figure 5 and Figure 6.

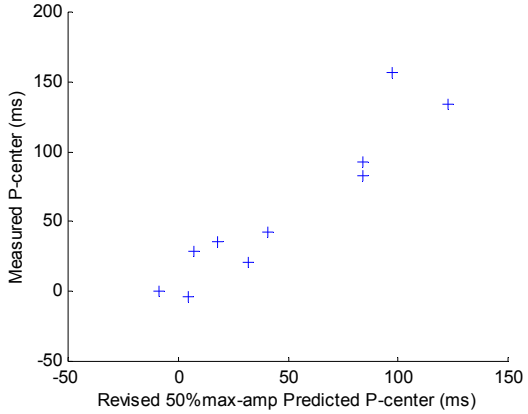

Figure 4 Comparison of P-Centres predicted by revised $50\%_{max\_amp}$ regression equation with measured P-Centres

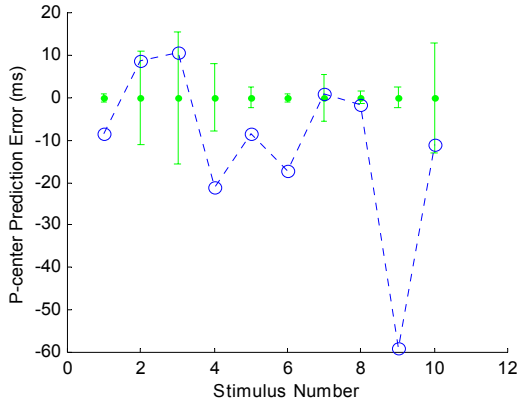

Figure 5 Comparison of revised 50%max_amp prediction error with measured P-Centres for stimulus set 2 excluding the "one" and "two" stimuli. Error bars represent measurement error range.

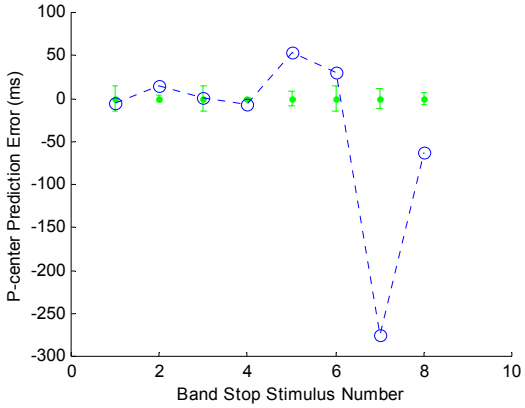

Figure 6 Comparison of revised 50%max_amp prediction error with measured P-Centres for stimulus set 3 (the band stop filtered stimulus set that removes energy in the band used by Scott's model). Error bars represent measurement error range.

There is a near linear relationship between the revised prediction model and the data and many predictions are within measurement error tolerances. It was the experience of the authors that P-Centre alignment errors of more than 20ms were generally perceptible. The number of such perceptible prediction errors suggests that Scott's model is too simplistic in its current form. Finally it is worth noting that the model (5) resulting from Scott's original linear regression did not perform well against the data used in this experiment while the new regression (7) performed better. This raises questions over the sensitivity of the $50\%_{max\_amp}$ feature which forms the basis of Scott's model to the data in use and whether it can in fact be the basis of a generally valid model.

b) Harsin's Model

Figure 7 illustrates the comparison of Harsin's predicted P-Centres with experimentally measured P-Centres.

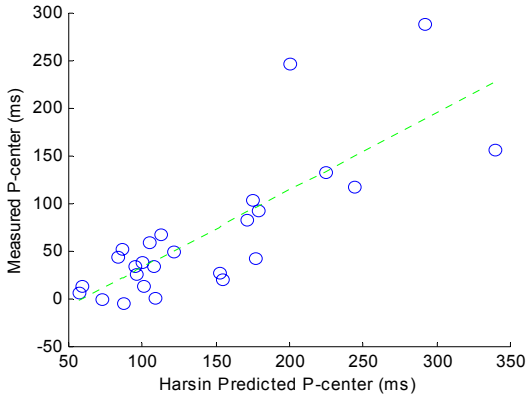

Figure 7 Comparison of P-Centres predicted by Harsin's Model and measured P-Centres for all stimuli

It is immediately clear that this model performs less well than Scott's model on the stimuli used. In particular the performance measured by Harsin was not replicated. There are a number of reasons why this might be the case. Harsin's model is ambiguous in a number of respects and so the possibility of implementation differences exists. Harsin's stimuli were primarily CV syllables ending with the vowel /a/ while the stimuli used in this experiment were primarily CVC syllables. Harsin's psychoacoustic envelope contains peaks at the offset of CVC syllables due to the filtering out of stable or slowly varying parts of the envelope (very low frequency modulations). These peaks at the offset tended to make the temporal center of gravity and hence the P-Centre prediction later. A linear regression of Harsin's prediction against measured P-Centres results in the regression equation:

$$y = -47.6 + 0.81(HarsinPrediction) \quad (8)$$

The linear fit is just moderate (R^2=0.638) and Harsin's model tends to predict P-Centres that are too late.

V DISCUSSION

The results of the previous section indicate that neither Scott's nor Harsin's models can be said to provide a consistent and robust technique for extracting P-Centres. The cochlea is known to decompose acoustic stimuli into frequency components along the length of the basilar membrane, a phenomenon known as tonotopic decomposition. It has been documented that modulations of the amplitude envelope are important for syllable perception [9]. Both Scott and Harsin's algorithms rely on analysing the amplitude envelope of the sound in filtered channels that are a coarse approximation to the cochlear filtering process. It is known that the nerve fibres emanating from a high frequency location in the cochlea "phase-lock" to the envelope of a stimulus around that frequency and

that to a first-order approximation, the pattern of nerve fibre activity around a tonotopic location conveys the AM or envelope information [8]. However, the complexities of this key property are not adequately incorporated into either algorithm.

Scott's primary feature is the time of a threshold crossing, where the threshold is set at 50% of the maximum amplitude for the stimulus. This threshold may be arbitrary and there is no explanation of why 50% should be used. For example the results of work by Vos and Rasch (cited in [5]) associate the P-Centre with the crossing of a threshold which lies between 6dB and 15dB below the maximum stimulus amplitude depending on the sound sensation level. Harsin's primary feature is slightly more complex but can be interpreted as a temporal center of gravity using weighted first derivative envelope peaks. Harsin's model is a global model which requires information from the complete stimulus before the P-Centre can be calculated. It would seem unlikely that the brain must wait for the stimulus end in order to determine the perceptual center. Both models are probably over-simplistic. It may be productive to incorporate known psychoacoustic effects such as equal loudness perception and masking in an effort to identify stimulus features which are naturally emphasised by the auditory system.

To support the invention of a more robust algorithm it would be worthwhile to include carefully controlled stimuli in the experimental procedure. There are also a number of experimental difficulties associated with isolating the acoustic correlates of P-Centres. P-Centres are perceptual and cannot yet be measured objectively. This tends to introduce measurement noise. Existing mechanisms for direct measurement suffer from a number of flaws while indirect measures only allow the calculation of a P-Centre relative to a reference signal. Reference signals such as a 5ms 1000Hz click [7] and 50ms ramped white noise [5] have been used but it is not clear where the absolute P-Centres of such stimuli lie. It is quite possible that the P-Centre of very short stimuli may be perceived after the stimulus has already ended. Researchers have tended to use a relatively small set of experimental stimuli such as the spoken digits [5],[6] and simple CV syllables using just one vowel type [7]. Some use has been made of synthetic stimuli when trying to isolate the effect of rise time on P-Centre but this use has been limited. It may be productive to perform the rhythm setting task with a set of synthetic stimuli whose envelope, timbre and pitch modulations could be carefully controlled. In particular, the effect of envelope modulations within a single critical bandwidth has not been adequately studied in the context of P-Centre research.

VI CONCLUSION

Using newly developed software for P-Centre testing, it was shown that two of the currently available algorithms for P-Centre extraction are unreliable. The shortcomings of each model were analysed and reasons for prediction errors proposed. It is apparent that the problem of automated P-Centre determination is still open and the authors aim to make progress in this area through the integration of additional results from psychophysiology and new experimental procedures.

REFERENCES

[1] O. Sayli, "Duration analysis and modelling for Turkish text-to-speech synthesis," M.S. thesis, Bogazici University, Turkey, 2002.

[2] P.A. Barbosa and G. Bailly. "Characterisation of rhythmic patterns for text-to-speech synthesis". *Speech Communication,* 15:127-137, 1994.

[3] P.A. Barbosa et al. "Airuete: a high quality concatenative text-to-speech system for Brazilian Portuguese with demisyllabic analysis-based units and a hierarchical model of rhythm". *Eurospeech 2001*, Aalborg, pp. 967-970, 2001.

[4] A. Patel et al. "The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-Centre problem". *ICPhS 1999*, San Francisco, pp. 405-408, 1999.

[5] S.K. Scott, "P-Centres in speech: an acoustic analysis," PhD thesis, University College London, 1993.

[6] S.M. Marcus. "Acoustic determinants of Perceptual-centre (P-Centre)," *Perception and Psychophysics.*, 30, pp. 247-256, 1981.

[7] C.A. Harsin. "Perceptual-centre modeling is affected by including acoustic rate-of-change modulations," *Perception and Psychophysics.*, 59, pp. 243-251, 1997.

[8] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech and Audio Proc.*, Vol.8, No. 3, pp. 240-254, May 2000.

[9] S. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech", *ICASSP 1997*, Munich, Germany, pp. 1647-1650, 1997.

[10] M. Slaney, "Auditory Toolbox Version 2," Interval Research Corporation, Technical Report #1998-010, 1998.