ROBUST VOICED/UNVOICED CLASSIFICATION USING NOVEL FEATURES AND GAUSSIAN MIXTURE MODEL

Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno
Speech Processing Lab / ECE Dept. / Temple University / 1947 N 12th St., Philadelphia, PA 19122-6077 USA
Email: shah@temple.edu, aniyer@temple.edu, bsmolens@temple.edu, robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

ABSTRACT

Need for deciding whether a given frame of a speech waveform should be classified as voiced speech or unvoiced speech arises in many speech analysis systems. Several approaches have been described in the literature for making this decision. In this paper, we have presented two novel approaches of using acoustical features and pattern recognition. The first method is based on Mel frequency cepstral coefficient with Gaussian mixture model classifier, which resulted in approximately 90% identification accuracy and the other is based on LPC coefficient and reduced dimensional LPC residual with Gaussian mixture model classifier, which resulted in 92% identification accuracy. The performances of both approaches were compared for various levels of noise and optimum condition for training is determined.

1 INTRODUCTION

The classification of the short-time speech segments into voiced or unvoiced is critical in many speech analyses – synthesis systems. The essence of classification is to determine whether the speech production system involves vibration of the vocal cords [1][2]. The V/UV classification can be performed using a single feature, whose behavior could be significantly affected by the presence or absence of voicing activity. The accuracy of such an approach would not go beyond a certain limit, because the range of values of any single parameter generally overlaps between different categories. Although V/UN classification has been traditionally tied to the problem of pitch frequency determination, the vibration of the vocal cords does not necessarily result in periodicity in the speech signal. Therefore, a failure in the detection of periodicity in some voiced regions would results in V/UN errors.

In the first approach we have presented V/UN decision based on the cepstral coefficient which are used to describe the short-term spectral envelop of a speech signal. The cepstrum is the inverse Fourier transform of the logarithm of the short-term power spectrum of the signal. Using the logarithmic operation, the vocal tract transfer function and the voice source are separated. There are two ways to obtain the cepstral coefficients, FFT cepstral and LPC cepstral coefficient. In [3] the FFT based cepstral coefficient, the use of the Mel scale in the derivation of cepstral coefficient was introduced. It was shown in this study that such a scale improves the performance of phonetic recognition system over the traditional linear scale. Experiments were performed using the fist Mel frequency cepstral coefficient which represents energy term using threshold criteria and 12 Mel frequency cepstral coefficients with GMM classifier has been used to obtain voiced and unvoiced classification.

In the second approach we have used the linear prediction coefficient which assumes all pole model of the vocal tract [2][4]. Linear discriminant (LD) has been used to obtain the additional feature by having remaining information from the LPC residual. Although the actual distribution of the LPC's of phonemes is not well Gaussian distributed, their estimate is [4]. By having these assumptions GMM classifier has been used for classification.

The organization of the remainder of this paper is as follows. In section 2, we describe the basic algorithm to calculate Mel frequency cepstral coefficient, linear discriminant analysis, and Gaussian mixture model. Experimental conditions are given in section 3. Results and observations are presented in section 4. Section 5 provides the conclusions and references are given in section 6.

2 BACKGROUND

2.1 Mel Frequency Cepstral Coefficient

Many experiment has shown that the ear's perception to the frequency components in the speech does not follow the linear scale but the mel-frequency scale, which should be understood as a linear frequency spacing below 1kHz and logarithmic spacing above 1kHz [6]. So filters spaced linearly at low frequency and logarithmic at high frequencies can be used to capture the phonetically important characteristics (voiced and unvoiced) of the speech. The common used formula to approximately reflex the relation between the mel – frequency and the physical frequency (the known variation of the ear's critical band – widths with frequency) is given by,

$$M(f) = 1125 \log_{10}(1 + f/700)$$

where f is frequency in hertz. Based on the assumption of ear's perception to the Mel frequency scale, the Mel – frequency cepstrum is proposed in [3]. The system diagram to compute the MFCC and classification scheme is shown in Figure 1 and is briefly explained below.

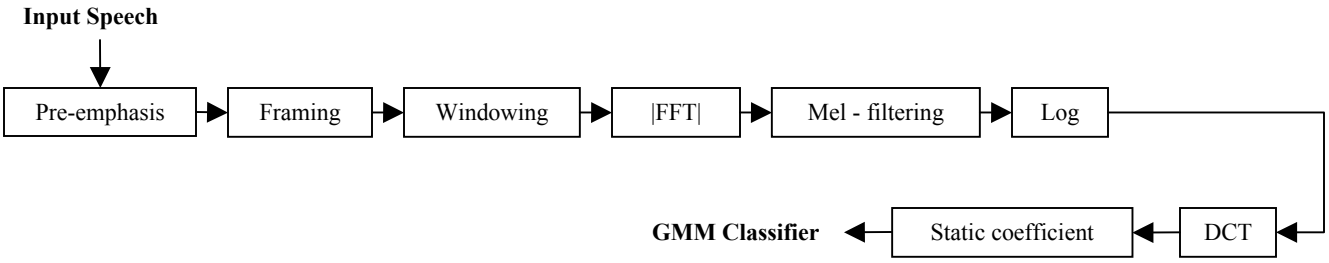

Figure 1: System Diagram Using the Mel Frequency Cepstrum Coefficient Feature With GMM Classifier.

A segment of speech (usually power of two to fit the FFT algorithm – in this experiment 256 sample points at 16kHz sampling rate) is hamming windowed and transformed to the frequency domain via the fast Fourier transform, and then the magnitude spectrum of the utterance is passed through a bank of triangular shaped filters whose center frequencies are spaced along the perceptually motivated mel frequency scale. The energy output from each filter is then log – compressed and transformed to the cepstral domain via the DCT. With assumption that phonetic information can be estimated using several Gaussian distributed functions, Gaussian mixture model (GMM) classifier has been used for classification.

2.2 LPC Feature with Linear Discriminant Analysis

In [5] it has been shown that residue as a whole carries richer information than the fundamental frequency alone. Linear discriminant analysis (LDA) was used in an attempt to capture all the remaining information left in the LPC residual to obtain the one additional feature. The goal of linear discriminant analysis is to use a linear transformation to project the set of raw testing data vectors onto a vector space of lower dimension such that some metric of class discriminant is maximized [7]. The technique attempts to maximize the between-class covariance S_b and minimize the within-class covariance S_w for a set of feature. The metric most often used is the ratio of the between class scatter to the within class scatter and that can be expressed as,

$$trace\{s_w^{-1} s_b\}$$

The LDA transform can be truncated to select only the *n* largest eigenvalues, the transformed features with the largest ratio of between class covariance to the within class covariance. By truncating the lower order LDA components, the dimensionality of the feature vector can be reduced. This transformation produces the 1-dimentisonal feature $\hat{y}$ from the LPC-residual data frames, which for this research where 256 samples (16 msec frames at 16 kHz sampling rate) in length. Hence, the transformation is from $\Re^{256}$ to $\Re$. The system block diagram of using LPC feature with linear discriminant analysis is shown in figure 2. We have used 4 LPC coefficients since by experiment it was determine that it contains the most useful information of vocal tract about voicing and their residual to obtain the remaining information. With the assumption that we can model this five dimensional feature vector into several Gaussians distributed functions, GMM classifier has been used for the classification.

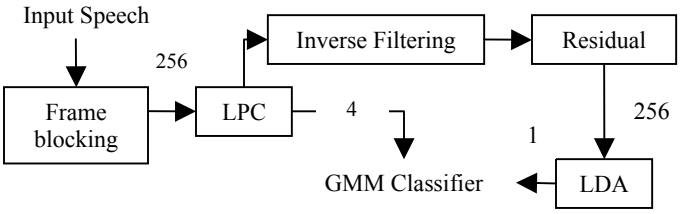

Figure 2: System Diagram of LPC Feature with LDA and GMM Classifier**.**

2.3 Gaussian Mixture Model

In a mixture model, a probability density function is expressed as a linear combination of several components (*pdfs*) [8]. A model with M components can be written in the form of

$$p(x) = \sum_{k=1}^{M} p(k)\, p(x \mid k)$$

Where the parameters $p(k)$ are called the mixing coefficients (weights) and the component density function $p(x|k)$ typically vary with k. A mixture of weights $p(k)$ has the form of an *a priori* probability of relatively importance of each component in mixture of *pdf*.

To obtain the parameters $p(k)$, the expectation maximization (EM) algorithm was used, which is an iterative implementation of maximum likelihood estimation using the incomplete information about the underlying probability distributions. 50 mixture components (M=50) were used, since this amount produced the lowest detection error rate.

3 EXPERIMENTAL CONDITIONS

Experiments were performed on 26 utterances, equal male and female from TIMIT labeled data. 17 utterances at 16 kHz sampling rate totals of 1501 frames used for training and 9 utterances containing 881 frames each of 256 sample points were used for testing. Labeled voiced data contains voiced and weak voiced regions. Unvoiced data includes unvoiced and silence regions.

4 RESULTS AND OBSERVATIONS

First Mel frequency cepstral coefficient which represents energy term provides good amount of information about voicing. Figure 3 shows the probability density distribution of first Mel frequency cepstral coefficient for voiced and unvoiced on clean speech data. Notice that the threshold of 10 provides a good

amount of separation between voiced and unvoiced distribution. Comparison of results using first Mel frequency cepstral coefficient with 10 thresholds and 12 Mel frequency cepstral coefficients with GMM classifier is shown in figure 4 for clean speech data.

Confusion matrix of using first Mel frequency cepstral coefficient for voiced and unvoiced is obtained as,

$$\begin{bmatrix} .91 & .9 \\ .90 & .10 \end{bmatrix}$$

Confusion matrix of using 12 Mel frequency cepstral coefficients and GMM classifier for voiced and unvoiced is obtained as,

$$\begin{bmatrix} .96 & .4 \\ .91 & .9 \end{bmatrix}$$

Notice that by using the 12 MFCC and GMM classifier we have increase the classification rate by 4% however the tradeoff is computational complexity.

Classification rate was obtained using the confusion matrix and in such a way that both voiced and unvoiced detection rate can be considered simultaneously. Classification rate can be obtained as: Classification rate = (Correctly detected voiced + correctly detected unvoiced) / (Correctly detected voiced + false rate of detecting voiced + correctly detected unvoiced + false rate of detecting unvoiced).

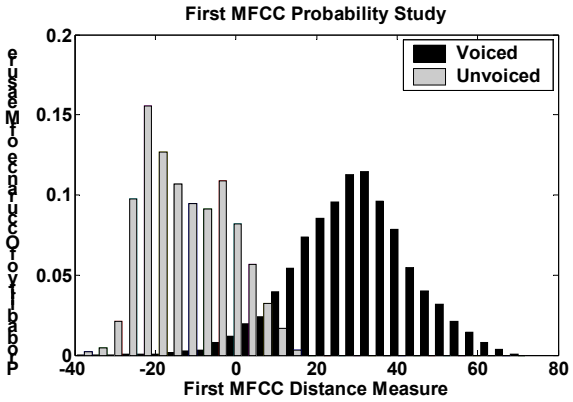

Figure 3: Probability Density of First Mel Frequency Cepstral Coefficient: Voiced speech (Black bars) and Unvoiced speech (Gray bars).

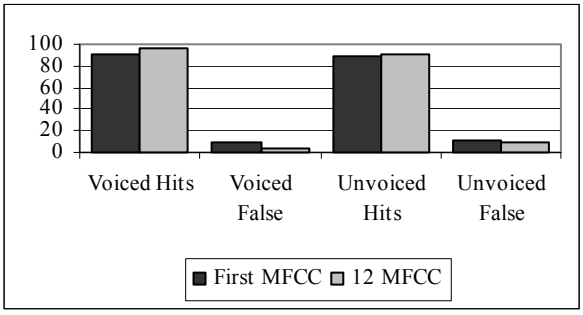

Figure 4: Comparison of Results Using First Mel Frequency Cepstral Coefficient and 12 Mel Frequency Cepstral Coefficients. First MFCC (Black bars) and 12 MFCC (Gray bars).

To verify the effect of noise on the algorithm, experiments were performed using 12 mfcc with various levels of noise. The classification rate is shown in Table 1. Notice that if one train the algorithm using clean speech and test under noisy condition may not obtain result very well instead we have determine if one wishes to use the algorithm under noisy situation, 15 dB SNR is reasonable lower limit for training. Notice from Table 1 that for training under 15dB SNR and testing under clean speech gives 85% and test under 15dB SNR gives 88% classification accuracy which is significantly better than the train under clean speech and test under noisy speech.

Table 2 shows the comparison of results of using 4 LPC coefficients with GMM classifier and 5 dimensional LPC feature with GMM classifier. Note that 5 dimensional LPC feature contains 4 LPC coefficient vectors and 1 feature obtained from LPC residual using Linear discriminant analysis. The algorithm was trained using 15 dB SNR since it was determined it is reasonable lower limit so that one can use under clean as well noisy environment.

Table 1: Comparison of results using 12 MFCC with GMM under various noisy conditions: Noisy signal obtained using white Gaussian noise**.**

		Testing			
Training	12 Dim MFCC with GMM		Clean	30dB SNR	15dB SNR
		Clean	94%	85%	62%
		30dB SNR	90%	90%	72%
		15dB SNR	85%	88%	88%

Table 2: Comparison of Results of 4 LPC and 5 Dimensional LPC Feature: GMM classifier has been used for classification. Noisy signal obtained using white Gaussian noise.

	Testing		
	Clean	30dB SNR	15dB SNR
4 LPC	80%	78%	60%
4 LPC + 1 LPC Residual	85%	88%	92%

It should be noted from Table 2 that LPC residual is information rich and can be used along with LPC coefficients to obtain phonetic information.

5 CONCLUSION

We have presented two approaches of detecting voiced and unvoiced speech. First is using 12 Mel frequency cepstral coefficient which represents short term energy spectrum expressed on a Mel – frequency scale and second is using LPC coefficient with novel method of reduced dimensional LPC residual feature. It was noticed that one can train the algorithm using 15dB SNR to obtain the optimum results instead of train using clean speech and test under noisy environment.

Further improvement can be made possible by using the by the mel scale discrete time energy operator so that one can

observe the changes in amplitude and frequency simultaneously while having mel frequency scale.

In our experiments we have obtained MFCC and LPC coefficient frame by frame basis. In a frame where there are less than few pitch periods in duration may give a false detection. Instead if one can identify each segment by detecting endpoint, one can obtain the coefficient on entire segment instead of frame by frame basis and hence may be able to improve the detection rate.

ACKNOWLEDGEMENT

The effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-02-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose notwithstanding any copyright annotation thereon. We extend out thanks to Uchechukw Ofoegbu for critical review and suggestions.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily represents the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

6 REFERENCES

[1] B. S. Atal and L. R. Rabiner, *"A Pattern Recognition Approach to Voiced – Unvoiced – Silence Classification with Applications to Speech Recognition",* IEEE Transaction on Acoustics, Speech and Signal Processing (1976), Vol: 24(3), pp: 201-212.

[2] L. R. Rabiner and R. W. Schafer, *"Digital Processing of Speech Signals"*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[3] S. B. Davis and P. Mermelstein, *"Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences",* IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol 28(4): pp: 357-366.

[4] S. M. Kay, *"Fundamentals of Statistical Signal Processing"*, Englewood Cliffs, NJ: Prentice-Hall, 1998.

[5] M. Faundez-Zamuy and D. Rodriguez-Porcheron, *"Speaker Recognition Using Residual Signal of linear and Nonlinear Prediction Models"*, Vol 2: pp: 121-124, ICSLP 98, Sydney.

[6] T. F. Quatieri, *"Discrete – Time Speech Signal Processing: Principles and Practice",* Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[7] K. Fukunaga, *"Introduction to Statistical Pattern Recognition"*, San Diego, CA: Academic Press, 1990.

[8] R. O. Duda, "*Pattern Classification",* Second Edition, New York, NY: John Wiley & Sons, Inc, 2001.