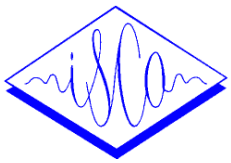

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# MINIMUM MEAN SQUARE ERROR SPECTRAL PEAK ENVELOPE ESTIMATION FOR AUTOMATIC VOWEL CLASSIFICATION

*Jaishree Venugopal, Stephen A. Zahorian and Montri Karnjanadecha*

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529, USA

10.21437/ICSLP.2000-461

## ABSTRACT

Spectral feature computations continue to be a very difficult problem for accurate machine recognition of vowels especially in the presence of noise or for otherwise degraded acoustic signals. In this work, a new peak envelope method for vowel classification is developed, based on a missing frequency components model of speech recognition. According to this model, vowel recognition depends only on the location of spectral peaks. Also, smoothing and interpolation of the sampled spectra, performed in the cepstral analysis method commonly used in automatic speech recognition results in a loss of valuable information. The new method for feature extraction presented in this paper is based on minimum mean square error curve fitting of cosine-like basis vectors to all peaks in the speech spectrum. A mathematical model for smoothly tracking spectral envelopes using only spectral peak information and ignoring other parts of the spectrum is presented. A software algorithm for the model was developed and tested for various speaker types using a neural network classifier. Vowel classification experiments were conducted based on the features derived from the spectral peaks. The classification rates of the peak method under various signal to noise ratios was also evaluated. The basic conclusion is that the new features perform the same as cepstral features for clean speech, but have advantages when the signal is degraded by noise.

## 1. INTRODUCTION AND BACKGROUND

Ever since the time of Peterson and Barney [1], the first three formants (F1, F2 and F3) have been regarded as the primary source of spectral information. Motivated by the idea that vowel information is primarily contained in the spectral peaks, but wanting a more robust method than formant tracking, Paul [2] tracked the spectral peaks by first computing the fundamental frequency. He then linearly interpolated between these peaks in the frequency domain to derive a spectral envelope, which better fitted the harmonic peaks in the spectrum than do smoothing methods based on the entire spectrum.

Very recently, a new model for vowel identification was proposed by Cheveigne and Kawahara [3]. They argued against smoothing and interpolation of the spectrum since it attempts to "guess" missing samples based on a predefined model and thus may be misleading. According to their theory, vowel identification (by humans) is a process of pattern recognition where matching is restricted to available data, and missing data are ignored using an F0-dependent weighting function that emphasizes regions near harmonics. Their theoretical arguments are based on human perceptual considerations. They did not extend or test their theory in the context of an automatic vowel classifier or recognizer.

The general objectives of this paper are to investigate a model for vowel identification based only on harmonic spectral peaks. More specifically, a mathematical model for smoothly tracking spectral envelopes using only harmonic spectral peak information, and ignoring other parts of the spectrum, is developed and presented. The model results in a set of cepstral-like features, except that the features are only derived from the spectral peaks. This model is illustrated with several spectral plots. The model is then tested for both clean and noisy speech for vowels for men, women, and children, using a neural network classifier. The theory and results are used to provide additional information regarding, the theory of Cheveigne and Kawahara [3].

## 2. ALGORITHM

The algorithm for curve fitting with cosine like basis vectors to peaks in the spectrum can be formulated as follows: First, let $x(n)$ be a vector of log spectral magnitudes, where typically $N$ is one half of the FFT length used for spectral estimation. Let $f_k(n)$ be an $N$ by $P$ matrix of Cosine Basis Vectors (CBV's). Then the goal is to approximate $x(n)$ using $\hat{x}(n)$,

$$\hat{x}(n) = \sum_{k=1}^{P} c_k f_k(n). \quad (1)$$

Selection of the coefficients $c_k$ (which we call Discrete Cosine Transform Coefficients--DCTCs, as in our previous work [4]) is based on minimizing the error between the original and the approximation. The Weighted Mean Squared Error $E$ between $\hat{x}(n)$ and $x(n)$ is given by

$$E = \sum_{n=1}^{N} [x(n) - \hat{x}(n)]^2 \, index(n) \quad (2)$$

where $index(n)$ is a vector of 0's and 1's used to select peaks in the spectrum. In particular, $index(n) = 0$, for those $n$ such that $x(n)$ is not a peak in the spectrum, and = 1, for those $n$ such that $x(n)$ is a peak in the spectrum.

The use of this index term is what differentiates this new method from methods which use the entire spectrum. If *index*(*n*) is 1 for all *n*, then the method is identical to that in Zahorian and Nossair [4], or essentially very similar to cepstral coefficients which depend on the entire spectrum. Note that as in this previous work, the basis vectors are "warped" to provide a Mel-like frequency resolution in this new work.

The objective is to compute the coefficients $c_k$ such that $E$ is minimized. Differentiating with respect to each of the coefficients and setting these derivatives equal to zero is used to solve this problem. Substituting for $\hat{x}(n)$ in (2) we first obtain

$$E = \sum_{n=1}^{N} \left[ x(n) - \sum_{k=1}^{P} c_k f_k(n) \right]^2 index(n) \tag{3}$$

Differentiating (3) with respect to the coefficients $c_m$, and letting each term equal zero. We obtain $P$ equations ($1 \le m \le P$)

$$-\sum_{n=1}^{N} 2 \left\{ x(n) - \sum_{k=1}^{P} c_k f_k(n) \right\} f_m(n) index(n) = 0 \tag{4}$$

Expanding and rearranging terms,

$$\sum_{n=1}^{N} x(n) f_m(n) index(n) = \sum_{n=1}^{N} c_k \sum_{k=1}^{P} f_k(n) f_m(n) index(n) \tag{5}$$

The $P$ equations from (5) can be further organized to obtain the matrix equation

$$\begin{bmatrix} B_1 \\ B_2 \\ .. \\ .. \\ B_p \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ .. \\ .. \\ c_p \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & .... & A_{1k} \\ A_{21} & .... & .... & A_{2k} \\ & & & \\ & & & \\ A_{p1} & .... & ..... & A_{pp} \end{bmatrix} \tag{6}$$

where

$$\begin{aligned} A_{ij} &= \sum_{n=1}^{N} f_j(n) f_i(n) index(n) \\ B_i &= \sum_{n=1}^{N} x(n) f_i(n) index(n) \\ &for \\ 1 &\le i \le P \\ 1 &\le j \le P \end{aligned} \tag{7}$$

"Standard" matrix methods can then be used to solve for the coefficient vector. Note that, unlike whole spectrum methods for computing cepstral (or DCTC coefficients), a matrix inverse is required. Although stability is, of course, an issue, if this algorithm is used with the "broad" peak method described in the next section, stability has not been a problem in practice. Stability will become a problem if an excessive number of coefficients are chosen relative to the number of peaks selected from the spectrum. The resultant problems are very similar to those encountered in polynomial curve fitting using a small number of data points and a higher order polynomial [5].

## Peak Picking

The method used to select peak regions in the spectrum can be described algorithmically as follows.

Consider a spectral vector $X(k)$, $1 \le k \le N$, as the speech spectrum for which the peak regions are to be identified. $k$ is considered as the frequency index. Define a frequency dependent window width for finding maxima in $X(k)$. This function is referred as $W(k)$. For each frequency index $k$, determine the maximum of $X(k)$ over the width determined by $W(k)$. That is, compute $Y(k) = \max(X(j), k - W(k)/2 \le j \le k + W(k)/2)$, for $W(1)/2 < k < W(N)/2$. For $k$ outside the range given above, define $Y(k) = X(k)$.

Next compare $X(k)$ and $Y(k)$, for $1 \le k \le N$, using delta as a parameter to compare closeness. If $(Y(k) - X(k)) < delta$, then $X(k)$ is close to a peak, and $index(k) = 1.0$. If $(Y(k) - X(k)) > delta$, then it is assumed that $\boldsymbol{X}(k)$ is not close to a peak, and $index(k) = 0.0$. Note that with very small modifications, this algorithm could also be used to find spectral dips, which it was used for a control case in some of the experiments reported below.

This overall method was implemented using three parameters: *delta*, which controls the width of each peak; *freq_rang_min*, which sets the width of the frequency window at low frequencies; and *freq_rang_max*, which sets the width of the frequency window at high frequencies. Typical values used were *delta* = 6 dB, *freq_rang_min* = 200 Hz, and *freq_rang_max* = 300 Hz. Note the frequency window width was linearly interpolated between the lowest and highest frequencies.

The overall algorithm is illustrated in Figure 1. The figure shows the FFT spectrum, the peaks (heavy dots), the spectrum smoothed by a DCTC analysis which uses the entire spectrum, and the spectrum smoothed by the peak-DCTC analysis method described in this paper. Note, as mentioned above, the algorithm used for peak picking selects broad peak regions (typically 3 to 5 points centered at each spectral peak), rather than individual points for each peak. For this particular example, the main difference between the two DCTC smoothed spectra is a level shift; however, there are also some small differences in the regions around the formant peaks.

# 3. EXPERIMENTAL VERIFICATION

## 3.1 Introduction

Vowel classification experiments were conducted to compare the peak DCTC features with our "standard" whole spectrum DCTC features. In each case, 12 DCTC features were computed for each frame of each vowel data scaled, and classified using a neural network classifier. The neural network classifier was a feed-forward, 1 hidden layer (12-25-10) network trained with 100,000 updates using back propagation. The vowels were spoken in isolation by adult male speakers (90 for training and 24 for testing), adult female speakers (100 training speakers and 24 test speakers), and child speakers (54 for training and 24 for

testing). More details of the database, signal processing and classification algorithm can be found in Zimmer et al.[6].

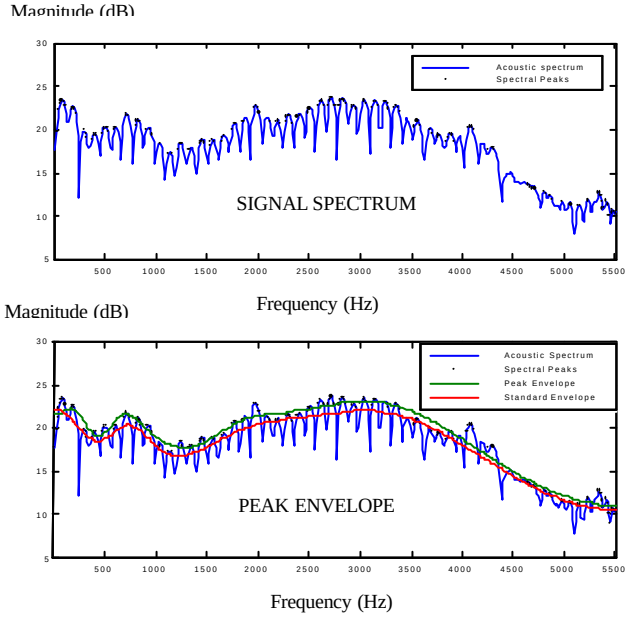


**Figure 1:** Illustration of harmonic spectral peaking and whole spectrum and peak spectrum DCTC analysis.

Several pilot experiments were conducted to establish good parameter settings for the adjustable variables in the processing. None of these variables caused large changes in results, and the results of those pilot experiments are not reported here. In the present paper we focus only the differences in results between the whole spectrum DCTCs and the peak spectrum DCTC features. Note that only test results are reported.

## 3.2 Experiment 1

The objectives of the first experiment were to evaluate some basic control and test conditions. Specifically, classification rates were obtained for the standard DCTC method without time smoothing, standard DCTC method with time smoothing, peak method without time smoothing, peak method with time smoothing, peak method tracking the valleys of the spectrum without and with time smoothing. Time smoothing was done by averaging over five frames for each token.

The testing was done for male, female, children, male&female vowels and the results are presented in Table 1. From the table, it can be concluded that the peak method performs comparable to the standard DCTC method. However classification rates obtained from valley envelope parameters, without time smoothing, were distinctly worse. Time smoothing of valley envelope information improved it considerably. Since time smoothing was found to generally improve classification rates, it was used in all further experiments.

## 3.3 Experiment 2

The objective of the second experiment was to examine performance as a function of signal-to-noise ratio (SNR) for the whole spectrum method and the peak envelope method. The signal to noise ratio used was varied in steps of 5 dB, from –10 dB to +25 dB, for each speaker type. The 25 dB SNR can be considered as clean speech since the noise level is still very low. This was done for the whole spectrum DCTC method and peak spectrum DCTC method for the male, female, and child speaker populations.

The plots of test classification rates versus SNR are shown in the figure 2 (male speakers), figure 3 (female speakers), and figure 4, (child speakers). It is observed that the peak method is superior to the whole spectrum DCTC method, particularly for low SNR values and for the female speakers. For the case of high SNR values, the two methods perform very similarly. Thus the biggest advantage to the peak spectrum features are for the case of noisy speech and high F0 voices, with widely spaced harmonics.

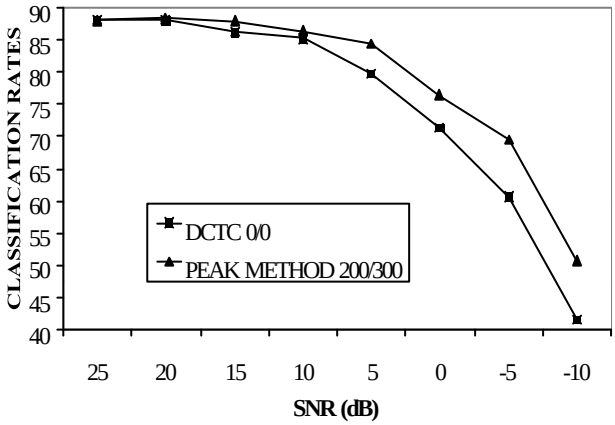


**Figure 2:** Vowel classification rates for male speakers for whole spectrum (DCTC) and peak spectrum (PEAK METHOD) as a function of SNR.

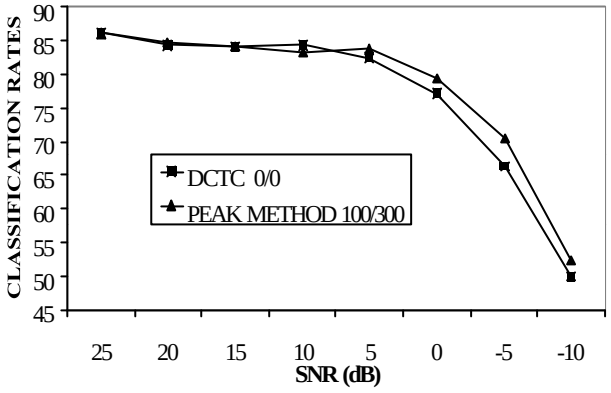


**Figure 3:** Vowel classification rates for female speakers for whole spectrum (DCTC) and peak spectrum (PEAK METHOD) as a function of SNR.

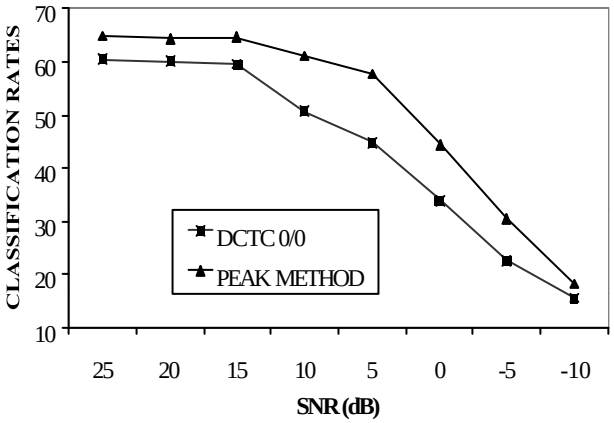


**Figure 4**: Vowel classification rates for child speakers for whole spectrum (DCTC) and peak spectrum (PEAK METHOD) as a function of SNR.

## 4. CONCLUSIONS

A mathematical model for curve fitting of cosine-like basis vectors to the peaks in speech spectra has been derived and tested in a series of vowel classification experiments. For the case of clean speech and closely spaced harmonics (male speech), vowel classification rates are nearly identical with either whole spectrum coefficients or peak-spectrum coefficients. However, for noisy speech, and particularly with more widely spaced harmonics (females and children), the peak method for coefficient calculation appears to be more robust. These data partially support the missing frequency components theory of Cheveigne and Kawahara [3], but not conclusively. A more complete test would involve the use of an accurate F0-tracking algorithm, so that peaks in the spectrum could be more accurately located.

## 5. REFERENCES

[1] Peterson G. and Barney H., "Control Methods Used in a Study of the Vowels," *J. Acoust. Soc. Am*., 24 (2),: pp.175-194, March 1952.

[2] Paul, D., " The Spectral Envelope Estimation Vocoder", in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol.29, No.4, pp. 786-793, August, 1981.

[3] Cheveigne, A. and Kawahara, H., " Missing-data model of vowel identification", *J. Acoust. Soc. Am*., 105 (6), pp. 3497-3508, June 1999.

[4] Zahorian S. A., and Nossair Z. B., "A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 414-425, July 1999.

[5] Lancaster, P. and Salkauskas, K., "Curve and Surface Fitting An Introduction" (Academic Press Ltd., Oval Road, London) 1986.

[6] Zimmer A., Dai, B., and Zahorian, S., "Personal Computer Software Vowel Training Aid for the Hearing Impaired", *Proc. ICASSP 98*, Vol. 6, pp. 3625-3628, 1998.

## 6. ACKNOWLEDGEMENTS

This work was partially supported by NSF grant BES-9977260.

| | Standard DCTC method | Standard DCTC method w/ time smoothing | Peak method | Peak method w/ time smoothing | Peak method tracking valleys | Peak method tracking valleys w/ time smoothing |
|---|---|---|---|---|---|---|
| male | 86.8 | 87.2 | 86.0 | 87.1 | 78.9 | 84.3 |
| female | 87.6 | 89.6 | 880 | 87.7 | 70.8 | 82.8 |
| children | 64.8 | 64.9 | 65.0 | 64.7 | 55.1 | 59.4 |
| male & female | 84.6 | 83.0 | 83.6 | 82.8 | 70.8 | 78.5 |

**Tabel 1**: Vowel classification rates for various speaker groups and various signal processing configurations.