KURDISH RECOGNITION SYSTEM DIGIT

Özlem BATUR DİNLER

Siirt University, Department of Computer Engineering, Siirt-Turkey

o.b.dinler@siirt.edu.tr

Nizamettin AYDIN

Yildiz Tecnical University, Department of Computer Engineering, Istanbul-Turkey

nizamettin@ce.yildiz.edu.tr

Abstract: The security problems arising with the developing technology have put secure systems in quest of being developed with new techniques. The voice processing technology among these techniques has gained great importance. This situation has directed the reliability ensuring the man-machine interaction together with the voice processing applications such as speech recognition, speaker recognition, converting audio to text or converting text to audio to the design of high and correct systems. In this study, a digit speech recognition system consisting of the audio examples of digits from 0 to 9 in Kurdish spoken in Turkey was realized. For this purpose, audio recordings were obtained from 102 adult speakers who speak Kurdish. The feature vectors of the audio data acquired from these recordings were obtained with the Mel-Frequency Cepstral Coefficient by using different windowing methods, and the success rates were compared by conducting the recognition process with Dynamic Time Warping according to the feature vectors obtained.

Keywords: Speech processing, Kurdish, MFCC, DTW.

Introduction

Nowadays, speech recognition systems play an important role especially in personal applications that require password verification. Therefore, such applications have led to the widespread use of speech recognition systems and the development of speech recognition systems of different languages and dialects based on different voice features. This development first occurred in English, and then its applications were developed in French, Spanish, German, Danish, Japanese and Chinese. The aim of this study is to develop a speech recognition system in the Kurdish language.

Kurdish is an inflected language within the Indo-Iranian language group. Kurdish is the most spoken language after Turkish in Turkey, after Arabic, Turkish and Persian in the Middle East and Asia (Khan & Lescot., 1971, Wikipedia). The fact that a Kurdish speech recognition system to be created in this context can be used by wide masses also makes this article important.

Speech is a complex biometric signal produced as a result of various transformations that occur at the acoustic and articulation level (Trivedi., 2013). Speech recognition is the automatic extraction of linguistic information by obtaining acoustic models from speech signals (Karadaş., 2014). Speech recognition systems design systems with the capability to make easier, quicker, and more effective and more reliable processes with the human and machine interaction that supports artificial intelligence, speech-to-text and text-to-speech application areas. The most frequently used speech recognition applications in daily life are disabled aid applications, robotics applications, telephone banking applications and automatic pager applications.

The speech recognition process consists of the procedures of receiving the voice signal as an input, processing the voice and recognizing the voice signal as an output. In a general speech recognition system, a speech signal received by a microphone or telephone is first passed through a pre-treatment stage, and the distinguishing feature (property) vectors are extracted to perform the recognition of the speech expression, and a model is created by training these feature vectors extracted. Then, the model of the test speech signal, which will be questioned in terms of which speech expression it is, is extracted and then whether there is a match is determined by comparing it with the previously modelled (by training) templates. The block diagram of a speech recognition system is presented in Figure 1.

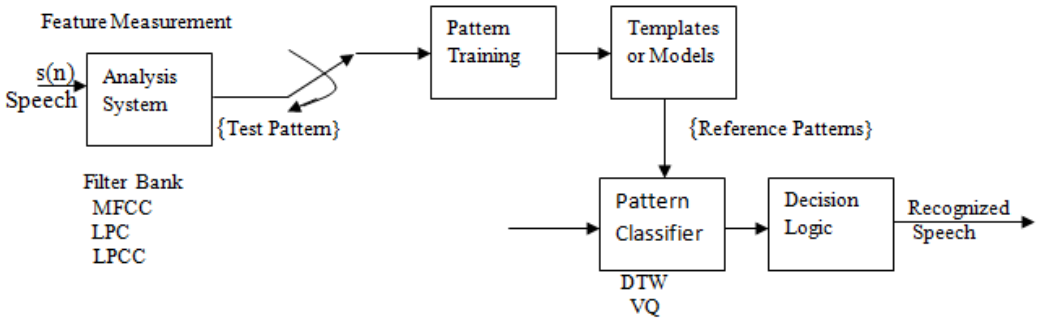

Figure 1: Block diagram of a speech recognition system (Rabiner & Juang., 1993).

The process steps that make up a typical speech recognition system defined in Figure 1 consist of 4 main blocks:
Feature Measurement (Extraction):
The most basic process to be performed in order to process speech signals of different individuals in speech recognition systems is to extract important information (distinguishing features specific to the speech expression) from the speech patterns of each speaker. However, speech patterns should be filtered from the noise effect resulting from various factors to perform this process. For this purpose, the speech patterns should be first put through a pre-treatment stage by being subjected to a filtering method before the feature extraction process.

The most frequently used Filtering and feature extraction methods are as follows:

- Wiener Filter
- Spectral Subtractive
- Kalman Filter
- Mel Frequency Cepstrum Coefficients (MFCC)
- Linear Predictive Coding (LPC)
- Linear Predictive Cepstral Coefficients (LPCC)

Pattern Training:
The feature that makes up the template corresponding to the speech expression of the same class is used in obtaining the parameters. A whole template that represents the feature parameters of each speech class is called a Reference Template.

Pattern Classification:
The classification process is the recognition of the speech signal by comparing the test speech signal with Reference Template examples. The most frequently used techniques to perform this process block are as follows:

- Dynamic Time Warping (DTW),
- Vector Quantization *(VQ)*

Decision Logic:
A similarity score is obtained between the speech signal put through the test process and the Reference Template in the final block of the speech recognition system. The highest similarity score will perform the recognition process. The materials and methods used for the development of an automatic digit recognition system were theoretically summarized in the remaining parts of this study, and then it was concluded with the conclusion part after presenting the experimental results obtained.

Materials and Methods

In this study, database sampling that consists of Kurdish speech samples was created to recognize Kurdish digits. With this database created, distinguishing features (feature vectors) of each speech signal were obtained using the MFCC method. Feature vectors were obtained by using the MFCC method together with Hamming, Hanning and Rectangular windowing techniques. Afterwards, the feature vectors obtained were modelled by training with DTW. At the test stage, whether there is any speech was questioned by comparing the given speech signal with the templates in the training set.

As a result of this study, a digit recognition system was developed in Kurdish, and its effects on the Kurdish speech recognition system were investigated by applying different windowing approaches to the MFCC feature extraction method.

Database Sampling

For an automated speech recognition system to be realized, it is primarily required to create databases that contain the voice signals suitable for the systems that are aimed to be realized. Detailed information on the database features used in this study is presented in Table 1.

Table 1: The features that make up the dataset.

	Number of speakers	Age interval	Number of words	Number of repetitions	Words told
Adult Male	76	19-55	10	3	0,1,2,3,4,5,6,7,8,9.
Adult Female	26	19-52			

As is seen in Table 1, the numbers were taken from 102 adult speakers from different age and gender groups. Speeches taken from each speaker from 0 to 9 were recorded as a single file. 3 records were taken from each speaker. Therefore, three (Record1, Record2 and Record3) speech files were collected for each speaker. In this study, each record file was used as training and test data alternately as is shown in Table 2. So, for each record, the DTW algorithm is repeatedly applied to one out of the 3 folds, while two different folds are held out each time.

Table 2: Training and Test data sampling.

Training Data	Test Data
Record1	Record2 and Record3
Record2	Record1 and Record3
Record3	Record1 and Record2

It is necessary to determine the voiced (spoken) parts in any file recorded, and remove the silent parts. In this study, the determination of the voiced and silent parts was made with signal energy and spectral centre parameters.

The voice records were obtained from silent environments or office environments with little noise. General Mobile Discovery II+ brand cell phone and the Easy Voice Recorder Pro voice recording programme loaded in this phone were used in taking the voice signals. The digital recording format is mono, 16 bit, 22050 Hz and PCM Wave file format.

Methods Used

This part is about the methods used.

I. Wiener

Wiener filter is a method performed in the frequency space.

The Wiener filtering method is expressed by the following Equation 1 equality:

$$x(m) = \sum_{k=0}^{P-1} w_k y(m-k) \qquad (1)$$

- m: means the time series,
- $y^T = [y(m), \dots, y(m-P-1)]$: input value,
- x(m) : filter output value, and
- $w^T = [w_{0,} w_{1,} \dots., w_{P-1,}]$: means the coefficient vector of Wiener filter.

Error signal is expressed with the following Equation 2 equality:

$$e(m) = x(m) - w^T y \qquad (2)$$

The Wiener filter coefficients are obtained with the value that minimizes the average squared error expressed with $e^2(m)$.

II. MFCC

MFCC is one of the most widely used feature extraction methods in speaker recognition systems. MFCC is a numerical technique analysis that imitates the perception of human ears and is calculated based on FFT (Fast Fourier Transform) (Karasartova., 2011). The flow diagram of the steps that are necessary to obtain the MFCC coefficients is presented in Figure 2.

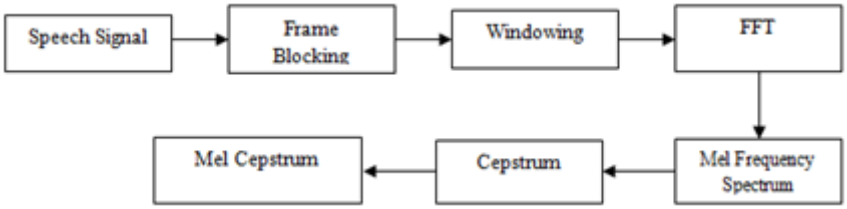

Figure 2: MFCC block diagram.

1) Frame Blocking: It means obtaining the characteristic features of the speech signal in a more stable state by processing it at small time intervals.

2) Windowing: Windowing can be defined as dividing a speech signal into a time period of a particular length. More generally, windowing is the multiplication of a framed signal with a special function.

Windowing methods,

- Hamming,
- Hanning,
- Rectangular,
- Barlett and Kaisser.

The most frequently used methods for windowing are Hamming and Rectangular windowing examples. Hamming, Hanning and Rectangular windowing methods were used in this study. The functions of windowing methods are in the following equations (N: means the number of samples).

Hamming Window:

The Hamming window is expressed by the following Equation 3equality:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1 \qquad (3)$$
$$w(n) = 0, \quad otherwise$$

Hanning Window:

The Hanning window is expressed by the following Equation 4 equality:

$$w(n) = \frac{1}{2}\left\{1 - \cos\left(\frac{2\pi n}{N-1}\right)\right\}, \quad 0 \le n \le N-1 \qquad (4)$$
$$w(n) = 0, \quad otherwise$$

Rectangular Window:

The Rectangular window is expressed by the following Equation 5 equality:

$$w(n) = 1,\ 0 \le n \le N-1 \qquad (5)$$
$$w(n) = 0, \quad otherwise$$

3) FFT (Fast Fourier Transform): Fast Fourier Transform (FFT) takes the windowed frame from the time domain to the frequency domain. The Fast Fourier Transform signal is expressed by the following Equation 6 equality:

$$X_n = \sum_{k=0}^{N-1} x_k\, e^{-2\pi jkn/N}, 0 \le n \le N-1 \qquad (6)$$

4) Mel Frequency Spectrum: In this process block, the signal with "f" frequency in Hz unit is taken from Mel filter bank that consists of triangular waves that permeate the band. Therefore, the value of the signal in M(f) frequency unit is obtained with this process. The number of filters in Mel filter bank defines MFCC coefficient value.

The algebraic equality of the process of transforming the Mel spectrum and FFT frequency values in Hz into Mel frequency unit is as follows:$M(f) = 2595.log_{10}\ \left(1 + \frac{f}{700}\right)$.

5) Mel Cepstrum: The signal obtained from Mel Filter bank is taken to the time domain from the frequency domain using DCT (Discrete Cosinus Transform) after taking its logarithm, and Mel frequency Cepstrum Coefficients values are obtained.

III. DTW (Dynamic Time Warping)

DTW is a classification method based on dynamic programming. This method is used to find the similarity between the same speech expression told by the speaker at different times and speeds. In the DTW method, the distance values that give the best match are calculated by comparing the model obtained from the feature vector of the test speech signal with the Reference Template models created from the training dataset. Afterwards, the Reference Template that is at a minimum distance to the test signal from these distance values calculated decides on the match model.

Results and Discussion

Kurdish digits from 0 to 9 were used in this study. The success rates of the MFFC method using Hamming, Hanning and Rectangular windowing techniques are compared with each other. The percentage of correct recognition of each technique based on its recognition result is shown in Table 3, and the Computational process time is shown in Table 4.

Table 3: The recognition accuracy results.

	DTW		
MFCC	**Hamming**	**Hanning**	**Rectangular**
	%99.03	%98.96	%99.26

Table 4: Comparation computational process results.

	Computation Time (Hours)		
MFCC	**Hamming**	**Hanning**	**Rectangular**
	23.363	22.790	22.234

According to the results obtained in Table 3 and Table 4, the Rectangular windowing technique yields better results than other windowing techniques in terms of recognition and computational process time in practice.

Conclusion

Within the scope of this study, it was aimed to develop a digit recognition system based on Kurdish speech patterns by modelling the feature vectors obtained with the Hamming, Hanning and Rectangular windowing approaches of the MFCC feature extraction method using the DTW method. Therefore, the analysis of different windowing techniques on the Kurdish voice system was examined. Furthermore, general information is provided on the methods used in the speech recognition field in this study. Following this study, it was aimed to reveal the speech recognition system that gives the best performance based on the voice structure of Kurdish by training MFCC and different windowing samples with the HMM (Hidden Markov Model) method.

References

Khan, E.D.B & Lescot, R. (1971). Grammaire Kürde (Dialecte Kurmandji), Paris, Avesta Press.
https://tr.wikipedia.org/wiki/K%C3%BCrt%C3%A7e.
Trivedi,V. (2013). A Survey on English Digit Speech Recognition using HMM, International Journal of Science and Research (pp. 247-253). India.
Karadaş, M. A. (2014). Bilişim teknolojisi (bt) sınıflarında konuşma tanıma teknolojisi ile sınıf otomasyonu, Ankara, MA: Gazi University, Master Thesis.
Rabiner, L. & Juang B.H. (1993). Fundemental s of speech recognition, United States of America, MA: Ptr Prentice-Hall Publishers.
Karasartova, S. (2011). Metinden bağımsız konuşmacı tanıma sistemlerinin incelenmesi ve gerçekleştirilmesi, Ankara, MA: Ankara University, Master Thesis.

