Word and Speaker Recognition using Concept of MFCC and Correlation Coefficients

Akshay Ramesh[1] Jaykumar Chaudhary[2] Sayantan Chakraborty[3] Nieves Crasto[4]
[1,2,3,4]Department of Electronics & Telecommunication Engineering
[1,2,3,4]Mukesh Patel School of Technology Management and Engineering, NMIMS (Deemed-to-be University)

Abstract— Speech and speaker recognition has permeated into everyday technology. Most recognition systems are based on machine learning techniques employing deep neural networks for the classification. This paper deals with providing a simple solution to yes/no speech classification using Fourier Transform and histogram based thresholding. Most automatic speech recognition system use Mel Frequency Coefficients (MFCCs) are features for classification. In this paper instead of using MFCCs as features directly, correlation coefficients between MFCCs of different frames are used as features and classification is carried out using Euclidean distance based template matching. An accuracy of 85% and 98.375% was achieved yes/no classification and speaker recognition respectively.
Key words: Word and Speaker Recognition, MFCC, Correlation Coefficients

I. INTRODUCTION

Speech signals have enormous capacity of carrying information. The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 5 kHz. The property of speech signal changes markedly as a function of time. Speech and speaker recognition is based on an individual's speech characteristics that cannot be stolen, forgotten or lost, thus allowing for a secure method of authenticating speakers.

During the past six decades various feature extraction and feature matching methods were introduced to represent the voice signal. The method for modeling the human auditory perception system, Mel Frequency Cepstral Coefficients (MFCCs) is discussed in this paper as feature extraction technique. Due to its advantage of less complexity in implementation of feature extraction algorithm, certain coefficients of MFCC corresponding to the Mel scale frequencies of speech Cepstrum are extracted from spoken word samples in database [7]. Also in this paper we have used Fast Fourier Transform method for specific word recognition.

II. YES/NO CLASSIFICATION

The basic concept in classifying 'yes' and 'no' speech signals is based on the fact that the power spectral density of the 'yes' signal has more energy in the high frequencies compared to the 'no' signal because of the 's' sound in yes (Figure 1). Therefore the sum of the magnitude of the low frequency DFT components to those of the high frequency components has been used as the classifying feature a. If the

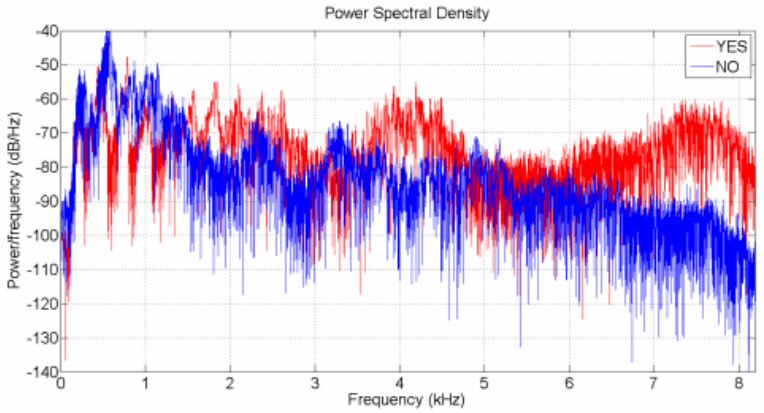

Fig. 1: PSD of yes and no speech signals.

	Yes	No
Yes	9	1
No	2	8

Table 1: Yes/N O Classification Confusion Matrix.

N-point DFT of an input signal x[n] where n = 0; 1; 2;:::; N-1 is given by,

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2\pi kn}{N-1}} \qquad k = 0, 1, 2, \ldots, \text{N-1} \tag{2.1}$$

Then the classifying feature a is computed by,

$$a = \frac{\sum_{k=0}^{N/4} |X[k]|}{\sum_{k=N/4}^{N/2} |X[k]|} \quad (2.2)$$

A recording that contains the word 'no' will usually have a higher value than would be found if the recording contained 'yes' i.e. a no > a yes. Since this feature is a ratio, its value is not affected by different volume levels. For classification, 50 samples each of 'yes' and 'no' from 6 different people in different environments were recorded. 80% of these samples were used for determining the thresh-old and the remaining were used for testing the accuracy of the system. The histogram, shown in figure 2, clearly shows a marked difference in the feature values for 'yes' and 'no'. Thus the data can be easily separated by using a single threshold value. The calculated threshold value was tested on the remaining 20% of data samples and it provided an accuracy of 85%.

III. SPEAKER RECOGNITION

Mel frequency cepstral coefficients (MFCC) is probably the best known and most widely used for both speech and speaker recognition [1].

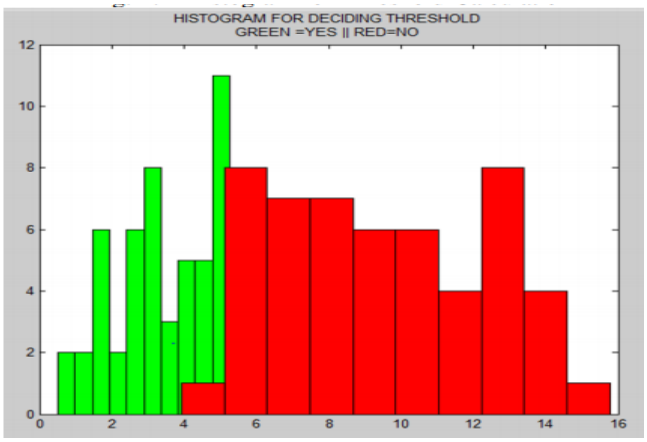

Fig. 2: Histogram for Threshold Calculation.

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1 kHz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [1, 3, 4]. MFCC has two types of filters which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000 Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The following formula is used to compute the Mels for a particular frequency, f

$$\text{mel(f)} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (3.1)$$

A. Framing and Windowing

The input speech signal is segmented into frames of 20 ms with overlap of 50% of the frame size. Usually the frame size (in terms of sample points) is equal to power of two in order to facilitate the use of FFT [2]. If this is not the case, we need to do zero padding to the nearest length of power of two. Windows are basically used in speaker recognition to remove discontinuities in speech. While extracting MFCC the window attenuates both ends of the frame which removes the abrupt changes at the ends. The window function is convoluted with the input signal. The Hamming window (Eq. 3.2) is used which provides reasonable side lobe and main lobe characteristics which are required for the DFT computation.

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (3.2)$$

B. Fast Fourier Transform

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame. When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around.

C. Mel Frequency Warping

In this step, DCT is applied on the energy E_k obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients.

$$C_m = \sum_{k=1}^{N} \cos\left[\frac{m(k-0.5)\pi}{N}\right] E_k \qquad m = 1, 2, ..L \quad (3.3)$$

Where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. In all 24 mel frequency coefficients were chosen. Since FFT has been performed, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients [12], or MFCC. MFCC alone can be used as the feature for speech recognition. For better performance, the correlation coefficients are calculated on the input matrix X whose rows are the frames and columns represent the 24 MFCCs.

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \quad (3.4)$$

Where C(i; j) is the covariance between the MFCCs of the I th and j th frame and is calculated by,

$$C(i,j) = \mathrm{E}[(i-\mu_i)(j-\mu_j)] \quad (3.5)$$

The mean across all the frames then constitutes the feature vector which is used for speaker recognition using template matching. The Euclidean distance between an input feature vector X and a template feature vector, Y referred to as a global distance[10, 9] is calculated using,

$$d = \sqrt{(X-Y)(X-Y)^T} \quad (3.6)$$

IV. Results And Conclusion

The template feature vector of 'yes' and 'no' for two speakers were stored. After classifying the input signal as yes/no, the Euclidean distance of the input from the template of speaker 1 d 1 and speaker 2 d 2 is calculated. The minimum of these two values determines the speaker's identity. By using just the MFCCs as the feature vector, the classification accuracy for speaker 1 was 100% while that of speaker 2 was 91.6%. Istaed by using the correlation coefficients an overall accuracy of 98.375% was achieved with speaker 1 having an accuracy of 99.6% and speaker 2 having an accuracy of 97.15%. Using the correlation coefficients of MFCCs as features, an improvement of 1.375% was observed in the identification of speaker 2 with a marginal decrease in the accuracy of speaker 1 identification. Testing these results on a larger dataset will help in consolidating the fact if correlation coefficients of MFCCs provide a higher accuracy in speaker recognition.

References

[1] S. Furui, "Speaker-independent isolated word recognition using dy-namic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, pp. 52-9, Feb. 1986.
[2] C.D. Bei and R.M. Gray. An improvement of the minimum distortion encoding algorithm for vector quantization. 'IEEE Transactions on Communications., October (1998).
[3] R. Vergin, "An algorithm for robust signal modeling in speech recog-nition," Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), Vol. 2, pp. 969-972, May, 1998.
[4] B. P. Lathi, Modern Digital and Analog Communication Systems, California state universtiy, 1998.
[5] J. R. Deller, J. H. L. Hansen and J. G. Proakis, Discrete Time Processing of Speech Signals, IEEE Press, 2000, 56-63 380-385 .
[6] Premakanthan and W.B. Mikhael, Speaker verification/recognition and the importance of selective feature extraction: Review, Proceedings of the 44th IEEE 2001,Midwest Symposium, 1: 14-17(2001).
[7] Molau, S, Pitz, M, Schluter, R, and Ney, H., Computing Mel-frequency coefficients on Power Spectrum, Proceedings of IEEE ICASSP-2001, 1: 73-76(2001).
[8] W. C. Chu, Speech Coding Algorithms, John Wiley and Sons, Wiley-IEEE, 2003.
[9] S. E. Levinson, Mathematical Models for Speech Technology, John Wiley & Sons Ltd., University of Illinois at Urbana-Champaign, USA, 2005.
[10] R. S. Kurcan, Isolated Word Recognition From in-ear Microphone Data Using Hidden Markov Models (HMM),M. Sc. Thesis, 2006.
[11] Remote Speaker and Speech Recognition A senior design project De-partment of Electrical Engineering University of California, Riverside Prepared by: Isaac Saldana David Ginsberg Faculty advisor: Yingbo Hua.
[12] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, S. Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition," IEEE signal processing letters, Maryland ,vol.12, 2011, pp 80-84.
[13] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthaworn-suk 'Speech Recognition using MFCC' International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July 28-29, 2012 Pattaya (Thailand)
[14] Nisha V.S, M.Jayasheela,? Survey on Feature Extraction and Matching Techniques for Speaker Recognition Systems?, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 2, Issue 3, March 2013.