*The Second International Conference “Problems of Cybernetics and Informatics”*
*September 10-12, 2008, Baku, Azerbaijan. Section #2 “Intellectual Systems”*
www.pci2008.science.az/2/07.pdf

# THE PRINCIPLES OF CONSTRUCTION OF THE AZERBAIJAN SPEECH RECOGNITION SYSTEM

**Кamil Aida-zade[1], Samir Rustamov[2]**

Institute of Cybernetic of ANAS, Baku, Azerbaijan
[1]*kamil_aydazade@rambler.ru*, [2]*samir.rustamov@gmail.com*

## INTRODUCTION

Recently as a result of wide development of computers, the various forms information exchange between man and computer are discovered. At present inputting the data into the computer by the speech and its recognition by the computer is one of the developed scientific fields. Because each language has its specific features, the various speech recognition systems are investigated for the different languages. This is why we propose speech recognition system the Azerbaijani language.

The subject of this paper is about the construction of structured Azerbaijan speech recognition system (ASRS), analysis of investigating the speech recognition system, and recognition result. The speech inputted to our system consists of finite number of words clearly expressed with definite time interval. The recognizable words (speech) depending on applied fields can be used for various purposes.

## PROBLEM STATEMENT

Automatic speech recognition by computer is a process where speech signals are automatically converted into the corresponding sequence of words in text.

Automatic speech recognition involves a number of disciplines such as physiology, acoustics, signal processing, pattern recognition, and linguistics. The difficulty of automatic speech recognition is coming from many aspects of these areas.

*Variability from speakers:* A word may be uttered differently by the same speaker because of illness or emotion. It may be articulated differently depending on whether it is planned read speech or spontaneous conversation. The speech produced in noise is different from the speech produced in a quiet environment because of the change in speech production in an effort to communicate more effectively across a noisy environment. Since no two persons share identical vocal cords and vocal tract, they cannot produce the same acoustic signals. Typically, females sound different from males. So do children from adults. Also, there is variability due to dialect foreign accent.

*Variability from environments:* The acoustical environment where recognizers are used introduces another layer of corruption in speech signals. This is because of background noise, reverberation, microphones, and transmission channels.

## THE METHODS OF SOLUTION

The main part of speech recognition system consists of training and recognition processes. Initially basic features characterizing speech signal are computed in both processes. The efficiency of this stage is one of the significant factors affecting behavior of the next stages and exactness of speech recognition. Using the time function of the signal as feature is ineffective. The reason for this is that when the same person says the same word, its time function varies significantly.

At present the methods of calculating MFCC (Mel Frequency Cepstral Coefficients) and LPC (Linear Predictive Coding) are widely used in speech recognition as speech features.

The LPC and MFCC cepstrals combined use in speech recognition system for calculating speech features. Calculation of the speech features algorithm is defined in the following form (fig.1).

- *Pre-emphasizing.* The amplitude spectrum of a speech signal is dominant at "low frequencies" (up to approximately 4kHs). The speech signals are passed through a first-order FIR high pass filter.
- *Voice activation detection (VAD).* The problem of locating the endpoints of an utterance in a speech signal is a major problem for the speech recognizer. An inaccurate endpoint detection will decrease the performance of the speech recognizer.
- *Framing.* To increase the recognition quality, the input signal is divided into overlapping frames.
- *Windowing.* There are a number of different window functions to choose between to minimize the signal discontinuities. One of the most commonly used for windowing a speech signal before Fourier transformation, is the Hamming window.

**Calculating of MFCC features.** To calculate Mel Frequency Cepstral Coefficients, the spectrum are calculated by applying the Fourier transformation to the windowing frames. The mel frequency spectrum reduces the amount of data without loosing vital information in a speech signal. By this aim a signal is passed through the Mel filter. After logarithming of signal by applying of Inverse Fourier transformation we are getting the Mel Frequency Cepstral Coefficients. Reminder 12 cepstrals are added to the feature vector after applying cepstral mean subtraction on the next stage.

*Cepstral Mean Subtraction (CMS).* A speech signal may be subjected to some channel noise when recorded, also referred to as the channel effect. A problem arises if the channel effect when recording training data for a given person is different from the channel effect in later recordings when the person uses the system. The problem is that a false distance between the training data and newly recorded data is introduced due to the different channel effects. The channel effect is eliminated by subtracting the mel-cepstrum coefficients with the mean mel-cepstrum coefficients.

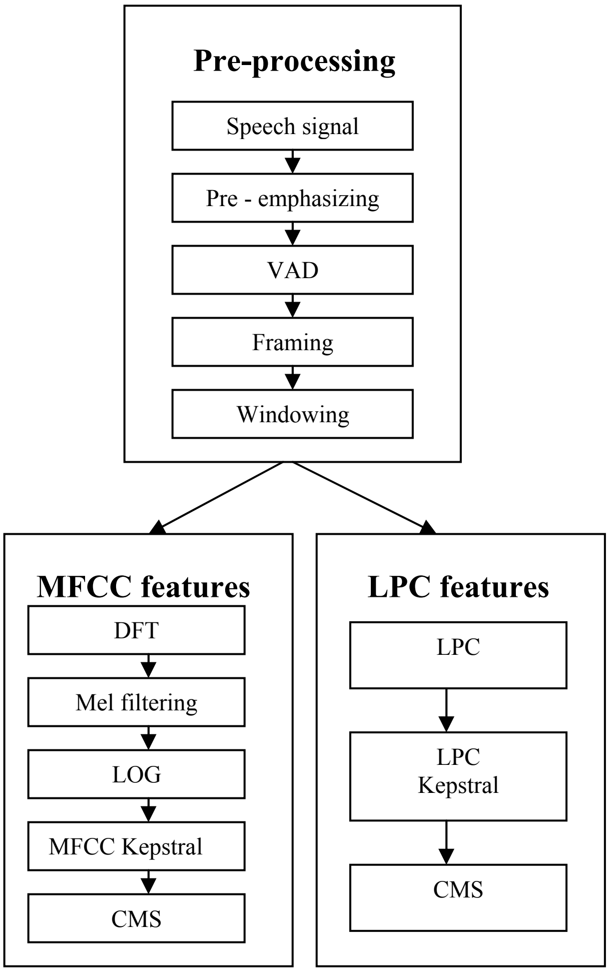

Fig.1. Scheme of speech features.

**Calculating of LPC features.** The LPC coefficients of each frame are found by applying Levinson-Durbin algorithm. Cepstrals of frames are calculated by means of found LPC coefficients. Getting 12 cepstrals are added to the feature vector after applying cepstral mean subtraction on the next stage.

## THE RECOGNITION PROCESS

One of main requirements in speech recognition system is reliability of recognition. To improve reliability of the system is offered combine using different structured or different features systems. This recognition systems can be work independently in one system and we called them conditionally subsystems of the main system (fig.2).

Speech signal is trained by different mathematical models in each subsystem separately. The recognition results of the subsystems passed to decision making block in during recognition process.

The speech recognition system depending on the aim of a user presents him a recognition system of different quality. The recognition systems with respect to the factor of error recognition percent are conditionally called strong, intermediate and weak reliability systems.

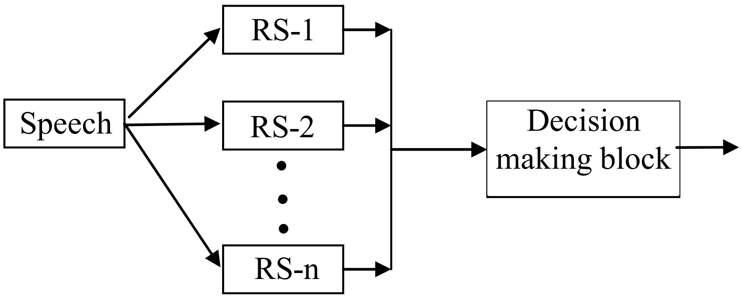

Fig. 2. Structure scheme of the ASRS.

*Strong reliability system.* This is a system confirming the recognition by each subsystem. If some of these subsystems discard the recognition, then the system doesn't accept any recognition. This system prevents the error in recognition process and therefore is more reliable.

*Intermediate reliability system.* This system uses voting between subsystems, and recognition system confirms the result of the voting. For example, if the number of subsystems are 3, then the system accepts the same result confirmed by some two subsystems of them. In spite of the fact that the confidence of the system is lower than "strong reliability system", the recognition percent is high.

*Weak reliability system.* Our suggested method in this system is a sequential method. Let's explain the main essence of the method. First subsystem is used initially, then second subsystem is applied to the unrecognized patterns by the first subsystem. Similarly the third subsystem is applied to the unrecognized patterns by the second subsystem and so on. This approach minimizes the number of unrecognized patterns. However, it has got weak reliability in terms of error rate.

As the subsystems of the ASRS are taken the neural networks trained from the different initial points. The results of the subsystems are compared in the decision making block and ASRS accepts this decision.

## EXPERIMENTS RESULTS

Using suggested principles and algorithms, the recognition system of the Azerbaijan speech with limited vocabulary has been developed. As the first experiment is taken the recognition of Azerbaijani digits and have been received high recognition. The other experiment the recognition of the names of 80 territories of Azerbaijan is realized. Note that, for training process from every speech are entered 3600 patterns, but for testing process 400 patterns to the system. There are given results of the recognition of names of Azerbaijan territories in the following tables.

Table 1. The results of the strong reliability system

| The types of features | The numbers of recognized patterns | The number of error recognized patterns | The number of unrecognized patterns |
|---|---|---|---|
| MFCC | 279 (69.75%) | 1 (0.25%) | 120 (30.0%) |
| LPC | 277 (69.25%) | 1 (0.25%) | 122 (30.5%) |
| MFCC and LPC | 268 (67.0%) | 0 (0.0%) | 132 (33.0%) |
| MFCC or LPC | 332 (83.0%) | 1 (0.25%) | 67 (16.75%) |

Table 2. The results of the intermediate reliability system

| The types of features | The numbers of recognized patterns | The number of error recognized patterns | The number of unrecognized patterns |
|---|---|---|---|
| MFCC | 353 (88.25%) | 2 (0.5%) | 45 (11.25%) |
| LPC | 350 (87.5%) | 2 (0.5%) | 48 (12.0%) |
| MFCC and LPC | 349 (87.25%) | 0 (0.0%) | 51 (12.75%) |
| MFCC or LPC | 370 (92.5%) | 4 (1.0%) | 26 (6.5%) |

Table 3. The results of the weak reliability system

| The types of features | The numbers of recognized patterns | The number of error recognized patterns | The number of unrecognized patterns |
|---|---|---|---|
| MFCC | 373 (93.25%) | 16 (4.0%) | 11 (2.75%) |
| LPC | 375 (93.75%) | 14 (3.5%) | 11 (2.75%) |
| MFCC and LPC | 368 (92.0%) | 9 (2.25%) | 23 (5.75%) |
| MFCC or LPC | 368 (92.0%) | 22 (5.5%) | 10 (2.5%) |

## References

1. K.R.Ayda-zade, S.S.Rustamov. Research of Cepstral Coefficients for Azerbaijan speech recognition system. (Azerbaijani). Transactions of Azerbaijan National Academy of sciences. "Informatics and control problems". Volume XXV, №3. Baku, 2005, pp.89-94.
2. K.R.Ayda-zade, S.S.Rustamov. On Azerbaijan Speech Recognition System. (Azerbaijani). Application of information-communication technologies in science and education. International conference. Baku, 2007. Volume II, pp. 670-677.
3. S.S.Rustamov. On using an ambiguity of training neural networks in systems of speech recognition. (Azerbaijani). Transactions of Azerbaijan National Academy of sciences. "Informatics and control problems". Volume XXVI, №2. Baku, 2006, p.256-260.
4. S.S.Rustamov. The establishment principles of speech recognition system. Problems of Cybernetics and Informatics. International conferance. Volume III. Baku, 2006, pp. 92-95.
5. Mikael Nilsson,Marcus Ejnarsson. "Speech Recognition using Hidden Markov Model". Department of Telecommunications and Speech Processing, Blekinge Institute of Technology. 2002. www.hh.se/staff/maej/publications/MSc Thesis - MiMa.pdf.