A Speech Feature Vector based on its Maximum Phase Component

J. Timoney, S. Feely, J.O'Kelly, and T. Lysaght
Department of Computer Science,
NUI Maynooth,
Maynooth,
Co. Kildare,
Ireland

Abstract
This paper examines the performance of a vowel classification scheme using a new form of feature vector derived from a decomposition of the speech segment into Maximum Phase and Minimum Phase components. Justification for this approach in terms of its perceptual relevance is first made, followed by a signal processing scheme to obtain the components. The form for the feature vector is then discussed. Lastly, experimental work compares the performance of this new feature vector under a variety of distortion conditions with the contemporary popular choice of Mel-Frequency Cepstral Coefficients.

1. Introduction

There have been suggestions in recent years that improvements in speech recognition technology can be attained if the dynamic properties of spoken language are modelled adequately [1] [2]. To achieve this requires a change from the traditional techniques of using Cepstral-based feature vectors. Occurring in parallel, there has been increased interest in analysing and modelling the amplitude and frequency modulation structure of speech as it attempts to overcome the deficiencies of linear speech models by indirectly describing the non-linear and time-varying phenomenon that occur during speech production [3] [4]. It is also motivated by better understanding of the signal processing function performed by the auditory periphery, particularly the cochlea. The cochlea is known to decompose acoustic stimuli into frequency components along the length of the basilar membrane. This phenomenon is called Tonotopic decomposition. It is also known that the nerve fibres emanating from a high-frequency location in the cochlea "phase-lock" to the envelope of the stimulus around that frequency, i.e. convey information about the envelope modulations in the signal. Thus, to a first-order approximation, it is often argued that the tonotopic location/place along the length of the basiliar membrane conveys the FM or frequency information about the signal, and the rate of nerve fibre activity around that location conveys the AM or envelope information [5]. In applying this AM-FM model to speech the approach in [6] further assumes that segments of the speech signal can be first decomposed into minimum phase (MinP) and maximum phase (MaxP) components. This decomposition was justified on the interpretation of particular phenomenon associated with the functioning of the auditory periphery. Evidence includes the auditory relevance of the left-sided spectrum of MaxP signals and the possibility that both the MinP and MaxP components can be represented in a discrete format by their zero/level crossings which could correspond to the information-bearing spikes in the auditory nerve fibres [6].

Although previous work has suggested that for vowel sounds the information carried by the MaxP component is very significant [6], this hypothesis was not thoroughly examined, and therefore, an attempt has been made to address it here. Thus, the intention in this paper is to examine the possibility of using a set of MaxP/MinP-based features extracted from the speech signal to the task of vowel classification, an essential component of the speech recognition process.

2. MinP/MaxP Speech Model

The model proposes that the speech signal can be represented as a periodic analytic signal $s(t)$ with period T seconds and of fundamental angular frequency $\Omega = 2\pi/T$ [5]. If $s(t)$ has finite bandwidth, it may be described for a sufficiently large M over an interval of T seconds by:

$$s(t) = e^{j\omega_i t} \sum_{k=0}^{M} a_k e^{jk\Omega t} \qquad (1)$$

where $e^{j\omega_i t}$ represents a frequency translation, and the a_k are the complex amplitudes of the sinusoids $e^{jk\Omega t}$. By analytic continuation $e^{j\Omega t}$ can be regarded as a complex variable (in the same fashion as the complex variable Z), that is, t the time variable is regarded as being complex valued [5]. It is possible to factor the M^{th} degree polynomial given by (1) into $M=P+Q$ factors, so that $s(t)$ can be rewritten as

$$s(t) = a_0 e^{j\omega_l t} \underbrace{\prod_{i=1}^{P} \left(1 - p_i e^{j\Omega t}\right)}_{\text{MinP}} \underbrace{\prod_{i=1}^{Q} \left(1 - q_i e^{j\Omega t}\right)}_{\text{MaxP}} \quad (2)$$

where the p_i's denote roots of the polynomial that lie inside the unit circle in the complex plane and the q_i's are roots that lie outside the unit circle. It is assumed that no roots actually lie on the circle so that $|p_i| < 1$ and $|q_i| > 1$. The factors corresponding to the zeros inside the unit circle, $\prod_{i=1}^{P} \left(1 - p_i e^{j\Omega t}\right)$, constitute a Minimum Phase signal while those corresponding to zeros outside the unit circle, $\prod_{i=1}^{Q} \left(1 - q_i e^{j\Omega t}\right)$, constitute the Maximum Phase signal [5]. These signals are the direct counterparts of the frequency responses of the minimum and maximum-phase FIR filters in discrete-time systems theory. This type of signal model has been referred to as a "product representation of signals" [4].

To apply this model to a real speech signal, it must be realised that the properties of the speech signal can change significantly over time. However, a reasonable assumption is that within a short-time interval these properties can be regarded as being stationary [7]. Therefore, successive overlapping *T*-second segments of a signal may be described using this model. Another issue is that a harmonic model can only be applied to what are termed voiced speech sounds, the spectrum of which exhibits a harmonic structure [7]. Since, the vowels belong to the category of voiced speech, it is justifiable to apply the model given by (1) to vowel sounds in order to derive features that can be used in a classification context.

3. Extraction of the MaxP component

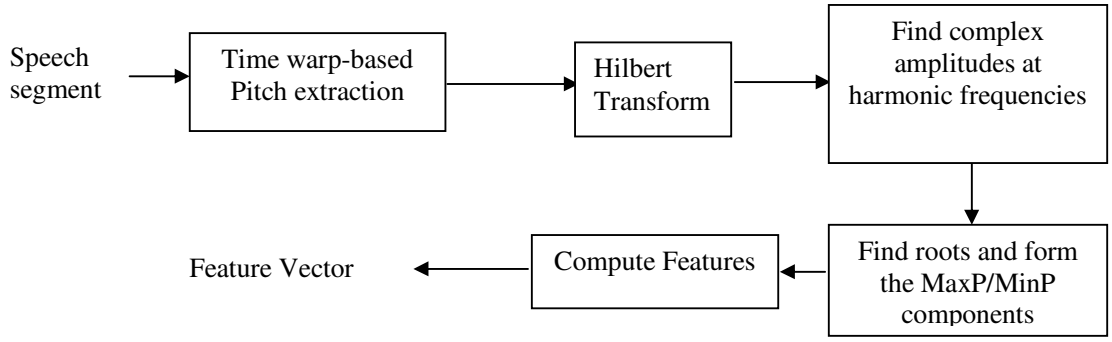

Fig. 1 Block Diagram for the conversion of speech segment to a MaxP/MinP feature vector

The procedure to extract the Maximum and Minimum Phase components from the speech and then create the feature vector representation is shown in Figure 1. The method is as follows. First, a high-resolution estimate of the fundamental frequency of the speech segment is obtained. The pitch detection algorithm is based on a parabolic time-warping procedure that effectively extracts the linear part of the pitch frequency variation from a voiced speech segment without affecting its time duration [8]. The form of the parabolic time warper is

$$\tau(t) = \frac{a}{T} t^2 + (1 - a)t, \quad 0 \le t \le T \quad (3)$$

in which *T* represents the duration of the speech segment, *t* represents real time, τ is warped time and *a* is the warping parameter.

The segment of speech is warped over a range of values of *a*, and the one producing the largest peak in an autcorrelation-based pitch detection scheme is retained [8]. The warped segment and the pitch value is then passed to the next stage where the complex amplitudes of the harmonic frequencies present in the segment are to be found. After taking the Hilbert transform of the warped segment, the Chirp z-transform [9] is employed to find the peak complex amplitude associated with each harmonic frequency. Furthermore, by taking the Chirp z-transform around the fundamental frequency a higher accuracy estimate of the pitch can be obtained. Once the complex amplitudes are known, the parameters of the speech model given by (1) can be filled. To generate the

MaxP and MinP components of the segment as given by (2), it is necessary then to find the roots of (1). This is achieved using a fast polynomial-rooting algorithm based on a combination of Muller's and Newton's method [10].

4. Creating a Feature Vector

Once the MaxP and MinP components are obtained the next step is to find a suitable form of feature vector that will capture their essential properties. By examining the MaxP component in the frequency domain it can be seen that it contains all the spectral information from DC to the greatest peak magnitude present, which in most cases corresponds to the location of the first formant. The MinP component therefore is formed from the peak magnitude and the remaining frequency components that exist in the spectrum, that is, those from the greatest peak up to the sampling frequency. Thus, the MaxP/MinP decomposition tends to split the spectrum of the waveform around the maximum peak present in it. Linking this with previous work [5], the MaxP/MinP model of (2) can also be described as

$$s(n) = A_c e^{\alpha(n)+\beta(n)+j\left(\hat{\alpha}(n)+\hat{\beta}(n)\right)} e^{j\left(\omega_c n - 2\hat{\beta}(n)\right)} \qquad (4)$$

where the "hat" stands for the Hilbert transform. A_c is a complex amplitude parameter, of the form $a_0 e^{j\phi}$. $\alpha(n)$ and $\beta(n)$ denote modulating quantities, and ω_c is a carrier frequency. From (4), it can be seen that the phases of the component signals, or equivalently the components' instantaneous frequencies, are essential aspects of the model. Thus, a possible feature set is the mean instantaneous frequency, that is the average of the time-derivative of the unwrapped phase, of the MaxP and MinP components for each segment. According to [5], the instantaneous frequency of the MaxP component will always be positive, a fact that is intuitively satisfying. This positivity is not necessarily the case for the instantaneous frequency of the MinP component. However, it was found that the trajectory of the instantaneous frequency curve of the MinP component can be much improved by first suppressing the frequencies lying close to DC. This MaxP/MinP feature vector was calculated for a synthetic version of the vowel /a/ which has its formant frequencies at 730, 1090 and 2440Hz respectively [7]. It was found that the mean instantaneous frequency of the MaxP component was approximately equal to the first formant frequency while the sum of the mean MaxP and MinP instantaneous frequencies was very close to the second formant frequency.

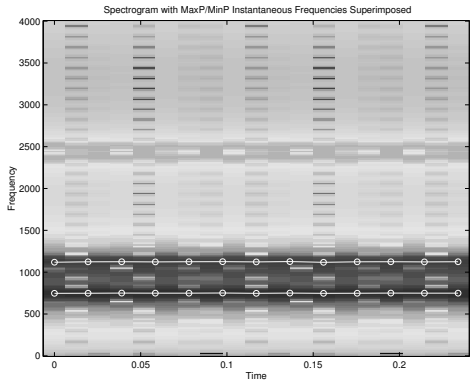

Figure 2 Spectrogram with MaxP/MinP Instantaneous Frequencies Superimposed

Figure 2 is a spectrogram of the synthetic vowel with these instantaneous frequencies superimposed on it as white lines marked with circles, and it is clear from the figure that the match is excellent.

A further inducement to use this form of feature vector comes from [1]. This work pointed out that there is evidence to show that complete knowledge of the formant frequencies is not required for accurate speech recognition. Moreover, [1] explains that results from perceptual experiments carried out by Fant and others appear to suggest that a two-formant approximation model (termed as perceptual effective formants) is a valid and robust framework for most vowels. Within these results, two prominent spectral peaks were found to be sufficient to describe all Swedish vowels. This effective formant model actually appeared to separate the vowel space better than a combination of the first two formants, and in addition, without the difficult requirement of an accurate formant tracking procedure. This format model was applied in a classifiaction procedure in [1] and the results indicated that the use of these perceptually effective formants conferred no disadvantages over any other choice of features. The use of the instantaneous frequencies of the MaxP and MinP components bears a

resemblance to the concept of perceptually effective formants as they are related to spectral maxima and also have perceptual validity. Thus, they should have potential for the vowel classification scenario.

5. Classification Task

To create a benchmark test within which the performance of the MaxP/MinP feature vector could be evaluated it was decided to use a set of synthesised vowels. The software to generate these vowels was found in [11] and the formant frequencies of the vowels were taken from [7]. Six different vowels were used and their formant frequencies are given in the table below

Vowel	F1 (Hz)	F2 (Hz)	F3 (Hz)
/i/	270	2290	3010
/E/	530	1840	2480
/a/	730	1090	2440
/c/	570	840	2410
/U/	440	1020	2240
/R/	490	1350	1690

Table 1: Formant Frequencies for the Synthetic Vowels

The pitch of these vowels was varied over the range 80 to 208 Hz in steps of 8Hz inclusive, and thus, eighteen copies of each vowel were generated in all. To classify the vowels the statistical Linear Discriminant Analysis procedure was used [12]. In order to rigorously test the performance of the classification, the vowels were subject to a series of distortions deemed to be typical of communications media: (1) additive noise, (2) peak clipping, (3) bandpass filtering and (4) reverberation [13]. These distortions were applied in various degrees, resulting in a total of 18 conditions, as specified in Table 2.

Distortion	Degree
Masking noise	SNR= 35, 25, 15, 10, 5 dB
Peak clipping	7,30,50,70, 90 % (cut-part/whole)
Band-pass filtering	0.8-1.3,1.3-1.9,1.9-2.6,1.4-3.2 kHz
Reverb	1.25ms (reflection coefficient 0.5 and 0.6), 6.25 and 12.5 ms (reflection coefficient 0.5)

Table 2: Distortions Applied to the Speech

By way of comparison, the performance of this feature vector was compared with the popular Mel Frequency Cepstral Coefficients (MFCC). Here, nine coefficients were generated for each speech frame in the segment, the first coefficient of each frame, which represents a transformation of the energy of the frame, was discarded, and their average was then taken over all the frames. The program to generate these MFCC feature vectors was obtained in [11].

6. Results

In carrying out the experiments, fifteen vectors, each of different pitch, for each vowel were used for training and the remaining three were retained for classification. Thus, the goodness of the vowel classification was evaluated using 18 input vectors and the percentage error was calculated to be the number of mis-classifications over the total input. This error percentage was calculated for each distortion condition and the results are shown in the four graphs in Figures 3 to 6. In each plot, the classification error for the MaxP/MinP feature vector is given by the solid line marked with triangles, while for the MFCC feature vector it is shown by the dashed line marked with circles.

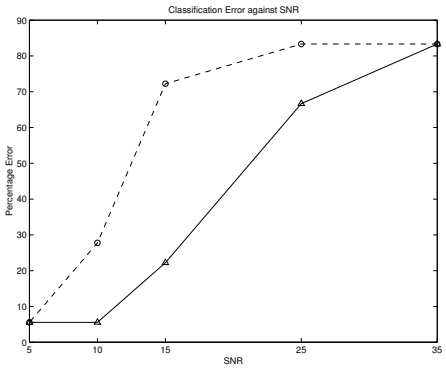

Fig.3 Classification Error for Noise

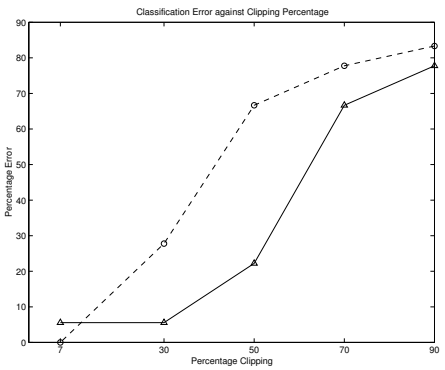

Fig. 4 Classification Error for Clipping

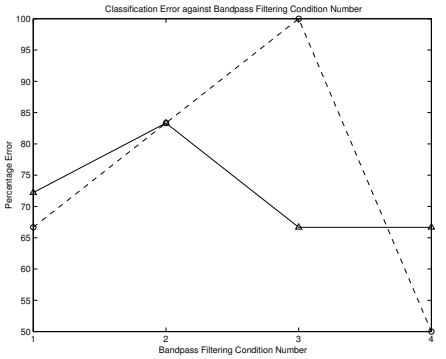

Fig.5 Classification Error for Filtering

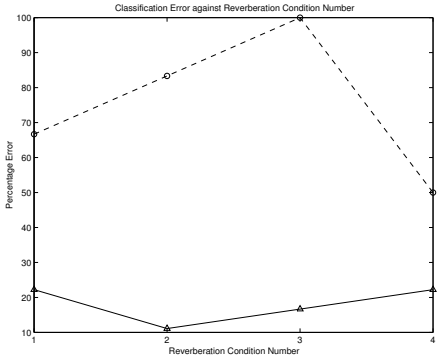

Fig. 6 Classification Error for Reverb

From the plots it can be seen that in most cases the MFCC-based feature vector outperforms the MaxP/MinP feature vector. The worst performance of the MaxP/MinP feature vector occurs in the presence of reverberation while its best performance, in relation to the MFCC feature vector, is for the bandpass filtering distortions. In the cases of noise masking and clipping the most redeeming quality of the MaxP/MinP feature vector is that in the worst case conditions its performance is either as poor as or better than that of the MFCC feature vector. Furthermore, in the case of noise masking, the classification performance of the two feature vectors is the same.

7. Conclusion

Overall, the results suggest that the MaxP/MinP feature vector is not as applicable as the MFCC feature vector to vowel classification under a range of common distortions, in particular for reverberation. However, given the relatively lower dimensionality of the MaxP/MinP feature vector, the results could be interpreted as showing that by extension of the feature vector with additional relevant information, it is possible that the performance could be brought to a level that is comparable with the MFCC feature vector.

Furthermore, it is possible that errors in the pitch detection process may be responsible for the lower performance level as a poor pitch estimate will result in inaccurate values for the complex harmonics amplitudes that are extracted in the subsequent processing stage. Immediate future work is to consider this and, if necessary, to find a pitch detection scheme that will overcome any problems found. A possible alternative to using the pitch detector-based scheme used in this work could be to find the harmonics using a Fractional Fourier transform approach [14]. Also of importance is to examine the augmentation of the MaxP/MinP feature vector with other relevant features derived from the model which may help to improve its classification performance. Lastly, another avenue for future work is to compare the performance of the MaxP/MinP feature vector with a feature vector derived from the perceptually effective formants [1]. Given the low dimensionality of both feature vectors, this would probably be a more fair comparison to make than with the MFCC feature vector chosen for this paper.

8. References

[1] Z. Hu and E. Barnard," Efficient estimation of perceptual features for speech recognition", *Eurospeech 1997*, Rhodes, Greece, vol. 1, pp. 493-496.

[2] L. Welling and H. Ney, "Formant Estimation for Speech Recognition," *IEEE Trans. Speech and Audio Proc.*, Vol.6, No. 1, Jan. 2000, pp. 36-48.

[3] A. Potamianos, '*Speech processing applications using an AM-FM modulation model*,' PhD. Thesis, Harvard University, Massachusettes, MA, Aug.1995.

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

[5] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journl. Acoust. Soc. Amer.*, vol. 105, no. 3, March 1999, pp. 1912-1924.

[6] R. Kumaresan and A. Rao, "Minimum/Maximum Phase decomposition of signals inspired by the auditory periphery," *Proc. Asilomar-29*, 1996, pp. 1239-1244.

[7] J. Deller, J. Proakis and J. Hansen, *Discrete-time processing of speech signals*, Macmillan, NewYork, 1993.

[8] R. Sluijter and A. Janessen, "A time warper for speech signals," *IEEE workshop on Speech Coding*, 1999.

[9] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[10] M. Lang and B. Frenzel, "Polynomial Root Finding," *IEEE Signal Processing Letters*, Oct. 1994.

[11] M. Slaney, *Auditory toolbox for Matlab*, http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/, 1998.

[12] The Mathworks inc., *Statistics toolbox for Matlab*, Prentice-Hall International, London, 2001.

[13] S. Wu and L. Pols, "A distance measure for objective quality evaluation of speech communication channels using also dynamic spectral features," *IFA Proc.*, Vol. 20, 1995, pp. 27-43.

[14] F. Zhang, Y. Chen and G. Bi, "Adaptive harmonic fractional Fourier transform," *IEEE Signal Processing Letters*, Vol. 6, No. 11, Nov. 1999.