Effect of Plosives on Isolated Speaker Recognition System Performance

Zekeriya ŞENTÜRK[1,2], Özgül SALOR[2]

[1]Department of Electronics Engineering, Turkish Military Academy, Ankara, Turkey
[2]Department of Electrical and Electronics Engineering, Gazi University, Ankara, Turkey
zsenturk@kho.edu.tr, salorozgul@gazi.edu.tr

Abstract

In this paper, the effect of keyword choice including and excluding plosive sounds on isolated speaker recognition system is investigated. In order to perform this study, a Turkish word database has been created consisting of 48 words including plosives and 7 words without plosives. Records are acquired at a sampling frequency of 16 kHz in a professional recording studio, with sound insulation. The records have been acquired during three or four sessions, achieved at different times of the day, for each participant to reflect the sound variability of the human vocal tract on the database. A speaker recognition system employing Mel-Frequency Cepstrum Coefficients (MFCC) for feature extraction and Dynamic Time Warping (DTW) for time equalization has been developed. After the system training stage, average speaker recognition performances for the keywords in the test set including plosives and excluding plosives has been found to be % 98.24 and % 91.76, respectively.

1. Introduction

Speaker recognition systems are used for checking up on who the speaker is. There are several areas using speaker recognition systems as security applications, banking transactions and so on and the system design can be different for different applications. Speaker recognition systems based on isolated word recognition are usually used for secure applications, which make robustness a must for these types of systems. Therefore, keyword choice for these systems makes up an important part of the operation as well as the choice of feature extraction and classification algorithms. Various speaker recognition systems have been reported in the literature [1-5]. In speaker recognition systems, there are many different methods for modelling speech. For instance, Hidden Markov Model (HMM) which is used for modelling speaker characteristics is very popular statistical method and has been used to achieve high recognition rate in speaker recognition systems for years [6]. Also, Gaussian Mixture Method (GMM) is another widely used method for modelling speaker identity [7].

In the study presented in this paper, the performance of the designed speaker recognition system is measured and compared using keywords including plosives and excluding plosives. In order to achieve this goal, an isolated speaker recognition system using Mel Frequency Cepstrum Coefficients (MFCC) as features has been designed and developed. Then, speaker recognition performance has been measured for the two separate sets of keywords, with and without plosives. It has been observed that keywords including plosives have better recognition performance than the keywords excluding plosives.

2. Speaker Recognition System

An isolated speaker recognition system has been designed in the scope of this work. Block diagram of the designed system is given in Fig. 1. In the training block, first records of the keyword set are acquired and MFCC features are extracted. First records are used as reference signal for Dynamic Time Warping (DTW) algorithm. Then, other training records are taken and MFCC features are acquired in block 4 after the equalization of the records with reference records via DTW algorithm (block 2). After that, in the block 4, mean of the features are calculated and every participant's each word's features are saved as a vector. By the step of 4, the train stage is done. Two thirds of the recordings are utilized for the train stage.

In the test process, size of test signal is equalized with reference signals of each participant for desired keyword respectively (Block 5). After equalizing, each equalized signal's feature vector is calculated in block 6 and compared with the feature vectors taken from train stage (block 7). In evaluating block, Euclidian distance is used for comparing the feature vectors. After evaluating, system decides who the owner of the test recording is.

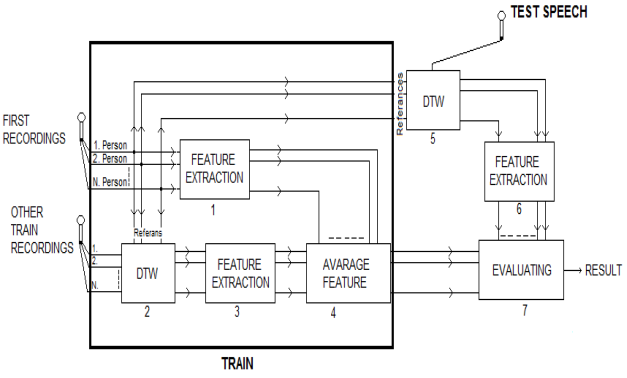

Fig. 1. Isolated Speaker Recognition System

Having studied the results, test keywords including plosives has better recognition rate than other keywords. 31 out of 48 keywords including plosives have one hundred percent recognition rate (Figure 2), although none of the 7 keywords which exclude plosive voice have hundred percent recognition rates (Figure 3).

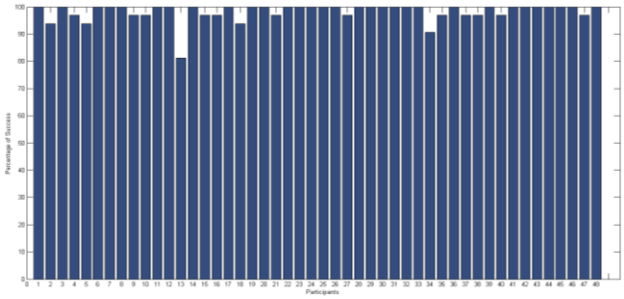

Fig. 2. Percentage Success of Keywords Including Plosives

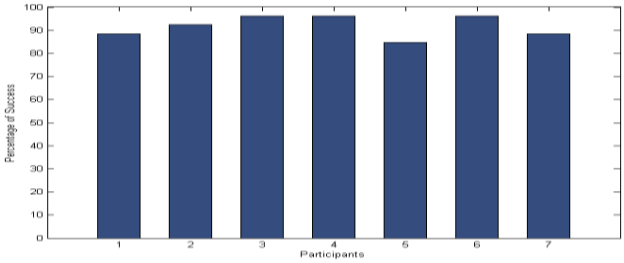

Fig. 3. Percentage Success of Keywords Excluding Plosives

2.1. Mel-Frequency Cepstrum Coefficient (MFCC)

Mel-frequency cepstrum coefficient (MFCC) which is commonly used for feature extraction in speech processing models human ear perception dissimilar to other technics using human vocal tract modeling (Reynolds 1995). Stevens and Volkman (1940) have formed Mel scale through experiments. The equation 1 is used for acquiring answer of any frequency in mel scale.

$$mel(f)=2595\log(1+f/700) \quad (1)$$

In MFCC feature extracting process, firstly, speech signal is segmented into small parts which can be assumed stationary. After that, every part are filtered with a finite impulse pre-emphasis filter of which transfer function is below.

$$H(z)=1-0{,}95z^{-1} \quad (2)$$

$$w(n)=0{,}54-0{,}46\cos(2\pi n/N-1),\ 1 \leq n < N \quad (3)$$

The filtered parts of signal are windowed by windowing function Hamming (eq. 3) due to prevent from spectral leakage during frequency transform operation. After windowing, signal segments are transformed to frequency domain via fast fourier algorithm. Then segments are filtered by mel filter bank (Fig. 4).

The number of filter used in filter bank gives the number of feature obtained from MFCC on every segmentation end every filter in the filter bank is located %50 overlap with neighboring segments (Figure 5). After mel filter bank filtering, the values extracted each filter in filter bank are transformed logarithmic scale. Finally, every logarithmic value is transformed by discrete cosine (DCT) algorithm and after all final MFCC's are acquired for each segment. Figure shows the MFCC extraction procedure.

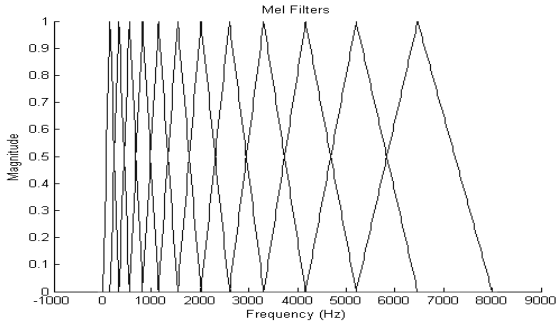

Fig. 4. Mel Filter Bank

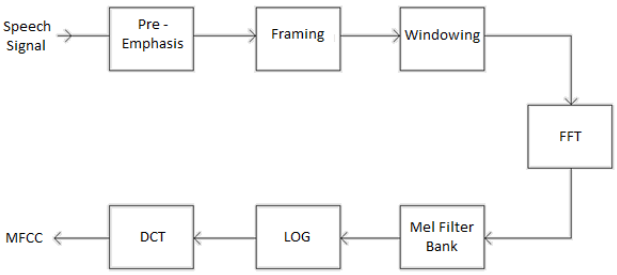

Fig. 5. MFCC Feature Extraction

In this study, mel filter bank is fitted 100Hz-8000Hz range and 13 filters are used for each filter bank. Hamming function is used for windowing the segments.

2.2. Dynamic Time Warping (DTW)

It is very common that the same speaker says same words in varied time durations. It poses feature matching problem in speaker recognition systems using fixed point feature extracting algorithm. To eliminate this problem, the recordings have to be fixed to the same size.

Dynamic time warping which is a pattern matching algorithm finds optimum alignment between two time sequences [8]. The algorithm investigates minimum cost path between two speech waveform having the same content (Figure 6). By this means, both training stage and testing phase feature vector sizes can be equalized.

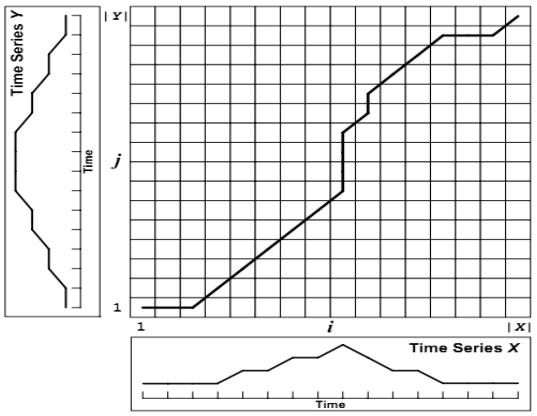

Fig. 6: Minimum-distance warping path for two time series [9]

3. Database

The Turkish Database used in this study has been created in a professional recording studio. In order to create the database, 48 words that include plosives and 7 words which exclude plosives are vocalized by 16 and 13 participants respectively. Each word is recorded 6 times by each participant at different times of day to reflect the sound variance on the database. All the recordings have been recorded 16 kHz sampling frequency and 16 bit quantization level.

4. Conclusions

In the study presented in this paper, the performance of the designed speaker recognition system is measured and compared using two different dataset including plosives and excluding plosives. In order to achieve this goal, an isolated speaker recognition system using Mel Frequency Cepstrum Coefficients (MFCC) as features and Dynamic Time Warping (DTW) for time equalization has been designed and developed. Then, speaker recognition performance has been measured for the two separate sets of keywords, with and without plosives. It has been observed that keywords including plosives have better recognition performance than the keywords excluding plosives.

It is obvious that, in speaker recognition systems, feature extraction and classification technics are crucial stages that influence performance of the systems. Besides that, choosing keywords used for isolated speaker recognition systems have importance alongside of feature extraction and classification algorithm. In this paper, it is shown that, keywords including plosives which have both unique spectral domain and time domain characteristic are more convenient than keywords which exclude plosives in isolated speaker recognition systems.

In addition to these studies, the keywords which have highest recognition rates can be analyzed phonetically and a database pool involving selected words can be created. Thus, created isolated speaker recognition systems can be more robust and can be used for more confidential security applications.

5. References

[1] T. May, S. van de Par, A. Kohlrausch, "Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling". *IEEE Transactions on Audio, Speech, and Language Processing*, vol. *20*, no. *1*, pp. *108-121*. January, 2012.

[2] M. Limkar, B.R. Rao, V. Sagvekar, "Speaker Recognition using VQ and DTW". in *International Conference on Advances in Communication and Computing Technologies*, ICACACT. 2012, pp.*18-20*.

[3] M. Sher, N. Ahmad, M.Sher, "TESPAR Feature Based Isolated Word Speaker Recognition System". in *International Conference on Automation and Computing*, ICAC . 2012, pp. *1-4*.

[4] G. Ekim, N. İkizler, A. Atasoy, İ.H. Çavdar, "A Speaker Recognition System Using by Correlation". in *IEEE Signal Processing, Communication and Applications Conference*, Aydın, SIU. 2008, pp. *1-4*.

[5] S. Nakagawa, W. Zhang, M. Takahashi, "Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM". in *Acoustics, Speech, and Signal Processing Conference*, ICASSP. 2004, pp. *I-81-I-84*.

[6] Savic, M., Gupta, S., "Variable parameter speaker verification system based on Hidden Markov Modeling, in proceedings of ICASSP'90, pp. 281–284, 1990.

[7] Tseng, B., Soong, F., Rosenberg, A., "Continuous probabilistic acoustic map for speaker recognition", in proceedings of ICASSP'92, vol.II, pp. 161–164, 1992.

[8] L. R. Rabiner, B. Juang, "Fundamentals of Speech Recognition ". Prentice-Hall, 1993.

[9] S. V. Chapaneri, "Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping". *International Journal of Computer Applications*, vol. *40*, no. *3*, pp. *6-12*. February, 2012.

[10] S.S. Stevens, J. Volkman, E Newman, "A Scale for the Measurement of the Psychological Magnitude of Pitch". *Journal of the Acoustical Society of America*, vol. *8*, no. *3*, pp. *185-190*. 1937.

[11] Wojcicki, K. (2011, September 19). *Mel frequency cepstral coefficient feature extraction in Matlab*. Avaible: http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab/content/mfcc/example.m

[12] Ellis, D. (2003). *Dynamic Time Warp (DTW) in Matlab*. Avaible:http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/