Investigation of Combined Use of MFCC and LPC Features in Speech Recognition Systems

K.R. Aida–Zade, C. Ardil, S.S. Rustamov

***Abstract*—**Statement of the automatic speech recognition problem, the assignment of speech recognition and the application fields are shown in the paper. At the same time as Azerbaijan speech, the establishment principles of speech recognition system and the problems arising in the system are investigated.

The computing algorithms of speech features, being the main part of speech recognition system, are analyzed. From this point of view, the determination algorithms of Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) coefficients expressing the basic speech features are developed. Combined use of cepstrals of MFCC and LPC in speech recognition system is suggested to improve the reliability of speech recognition system. To this end, the recognition system is divided into MFCC and LPC-based recognition subsystems. The training and recognition processes are realized in both subsystems separately, and recognition system gets the decision being the same results of each subsystems. This results in decrease of error rate during recognition.

The training and recognition processes are realized by artificial neural networks in the automatic speech recognition system. The neural networks are trained by the conjugate gradient method. In the paper the problems observed by the number of speech features at training the neural networks of MFCC and LPC-based speech recognition subsystems are investigated.

The variety of results of neural networks trained from different initial points in training process is analyzed. Methodology of combined use of neural networks trained from different initial points in speech recognition system is suggested to improve the reliability of recognition system and increase the recognition quality, and obtained practical results are shown.

***Keywords*—**speech recognition, cepstral analysis, Voice activation detection algorithm, Mel Frequency Cepstral Coefficients, features of speech, Cepstral Mean Subtraction, neural networks, Linear Predictive Coding

I. INTRODUCTION

RECENTLY as a result of wide development of computers, the various forms of 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 for the Azerbaijani language.

The subject of this paper is about the construction of structured Azerbaijan speech recognition system, 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.

II. 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 signal. Typically, females sound is 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 to introduce another layer of corruption in speech signals. This is because of background noise, reverberation, microphones, and transmission channels.

III. THE METHODS OF SOLUTION

At first the speech signal is transformed into electric oscillation by the sound recorders (for example, microphone). Later the signal passed over analog-digital converter is transformed into digital form at some sampling frequency f_d and quantization level. The sampling frequency - analog signal without losing its important information determines the necessary frequency for sampling.

The main part of speech recognition system consists of training and recognition processes. Initially basic features

Kamil Aida-Zade is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.
Cemal Ardil is with the National Academy of Aviation, Baku, Azerbaijan.
Samir Rustamov is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.

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.

Let's explain the essence of these methods, separately:

The model of speech generation consists of two parts: the generation of the excitation signal and the vocal tract filter. The excitation signal is spectrally shaped by a vocal tract equivalent filter. The outcome of this process is the speech. If $e(n)$ denotes a sequence of the excitation signal and $\theta(n)$ denotes the impulse response of the vocal tract equivalent filter, a sequence of the speech is then equal to the excitation signal convolved with the impulse response of the vocal tract filter as shown in equation (3.1).

$$s(n) = e(n) * \theta(n) \tag{1}$$

A convolution in the time domain corresponds to a multiplication in the frequency domain:

$$S(\omega) = E(\omega) \cdot \theta(\omega) \tag{2}$$

In MFCC method using the logarithm of equation (2), the multiplied spectra becomes additive

$$\log|S(\omega)| = \log|E(\omega)\cdot\theta(\omega)| = \log|E(\omega)| + \log|\theta(\omega)|.$$

It is possible to separate the excitation spectrum $E(\omega)$ from the vocal system spectrum $\theta(\omega)$ by remembering that: $E(\omega)$ is responsible for the "fast" spectral variations, $\theta(\omega)$ is responsible for the "slow" spectral variations. Frequency components corresponding to $E(\omega)$ appear at "large values" on the horizontal axis in the "new frequency domain", whereas frequency components corresponding to $\theta(\omega)$ appear at "small values". The new domain found after taking the logarithm and the inverse Fourier transform is called the cepstrum domain, and the word quefrency is used for describing the "frequencies" in the cepstrum domain.

As same way Z transform is applied to the convolution in the time domain in method LPC:

$$S(z) = E(z) \cdot \theta(z)$$

The main idea behind linear prediction is to extract the vocal tract parameters. Given a speech samples at time n, $s(n)$ can be modeled as a linear combination of the past p speech samples, such that:

$$\hat{s}(a;n) = \sum_{k=1}^{p} s(n-k)\cdot a_p(k)$$

where $a_p = (a_p(1), a_p(2), ..., a_p(p))$ are unknown LPC coefficients $(p \in [8,12])$.

Summing the real and predicted samples we get the following signal:

$$e(a;n) = s(n) + \hat{s}(a;n) = s(n) + \sum_{k=1}^{p} a_p(k)\cdot s(n-k) \tag{3}$$

Apply to this signal the $z-$transform: $R(z) = S(z)A(z)$.

The filter $A(z) = 1 + \sum_{k=1}^{p} a_p(k) z^{-k}$ is called the predicting error filter. This filter is equal to the inverse value of vocal tract equivalent filter.

$$A(z) = \frac{1}{\theta(z)}.$$

To find the vocal tract filter $\theta(z)$, we must first find the LPC coefficients a_p. By this aim, the following function is minimized

$$\varepsilon_p(a) = \sum_{n=1}^{M} |e(a;n)|^2 \to \min \tag{4}$$

where M is a number of frames.

We use the necessary condition of minimum to solve the problem:

$$\frac{\partial \varepsilon_p(a)}{\partial a_p(k)} = \frac{\partial}{\partial a_p(k)} \sum_{n=1}^{M} |e(a;n)|^2 = 2\sum_{n=1}^{M} e(a;n) \frac{\partial}{\partial a_p(k)} e(a;n) =$$

$$= 2\sum_{n=1}^{M} e(a;n) \frac{\partial}{\partial a_p(k)} \left[s(n) + \sum_{l=1}^{p} a_p(l) s(n-l) \right] =$$

$$= 2\sum_{n=1}^{M} e(a;n) s(n-k) = 0 , \qquad k = 1,2,..., p$$

Then we get

$$\sum_{n=1}^{M} \left[s(n) + \sum_{l=1}^{p} a_p(l) s(n-l) \right] s(n-k) = 0 . \tag{5}$$

Let's denote $r_x(k) = \sum_{n=1}^{M} s(n)s(n-k)$. Consequently we can write the equation (3.5) as following form.

$$r_x(k) + \sum_{l=1}^{p} a_p(l) r_x(l-k) = 0 \text{ or } \sum_{l=1}^{p} a_p(l) r_x(k-l) = -r_x(k), \quad k = 1,..., p. \tag{6}$$

The equation (6) is called the normal equation or the Yule-Walker equation.

Using the expression (3) in the functional (4), we get:

$$\varepsilon_p(a) = \sum_{n=1}^{M} |e(a;n)|^2 = \sum_{n=1}^{M} e(a;n)e(a;n) =$$

$$= \sum_{n=1}^{M} e(a;n) \left[s(n) + \sum_{k=1}^{p} a_p(k) s(n-k) \right] =$$

$$= \sum_{n=1}^{M} e(a;n)s(n) + \sum_{k=1}^{p} a_p(k) \sum_{n=1}^{M} e(a;n)s(n-k).$$

While $\sum_{n=1}^{M} e(a;n)s(n-k) = 0$, we can write the functional (4) as following form.

$$\varepsilon_{p,\min}(a) = \varepsilon_p(a) = \sum_{n=1}^{M} e(a;n)s(n) = \sum_{n=1}^{M} \left[s(n) + \sum_{k=1}^{p} a_p(k) s(n-k) \right] s(n) =$$

$$= r_x(0) + \sum_{k=1}^{p} a_p(k) r_x(k).$$

The coefficients $a_p(k)$, which giving the minimum to the functional is found by using following Levinson-Durbin recursion.

1. a) $a_0(0) = 1$ b) $E_0 = r_x(0)$
2. For $j = 0,1,...p-1$ calculated the following

expressions:

a) $\gamma_j = r_x(j+1) + \sum_{i=1}^{j} a_j(i) r_x(j-i+1)$

b) $\Gamma_{j+1} = -\gamma_j / E_j$

c) $i = 1,2,..., j$

$a_{j+1}(i) = a_j(i) + \Gamma_{j+1} a_j(j-i+1)$

d) $a_{j+1}(j+1) = \Gamma_{j+1}$

e) $E_{j+1} = E_j \left[1 - \left| \Gamma_{j+1} \right|^2 \right].$

IV. ALGORITHM OF CALCULATION OF SPEECH FEATURES

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

1. *Pre-processing.* The amplitude spectrum of a speech signal is dominant at "low frequencies" (up to approximately $4\,kHz$). The speech signals is passed through a first-order FIR high pass filter:

$$s_p(n) = s_{in}(n) - \alpha \cdot s_{in}(n-1)$$

where $\alpha -$ is the filter coefficient $(\alpha \in (0{,}95;1))$, $s_{in}(n) -$ is the input signal.

2. *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. Some commonly used measurements for finding speech are short-term energy estimate E_s, or short-term power estimate P_s, and short term zero crossing rate Z_s. For the speech signals $s_p(n)$ these measures are calculated as follows:

$$E_s(m) = \sum_{n=m-L+1}^{m} s_p^2(n), \qquad P_s(m) = \frac{1}{L} \sum_{n=m-L+1}^{m} s_p^2(n),$$

$$Z_s(m) = \frac{1}{L} \sum_{n=m-L+1}^{m} \frac{\left| \operatorname{sgn}(s_p(n)) - \operatorname{sgn}(s_p(n-1)) \right|}{2}$$

where

$$\operatorname{sgn}(s_p(n)) = \begin{cases} 1, & s_p(n) \geq 0, \\ -1, & s_p(n) < 0. \end{cases}$$

For each block of $L = 100$ samples these measures calculate some value. The short term zero crossing rate gives a measure of how many times the signal, $s_p(n)$, changes sign. This short term zero crossing rate tends to be larger during unvoiced regions.

These measures will need some triggers for making decision about where the utterances begin and end. To create a trigger, one needs some information about the background noise. This is done by assuming that the first 5 blocks are background noise. With this assumption, the mean and variance for the measures will be calculated. To make a more comfortable approach, the following function is used:

$$W_s(m) = P_s(m) \cdot (1 - Z_s(m)) \cdot S_c.$$

Using this function both the short-term power and the zero crossing rate will be taken into account. S_c is a scale factor for avoiding small values, in a typical application is $S_c = 1000$. The trigger for this function can be described as:

$$t_W = \mu_W + \alpha \delta_W$$

the μ_w is the mean and δ_w is the variance for $W_s(m)$ calculated for the first 5 blocks. The α term is constant that have to be fine tuned according to the characteristics of the signal. After some testing the following approximation of α will give a pretty good voice activation detection in various level of additive background noise.

$$\alpha = 0{,}2 \cdot \delta_W^{-0{,}4}.$$

The voice activation detection function, $VAD(m)$, can be found as:

$$VAD(m) = \begin{cases} 1, & W_s(m) \geq t_W, \\ 0, & W_s(m) < t_W. \end{cases}$$

By using this function we can detect the endpoints of an utterance.

3. *Framing.* The input signal is divided into overlapping frames of N samples.

$$s_{frame}(n) = s_p(n) \cdot w(n),$$

$$w(n) = \begin{cases} 1, & K \cdot r < n \leq K \cdot r + N, \quad r = 0,1,2,..., M-1, \\ 0, & otherwise, \end{cases}$$

where M is the number of frames, f_s is the sampling frequency, t_{frame} is the frame length measured in time, and K is the frame step.

$$N = f_s \cdot t_{frame}.$$

TABLE I
VALUES OF FRAME LENGTH AND FRAME STEP INTERVAL DEPENDING ON THE SAMPLING FREQUENCY

Sampling frequency (f_s)	$f_s = 16kHs$	$f_s = 11kHs$	$f_s = 8kHs$
Frame length (N)	400	256	200
Frame step (K)	160	110	80

We use the $f_s = 16kHs$ sampling frequency in our system.

4. *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:

$$s_w(n) = \left\{ 0{,}54 - 0{,}46 \cos\left(\frac{2\pi(n-1)}{N-1} \right) \right\} s_{frame}(n), \quad 1 \leq n \leq N.$$

Calculating of MFCC features.

Fast Fourier transform(FFT). Applying by FFT to windowing frames are calculated spectrum of frames.

$$bin_k = \left| \sum_{n=1}^{N} s_w(n) e^{-i(n-1)k\frac{2\pi}{N}} \right|, \quad k = 0,1,2,..., N-1.$$

Mel filtering. The low-frequency components of the magnitude spectrum are ignored. The useful frequency band lies between $64\,Hz$ and half of the actual sampling frequency. This band is divided into 23 channels equidistant in mel frequency domain. Each channel has triangular-shaped frequency window. Consecutive channels are half-overlapping.

The choice of the starting frequency of the filter bank, $f_{start} = 64Hz$, roughly corresponds to the case where the full frequency band is divided into 24 channels and the first channel is discarded using any of the three possible sampling frequencies.

The centre frequencies of the channels in terms of FFT bin indices ($cbin_i$ for the i-th channel) are calculated as follows:

$$Mel(x) = 2595\lg\left(1+\frac{x}{700}\right), \qquad x = 700\cdot\left(10^{\frac{mel}{2595}}-1\right),$$

$$f_{c_i} = Mel^{-1}\left\{Mel\{f_{start}\} + \frac{Mel\{f_s/2\} - Mel\{f_{start}\}}{NF}i\right\}, \quad , \quad i = 1,2,3,\ldots,NF-1$$

$$cbin_i = round\left\{\frac{f_{c_i}}{f_s}N\right\},$$

where $round(\cdot)$ stands for rounding towards the nearest integer. $NF = 24$ -is the number of channels of filter.

The output of the mel filter is the weighted sum of the FFT magnitude spectrum values (bin_i) in each band. Triangular, half-overlapped windowing is used as follows:

$$fbank_k = \sum_{i=cbin_{k-1}}^{cbin_k} \frac{i - cbin_{k-1}+1}{cbin_k - cbin_{k-1}+1} bin_i + $$
$$+ \sum_{i=cbin_k+1}^{cbin_{k+1}} \left(1 - \frac{i - cbin_k}{cbin_{k+1} - cbin_k + 1}\right) bin_i,\ k = 1,2,\ldots,NF-1.$$

where $cbin_0$ and $cbin_{24}$ denote the FFT bin indices corresponding to the starting frequency and half of the sampling frequency, respectively,

$$cbin_0 = round\left\{\frac{f_{start}}{f_s}N\right\};$$

$$cbin_{24} = round\left\{\frac{f_s/2}{f_s}N\right\} = \frac{N}{2}.$$

Non-linear transformation. The output of mel filtering is subjected to a logarithm function (natural logarithm)

$$f_i = \ln(fbank_i), \quad i = 1,2,\ldots,NF-1.$$

Cepstral coefficients. 12 cepstral coefficients are calculated from the output of the non-linear transformation block.

$$C_i = \sum_{j=1}^{NF-1} f_j \cdot \cos\left(\frac{\pi\cdot i}{NF-1}(j-0.5)\right), \quad i = 1,..,12.$$

We apply to these 12 LPC cepstrals the cepstral mean subtraction and enter to the feature vector in next step.

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:

$$mc_j(q) = C_j(q) - \frac{1}{M}\sum_{i=1}^{M} C_i(q), \quad q = 1,2,,\ldots,12$$

Calculating of LPC features.

The LPC coefficients of each frame are found by applying Levinson-Durbin algorithm and following cepstrals are calculated.

$$c(k) = -a_p(k) - \sum_{i=1}^{k-1}\left(1-\frac{i}{k}\right)a_p(i)c(k-i), \qquad k = 1,\ldots 12\cdot$$

We apply the cepstral mean subtraction to these 12 LPC cepstrals and enter to the feature vector in next step.

V. Construction of neural network

There are various mathematical models which form the basis of speech recognition systems. The widely used model is Multilayer Artificial Neural Network (MANN). Let's briefly describe the structure of MANN.

Generally, MANN is incompletely connected graph. Let $L-$ quantity of MANN's layers, N_ℓ- neuron quantity on layer l, $l = 1..L$; I_{lj}^- - set of neurons of layer $(l-1)$, which connected to the neuron j on layer l; θ_j^l - bias of neuron j on layer l; w_{ij}^ℓ - weighted coefficient (synapse) of connection between of neuron i on layer $(l-1)$ and neuron j on layer l; $s_{j,p}^\ell$ and $y_{j,p}^\ell$ - state and output value of neuron j on layer l for input signal $x_p \in X$ of MANN.

Forward propagation of MANN for $x_p \in X$ input signal has been described by the following expressions (figures 1,2):

$$s_{j,p}^l = \sum_{i\in I_{lj}^-} w_{ij}^l \cdot y_{i,p}^{l-1} + \theta_j^l, \tag{7}$$

$$y_{j,p}^l = f(s_{j,p}^l), \qquad j = 1,\ldots,N_l,\ l = 1,\ldots,L, \tag{8}$$

$$y_{j,p}^0 = x_{j,p}, \qquad j = 1,\ldots,N_0, \tag{9}$$

where $f(\cdot)$ - given nonlinear activation function. As activation function logistic or hyperbolic tangent functions can be used:

$$f_{\log}(z) = \frac{1}{1+e^{-\alpha z}}, \quad f_{\tan}(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}.$$

Their derivation can be calculated by function value:

$$\frac{df_{\log}(z)}{dz} = \alpha\cdot f_{\log}(z)\cdot(1-f_{\log}(z)), \quad \frac{df_{\tan}(z)}{dz} = 1 - f_{\tan}^2(z).$$

Let, the training set of $\{x_p, d_p\}, p = 1..P$ pairs are given, where $d_p = (d_{1,p},\ldots,d_{N_L,p})-$ desired output for x_p input signal. The training of MANN consists in finding such w_{ij}^ℓ and θ_j^ℓ $i \in I_{lj}^-$, $j = 1,\ldots,N_l, l = 1,\ldots,L$, herewith on x_p input signal that MANN has output y_p, which maximal closed to desired output d_p. Usually, training quality is defined by mean square error function:

$$E(w,\theta;x,s,y)=\frac{1}{P}\sum_{p=1}^{P}\eta_p E_p(w,\theta;x_p,s_p,y_p),$$
$$E_p(w,\theta;x_p,s_p,y_p)=\frac{1}{2}\sum_{j=1}^{N_L}\left(y_{j,p}^{L}-d_{j,p}\right)^2, \tag{10}$$

where η_p– coefficient, which determine the belonging "quality" of input x_p to its "ideal" pattern $p=1,...,P$, $j=1,...,N_L$.

The task of MANN training constitutes minimization of criterion (10) according to parameters (w,θ) with (7)-(9) conditions. The MANN of developed system was trained by conjugate gradient method.

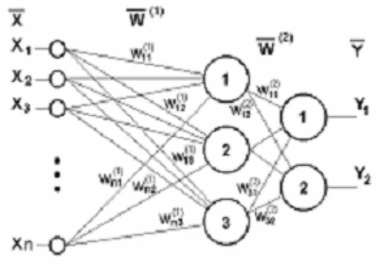

Fig. 1 MANN with two layers

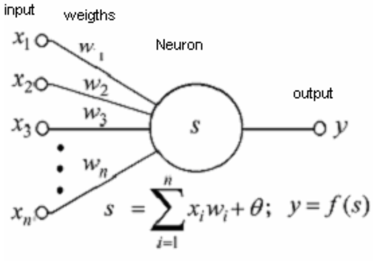

Fig. 2 One neuron description

VI. THE RECOGNITION PROCESS

The speech recognition system consists of MFCC and LPC-based two subsystems. These subsystems are trained by neural networks with MFCC and LPC features, respectively.

The recognition process is realized by two stages:

1. In MFCC and LPC–based recognition subsystems recognition processes are realized in parallel.
2. The recognition results of MFCC and LPC–based recognition subsystems are compared and the speech recognition system confirms the result, which confirmed by the both subsystems.

Since the MFCC and LPC methods are applied to the overlapping frames of speech signal, the dimension of feature vector depends on dimension of frames. At the same time, the number of frames depends on the length of speech signal, sampling frequency, frame step, frame length. In our system the sampling frequency is $16\,khs$, the frame step is 160 samples, and the frame length is 400 samples.

The other problem of speech recognition is the same speech has different time duration. Even when the same person repeats the same speech, it has the different time durations. For partially removing the problem, time durations are led to the same scale. When the dimension of scale defined for the speech signal increases, then the dimension of feature vector corresponding to the signal also increases.

The dimension of neural network is taken as a total number of weights and biases of neural network. The large dimension of the feature vector acts strongly on the dimension of neural network. For example, our neural network consists of 2 layers: a number of inputted parameters are 420, a number of neurons in the first layer are 50, a number of neurons in output layer are 10. The dimension of our neural network is $420\times 50+50\times 10+60=21560$.

Since the dimension of neural network is less than the number of trained samples**,** there exists a set of various weights and biases giving a minimum to minimization criterion (10), such that the application of these weights and biases to the recognition system gives the different results.

Here the construction of the following system is suggested by application of neural networks trained from different initial points. 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.

Strong reliability system. This is a system confirming the recognition by each neural network trained from different initial points. If some of these networks 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 neural networks trained from different initial points, and recognition system confirms the result of the voting. For example, if the number of neural networks trained by changing the initial points are 3, then the system accepts the same result confirmed by some two networks 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. The recognition system is trained from different initial points in the trained process. First trained neural network is used initially, then second neural network is applied to the by the first network. Similarly the third neural network is applied to the unrecognized patterns by the second neural network and so on. This approach minimizes the number of unrecognized patterns. However, it has got weak reliability in terms of error rate.

Note that the number of neural networks trained from the different initial points depends on the computer processing power. Apparently when the number of neural networks used in the system increases, the strong reliability system minimizes the error and the recognition reliability of the system increases. In weak reliability system despite of increasing the correct recognition percent, the error recognition percent also relatively increases and the reliability of the system decreases. The number of neural networks doesn't affect the results of intermediate reliability system.

VII. EXPERIMENTAL RESULTS

For training and recognition of the different scaled speech in neural network, it is necessary to lead them to the same scale. The dimension of scale has taken 5840 samples, which corresponds to the 35 frames. Every frame has 12 features.

Our neural network consists of 2 layers: a number of input parameters are 420, a number of neurons in the first layer is 50, a number of neurons in output layer is the number of testing words (10). Testing speech is taken by Azerbaijani digits.

For training process from every speech form digit are entered 140-150 patterns to the system. The neural networks of developed system were trained by conjugate gradient method. In following tables MFCC and LPC-based subsystems results are shown separately. Results of speech recognition system, which combined use the MFCC and LPC-based subsystems are also shown.

TABLE II
THE RESULTS OF THE MFCC AND LPC-BASED SPEECH RECOGNITION SUBSYSTEMS TRAINING FROM DIFFERENT INITIAL POINTS

Number of training	The numbers of testing patterns	The types of features	The numbers of recognized patterns	The number of error recognized patterns	The number of unrecognized patterns
1	312	MFCC	291 (93.27%)	7 (2.24%)	14 (4.49%)
		LPC	288 (92.31%)	5 (1.6%)	19 (6.09%)
2	312	MFCC	288 (92.31%)	9 (2.88%)	15 (4.81%)
		LPC	292 (93.59%)	7 (2.24%)	13 (4.17%)
3	312	MFCC	289 (92.63%)	5 (1.6%)	18 (5.77%)
		LPC	293 (93.91%)	6 (1.92%)	13 (4.17%)

TABLE III
THE RESULTS OF THE STRONG RELIABILITY SYSTEM

The numbers of testing patterns	The types of features	The numbers of recognized patterns	The number of error recognized patterns	The number of unrecognized patterns
312	MFCC	273 (87.5%)	1 (0.32%)	38 (12.18%)
	LPC	286 (91.61%)	2 (0.64%)	24 (7.69%)
	Combined	264 (84.6%)	1(0.32%)	47(15.1%)

TABLE IV
THE RESULTS OF THE INTERMEDIATE RELIABILITY SYSTEM

The numbers of testing patterns	The types of features	The numbers of recognized patterns	The number of error recognized patterns	The number of unrecognized patterns
312	MFCC	294 (94.23%)	3 (0.96%)	15 (4.8%)
	LPC	290 (92.95%)	5 (1.6%)	17 (5.45%)
	Combined	286(91.67%)	2(0.64%)	24(7.69%)

TABLE V
THE RESULTS OF THE WEAK RELIABILITY SYSTEM

The numbers of testing patterns	The types of features	The numbers of recognized patterns	The number of error recognized patterns	The number of unrecognized patterns
312	MFCC	299 (95.83%)	10 (3.2%)	3 (0.96%)
	LPC	296 (94.87%)	10 (3.2%)	6 (1.92%)
	Combined	294(94.23%)	4(1.28%)	14(4.49%)

REFERENCES

[1] K.R.Ayda-zade, S.S.Rustamov. Research of Cepstral Coefficients for Azerbaijan speech recognition system. Transactions of Azerbaijan National Academy of sciences."Informatics and control problems". Volume XXV, №3. Baku, 2005, p.89-94.

[2] К.Р.Айда-заде, Э.Э.Мустафаев. Об оптимизации параметров нейронной сети на этапе ее обучения / Труды Республиканской научной конференции «Современные проблемы информатизации, кибернетики и информационных технологий», том I, Баку, 2003, с. 118-121.

[3] Mikael Nilsson,Marcus Ejnarsson. "Speech Recognition using Hidden Markov Model".Department of Telecommunications and Speech Processing, Blekinge Institute of Technology. 2002. http://www.hh.se/staff/maej/publications/MSc Thesis - MiMa.pdf

[4] Group 622 "On Speaker Verification". 2004. 198 p. http://www.control.auc.dk/~jhve02/report_inf6.pdf

[5] А.Б.Сергиенко. Цифровая обработка сигналов. СПб.: Питер, 2002, 608 с.

[6] ETSI ES 201 108 v1.1.2 (2000-04). "Speech Processing, Transmission and Quality aspects(STQ); distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms". 20 p. http://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_13/docs/PDF/SP-010566.pdf

[7] Bengt Mandersson. Chapter 4. "Signal Modeling".Department of Electroscience. Lund University. August 2005. http://www.tde.lth.se/ugradcourses/osb/osb05_f2_a4.pdf

[8] Bengt Mandersson. Chapter 5. "Levinson-Durbin Recursion". Department of Electroscience. Lund University. September 2005. http://www.tde.lth.se/ugradcourses/osb/osb05_f3_a4.pdf

[9] Group 11. Tejaswini Hebalkar, Lee Hotraphinyo, Richard Tseng. "Voice Recognition and Identification System". Digital communications and Signal Processing Systems Design. June 2000. http://www.ece.cmu.edu/~ee551/Final_Reports/Gr11.551.S00.pdf

[10] Bengt Mandersson. Chapter 4. "Signal Modeling".Department of Electroscience. Lund University. August 2005. http://www.tde.lth.se/ugradcourses/osb/osb05_f2_a4.pdf

[11] Химмельблау Д. Прикладное нелинейное программирование. М.: Мир, 1975, 534 с.

Kamil Aida-Zade is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.
Cemal Ardil is with the National Academy of Aviation, Baku, Azerbaijan.
Samir Rustamov is with the institute of Cybernetics of the National Academy of Sciences, Baku, Azerbaijan.