RESEARCH ARTICLE OPEN ACCESS

An Improved Real-Time Speech Signal In Case Of Isolated Word Recognition

Takialddin Al Smadi
Department of Communication Engineering, Jerash University, Jordan

ABSTRACT
In modern computer systems, more attention is given to building natural interface information; directions of speech are the dialogue, which contains automatic speech recognition and synthesis. The system included many different applications, for example voice control, voice access to information resources, language training programming, incapable, accessing through voice verification/identification. Study a starting point on the speech recognition trouble used hidden Markov model, wavelets and neural network could be involved for making more precise prediction **are** implemented in as a library of Computer Simulation for use with digital signal processors tms320vc-5505 of Texas Instruments Brand. On this library there was created a test speaker dependent system of specific terms recognition with the small size of the dictionary.
Keywords - voice control, speech recognition, digital signal processing, real-time, providing API

I. Introduction

Speech recognition using control systems that use automatic speech recognition commands depends on hidden Markov models (HMM) on a computer. When fixed to date hardware-based systems of recognition and taking into consideration the tendencies of its development in the near future, is considered as the most important block of such systems- training unit training sequences. The successful solution of the tasks of learning Markov model depends directly on the excellence of the recognition system. The task of teaching HMM at the moment there are two serious problems: standard methods of its decision (Baum-Velča) method are local optimization methods that is not able to go beyond the local extreme of the function) and is heavily dependent on the starting parameters. In the search for solving this trouble is the improvement of software for speech-recognition systems. To achieve this goal in the work covers the following main tasks:

• Researched learning algorithms HMM training sequences.

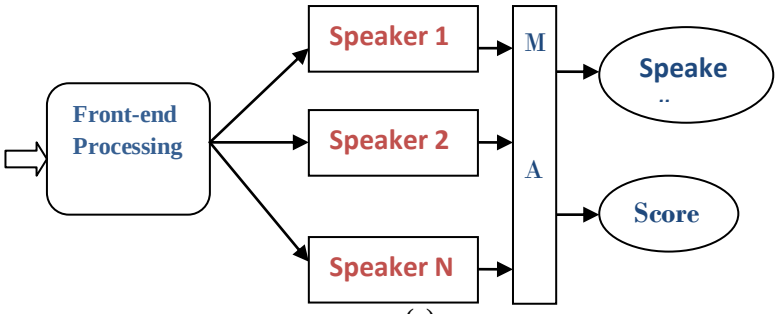
• Methods of improving the efficiency and performance of the algorithm in the situation of the problem.

Currently, work on speech recognition not only not lost relevance, but also develop a broad front, finding a large range of areas for practical application. Now, you can select 4 relatively isolated areas in development the speech technology [1].

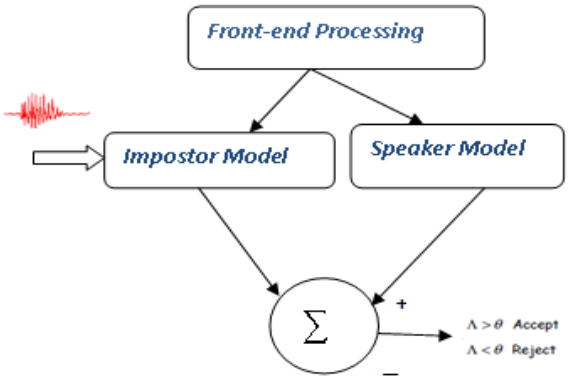
1. Speech recognition, the conversion of acoustic signal in the chain of narration, characters, words. These systems are able to be expressed on many parameters. First of all, it is the volume of the dictionary: small volumes up to 20 words, big-thousands or tens of thousands. Number of Speakers: one to arbitrary. Cast style from isolated teams to continuous speech and from reading to spontaneous speech. The branching factor means the value that identifies the number of hypotheses on each step of recognition: from small units (10 ÷ 15 <) to large (> 100 ÷ 200). Signal/noise
   Ratio of large (> 30 DB) to low (10 DB <). Quality links from high-quality microphone to the telephone Channel. The quality of speech recognition systems is generally characterized by reliability, word recognition, or what is the same, the percentage of errors.
2. Identification of the speaker's identity. These systems are divided into two classes: the verification of the speaker's identity and the speaker's identification of his personality from a limited number of people. Both of these classes can be divided into the following characteristic option-the volume of the passphrase. The other two as in speech: signal/noise ratio and the excellence of the link. Quality systems verification/identification of the speaker is characterized by two quantities: the probability of not "his" speaker identification and probability of a "foreign" announcer for his.
3. Speech synthesis. Essentially there are two classes:
   a) Reproduction of the recorded in one form or another, a limited number of messages;
   b) Speech synthesis on the text. Synthesizers are characterized by the following parameters: intelligibility word or syllable, natural sounding noise immunity.
4. Compression of speech. The main (and only) classification systems, it is an indication of the quantity of compression: low (32-16 kbit/s) to

high (1200-2400 KB/sec and below). The quality of speech characterized by compression systems, first of all additional features are very important in many applications are recognition of the speaker's voice and the ability to determine the stress level of the speaker.

These modes share the same functional part as shown in figure 1. If the system is in a learning mode, selections phase characteristics obtained values are saved in the template library. Once the system is able to recognize the values obtained are compared with sets from the library. The best result of the comparison is returned as the result of recognition [2].



(a)



(b)

Fig 1 (a) speaker identification & (b) speaker verification systems

Isolation of the word from the continuous flow of incoming information is difficult task due to specific characteristics of the voice, environment and equipment involved into speech signal recording process. A man can successfully recognize speech, the volume of which varies vary widely. The human brain is able to filter out the soft speech from ambient noise, such as music or noise produced by devices in the operation. Unlike the human brain, digital equipment is very sensitive to such external influences. The microphone is on the table, so when turning the head or changing the location of the body, the expanse between the microphone and the mouth will change [3]. This will alter the level of the output microphone signal and the signal/noise ratio, which damages the reliability of speech recognition. Changing in the intensity of speech pronunciation, softening of the beginning and ending sounds in words all these in practice lead to complicating to distinguish the endpoints from interference signal is always present in the signal [4].

II. Word recognition in continuous speech

For word recognition in continuous speech tested two different approaches. In the first case, the global approach must recognize a word that is compared with each term or word in the dictionary. The comparison is used, as an imperative, the spectral representation of each word. Among the different methods gave good results for this type of method of dynamic programming. In the second case, the analytical approach of each word or group of words first is segmented into smaller units. Segments are similar units, the same time retain parameters duration, energy, related to prosody and useful in the future. Segmentation can be based on transparent statements that are often located near maximum integrative energy spectrum. In this approach, the first criterion of segmentation is the energy change in time. Some consonants have the same energy as vowels. Therefore, supply additional parameters to determine the vowel sound in each previously defined segment.
For the identification of consonants is usually carried out division of explosive and non-explosive consonants. This is getting by detecting the convergence between the pauses before implementation. After the discovery of the bows are determined by changing the range and type of change. For each category of sounds typically use streamlined rules based on information independent of the phonetic and acoustic contexts. In continuous speech, phonetic realization of any particular statements depends on several factors, including dialect, speed of a speech has been delivered, the manner of pronouncing the speaker and others.

2.1 Application of neural networks for speech recognition.

Such processors typically are very simple, especially to the processors used in personal computers. Each processor is similar to the network only deals with signals, which he periodically receives, and the signals which he periodically sends the other processors, and yet, being united in a huge network that has a managed interoperability, such locally simple processors are capable of quite complex tasks. The concept initiated in studding the processes occurring in the brain when thinking, and if you try to simulate these processes. The models are called Artificial Neural Networks (ANNs).

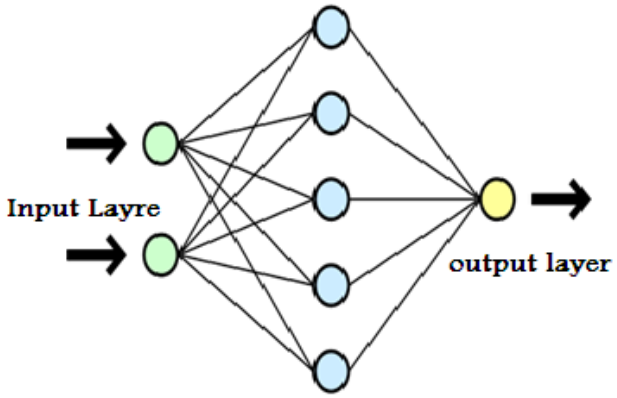


Figure 2: Diagram of simple neural network

The inputs are marked with the green circle, where the yellow circle marked as output component. Neural networks are not programmed in the usual sense of the word, they are trained. Education is one of the major advantages over traditional neural networks algorithms. Technical training is to discover the parameters of the connections between neurons. In the procedure of training a neural network is able to detect complex dependencies among the output and input, and perform synthesis. This means that, in the situation of successful learning, networking can give an exact integer result based on data that are not contained in the learning sample.

• ***Back propagation algorithms***:

More complicated is the multilayer networks case, as was initially unknown to the desired outputs of the network layers (except the last) and cannot be taught, guided only by the size of the network output errors, as was the case with the single-layer network. The most appropriate solution was the idea of signal propagation errors from network to its input, layer by layer. Algorithms that implement training network in this way were called back propagation algorithms. The most common variant of this algorithm we will continue to apply in the software implementation tasks.

III. Application of hidden Markov models for speech recognition.

Determination of word's endpoints, the problem of specification the initial and the finish of the sentence is very important task in the field of speech processing. Methods of speech's endpoints detecting are used for isolation of speech from background noise, as well as reducing the number of arithmetic operations because only the segments with speech signal are processed. The task of speech isolation from noise is very complicated, except the situations when signal/noise ratio is very High-quality recordings made in a studio. In this situation the energy of even the weakest sounds exceeds the energy of the noise and, therefore, to measure the energy of the signal is quite enough. However, such recording conditions usually do not occur in the real-life environment. In order to isolate a word from a continuous flow of information in real time, it is able to use a simple but very efficient method of Rabiner and Sambur for the determination of endpoints, based on estimating of frame power and zero-crossings rate. This method requires fewer amounts of calculations because of the requirement for more signal conversion from the time domain to the frequency one [6].

The frame energy, in this case, shall be referred as regularized sum of absolute quantities of the Amplitudes of discrete signal samples.

$$E = \frac{1}{K}\sum_{n+1}^{n} A_n \ldots\ldots(1)$$

Where K is the normalization factor, (*n*) is the frame's length. To determine the value of the energy, different methods of calculation can be used, such as Euclidean norm calculation, because the minimum memory unit is 2 bytes for the signal processor of tms320vc5505 series, and the

Resolution of the audio codec, used for digitizing audio signals, is 16 bits; the arrays consisting of double-byte elements are preferable as a structure for data storage. Normalization of the final value is needed to avoid the overloading of the bit grid [9-10]. The normalizing factor is to be chosen by the following reasons: because the codec resolution is 16 bits, in addition, The amount of the amplitude is able to be either negative or positive, the biggest absolute magnitude value, that can kept in two-byte character type, is equal to 216-1 = 32768, and the maximum sum of amplitudes' absolute value is equivalent to 32768 * 512. Based on the foregoing, and that the energy level stored in the two-byte character type, the normalizing factor is selected to be equal to the frame's length [7]. Zero-crossing rate is defined as the number of times when the source signal transforms its value and its sign is above the noise threshold. This value does not need any normalization, since the maximum value is equivalent to N-1.

Figure 3.presint show before the words "timing" with the presence of constant noise.

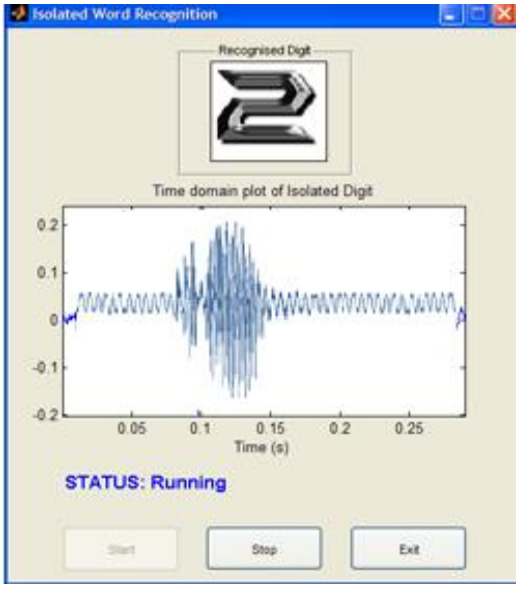


Figure 3: The timing diagram of the word.

Performing of only one does not guarantee the accuracy of determination of the end point. There are words consisting of periods of silence between phonemes, for figure 4 present show the word (“four”)

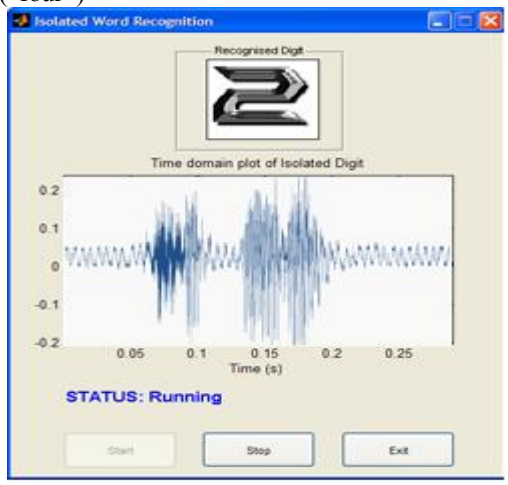
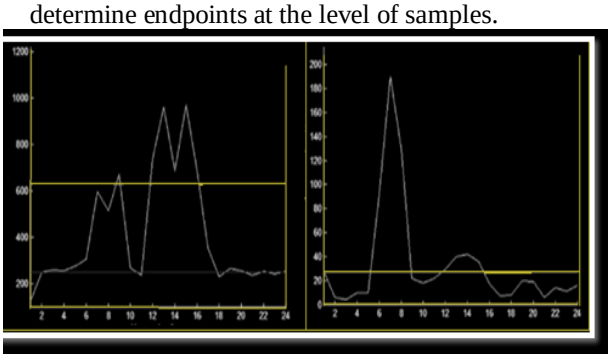


Figure 4: The timing diagram of the word (“four”)

These problems are resolved via introducing the maximum silence duration volume - the time interval when the parameter values may again exceed the respective levels within the interval [8].

If after the time expiration of this period the characteristics of the signal remain at the stage of the noise, then it is considered that the ending of the word is established and it locates in a fragment in which the first time energy level was less than ITL. Usually the time of silence is set to be ~ 0.1 s. In the current implementation, it equals to three frames, which keeps in touch with the time value ~ 130 ms. after finding the beginning and finish frames it is able to conduct a more analysis of the selected signal to determine endpoints at the level of samples.



(a) (b)

Figure 5: Changing of values: (a) the value of frame energy of the word ("four”), (b) - zero-crossings frames of the word.

IV. Isolation characteristics

Once the word was isolated from the input flow of data, the following step in the process of recognition is started, that means the process of characteristics isolation. There can be used a various methods for example a method of determination of Mel-frequency capacity coefficients, or linear prediction coefficients. The most important task in this step is isolation of some of signal’s parameters; provided that the quantity of these parameters shall be at minimum, so as to speed up the comparison with the parameter’s sets from the library, and simultaneously such parameters shall has such characteristics which are sufficient for perfect determination of a specific word [9].

4.1 Mel-frequency spectrum coefficients processor

The evolution of sensory systems possessed by creatures has passed the way ‘’to distinguish in order to survive”. The human hearing system as a sensory analyzer provides the distinction of sounds by their frequency content. However, the response to an acoustic stimulus should to be quick, and hence signal processing for the ear and the nervous system must be performed in a short time. Requirements of high frequency and time distinctiveness of analyzer are contradictory, but the result of the evolution was an optimal combination of these parameters [10]. Human acoustic organs have the ability of frequency masking, the situation where normally audible sound is covered by another loud sound with a close frequency. This feature based on the frequency of the signal and varies from 100 Hz for the low audible frequencies to 4000 Hz for the high frequencies. Consequently, audible frequency range can be divided into several critical bands division by 24 critical bands is generally accepted, which indicate a fall of ear sensitivity at higher frequencies. Critical bands can be considered as another sound characteristic, similar to its frequency. However, in difference to the frequency which is an absolute and does not dependent on acoustic organs, the critical bands are founded according to the auditory perception. As a result, they form some of the measures of frequencies

perceptions and for their measurement the measure units - bark and Mel - were coined [11].

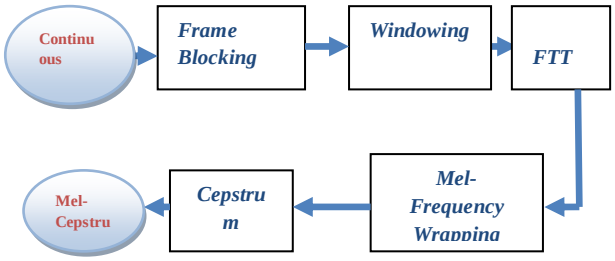


Figure 6: Structure of a processor.

4.2 Linear prediction coefficients

Linear prediction coding is a technique of coding of voice data by modeling the vocal tract. According to the model, a man can produce two types of sounds voiced and unvoiced. If the vocal cords vibrate when air passes there through from the lungs, thus the sounds produced are called voiced once. The sounds produced using the tongue, teeth and lips only are called unvoiced once.

For these two types of sounds the vocal tract may be signified as a series of cylinders of different radii with different energy value at the boundaries between the cylinders. Mathematically, this model can be shown as a linear filter, excited by the fundamental frequency to represent voiced sounds or by the white noise to represent unvoiced sounds [12]. The task of analysis depends on linear prediction is to obtain the parameters required to recreate the original sound: the sounds' type, the value of the fundamental frequency, the filter coefficients. The function of the linear prediction coding is to model the signal as a linear mixture of the previous samples (2). This method is quite effective, since the speech is a signal which is highly correlated over a short time interval, and hence a prediction can be made with minimal error. In the speech recognition task, this method is used to model the signal spectrum as autoregressive process [13-14].

$$s(n) = -\sum_{i=1}^{Nlp} aNlp(i) * s(n-i) + e_n \quad \ldots\ldots(2)$$

Where, Nlp is order of predictions (number of coefficients in the model, linear prediction coefficients, en - function of model error the difference between the predicted and actual values.
Providing that the quadratic error should be a minimal, the linear prediction coefficients are determined from the following system of normal equations shown in a matrix form:

$$R_{Nip} a_{Nip} = -r_{Nip} \quad \ldots\ldots(3)$$

where rk – autocorrelation coefficient of the speech signal weighted by windowing function w.
Linear prediction coefficients are calculated as follows:

$$aNlp = -R^{-1}{}_{Nlp} rNip \quad \ldots\ldots(4)$$

$$e_0 = r_0$$

$$for 1 \le m \le Nip$$

$$a_m(0) = 1$$

$$a_{(m)} = k_m = \frac{-r_m - \sum_{t=1}^{m-1} a_m - 1(i) * r_m - 1}{e_m - 1}$$

$$a_m(j) = a_m - 1(j) + k_m * a_{m-1}(m-j). \Rightarrow j = \overline{1.m-1}$$

$$e_m = e_m - 1 * (1 - k_m^2 \quad \ldots\ldots(5)$$

$$ci = \left[DCT(\ln(\ln(a_i{}^2))\right]^2 \quad \ldots\ldots(6)$$

Auto-correlation matrix has a Toeplitz structure, and for its solution there is an efficient Levinson-Durbin algorithm, the essence of which can be reviews in pseudo-code: To reduce the amount of stored parameters, the transformation (16) is applied to coefficients obtained in the previous step for calculating the spectral coefficients.

4.3 Dynamic Time Warping

The last step of the recognition is evaluation the input pattern with a set of standard patterns from the library. The recognition outcome is the index of the library template that is the most similar to the original block. But different implementations of speech patterns related to the similar class can be different considerably from each other in duration: this is according to the instability of the speaker's speech tempo caused by influence of intonation, accent, etc. For an accurate comparison of the speech pattern it is required to make their warping along the length. The warping by linear compression or stretching of one word realization till the value of another one does not resolve this mission, since a speech signal flows unevenly over time. This property of the speech is expressed by unevenly change of word's sounds duration when changing the word's length in general, so it is advisable to carry out a comparison with the non-linear time normalization [15]. For non-linear alignment of the compared patterns using an algorithm depends on the determination of the best match of input and reference speech patterns, known as the method of Dynamic Time Warping. The essence of the algorithm is like this, we denote the expanse among the"1"-element of the array of the input pattern parameters and the "j"-element of the array of the reference parameter Dij. In order to define the elements of input vector, which Symbolize the reference elements in the best way, the matrix C of size M * N is calculated by the following formulas:

$$c_{1.1} = D_{1,1}$$
$$c_{i,1} = D_{i,1} + c_{i-1.1,}$$
$$\rightarrow j = \overline{2..M}$$
$$c_{i,j} = D +$$
$$\min\left[c_{i-1,jc}, c_{i-1}, c_{i,j-1}\right],$$
$$\rightarrow i = \overline{2..M}, j = \overline{2..N} \ldots\ldots\ldots\ldots(7)$$

Where M is the quantity of elements in the input pattern; N is the quantity of elements of the reference. Distance Cij can be calculated in various behaviors. As a Euclidean distance (18), Manhattan distance (19), or Itakura-Saito distance (20). The final one is used if the feature vector consists of linear prediction coefficients.

$$D_{ij} = \sqrt{x_i^2 = x_j^2}$$
$$D_{ij} = \left|x_i - x_j\right|$$
$$\Rightarrow D_{ij} = \frac{xj}{xi} - \ln(\frac{x_j}{x_i}) - 1 \ldots\ldots\ldots\ldots(8)$$

Figure 6 shows elements of the matrix C connected by a broken line, corresponding to the most similar elements of the input term or word and the reference. The vertical line shows the state when the number of elements the links corresponds to a particular element of the input vector.

The horizontal line shows the state when the number of elements in the array of current templates is one element of the link. SEE, N contains a summary of the assessment of the similarity of the two characteristics of the vectors. After making comparison the input word with models from the library, the minimum score is chosen among all received summary and index model that depends on the minimum estimated appears as the outcome of recognition. Signals processing techniques provided in this article have been realized as a library of Matlab

Function for use with digital signal processors tms320vc5505 of Texas Instruments Brand. On the basis of this library there was created a test speaker dependent system of isolated words recognition with the small size of the dictionary, ``time-time" matrix is used to visualize the alignment. As with all the time alignment examples the reference pattern (template) goes up the side and the input pattern along the bottom.

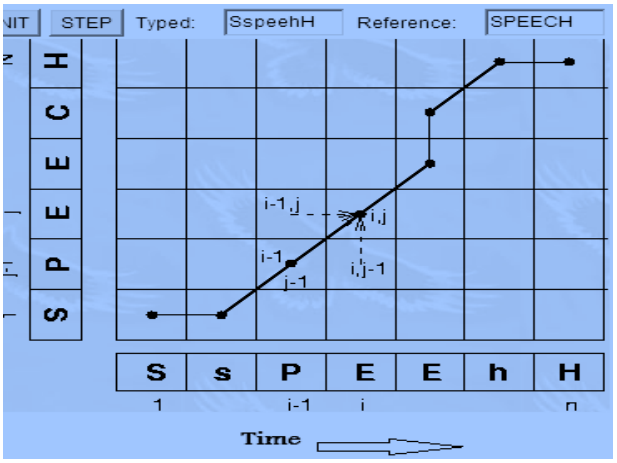


Figure 7: Dynamic time warping

The input "SsPEEhH", "noisy" version of the template ‘speech '. The idea that "h" is a closer match H compared to anything else in the template. Entering "SsPEEhH" will be matched against all templates in the repository system. Best of all it is a template that has a low distance path, aligning the template type in the template. Simple global distance estimation for path is just the sum of the distances that make way.

V. Conclusion

Provides a new idea for further research on speech recognition, a study of the hidden wavelet and neural network models are able to be used to produce more accurate forecasts. A new software library providing, an API for digital signal processing. It gives the essential functionality needed to build a simple recognition, real-time systems, combined with tools derived from digital signal processing. Speech system on which you be able to find extra complex solutions this library was created a recognition system depends on isolated words the speaker test with the small size of the dictionary.

A programmed has been developed for high-level programming languages Mat lab implements contained simulation algorithm of recognition of speech signals. The results have shown the ability to use the parameters of speech signals for speech recognition.

References

[1] An Automatic Speaker Recognition System. http://www.ifp.uiuc.edu/~minhdo/teaching/speaker_recognition.

[2] Siva P N ,, T. Kishore Kumar, Real Time Isolated Word Recognition using Adaptive Algorithm, 2012 International Conference on Industrial and Intelligent Information (ICIII 2012)IPCSIT vol.31 (2012) © (2012) IACSIT Press, Singapore, http://www.ipcsit.com/vol31/028-ICIII2012-30005.pdf

[3] Vucutury S. Multipath routing mechanisms for traffic engineering and quality of service in the Internet // PhD. Dissertation. - University of California - 2008. - 152 p.

[4] Wang Y., Wang Z. Explicit routing algorithms for Internet Traffic Engineering // Proc. of 8th International Conference on Computer Communications and Networks - Paris. - 2009. - P. 582 - 588.

[5] Oliver G, Carl J. Debono, Micallef P, "A Reproducing Kernel Hilbert Space Approach for Speech Enhancement" ISCCSP 2008, Malta, 12-14 March 2008

[6] Solera-Urena, Padrell-Sendra J, Martın-Iglesias J, Gallardo-Antolın A, Pelaez-Moreno A, and Dıaz-deMarıa F ,"SVMs for Automatic Speech Recognition: A Survey ", Progress in nonlinear speech processing Pages: 190-216, 2007, http://link.springer.com/chapter/10.1007%2F978-3-540-71505-4_11

[7] Wang Y., Wang Z. Explicit routing algorithms for Internet Traffic Engineering // Proc. of 8th International Conference on Computer Communications and Networks. - Paris. - 2009. - P. 582 - 588.

[8] Dr. Kekre HP, Ms. Vaishali K, "Speaker Identification by using Vector Quantization", International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1325-1331, http://www.ijarcsse.com/docs/papers/May2012/Volum2_issue5/V2I500451.pdf

[9] Anil K. Jain, Robert P.W. Duin, Jianchang Mao, (2000): Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22. pp.4-37.

[10] Sonia S, David P S, K Poulose Jacob, (2012): Optimal Daubechies Wavelets for Recognizing Isolated Spoken Words with Artificial Neural Networks Classifier, International Journal of Wisdom Based Computing, Vol. 2(1), pp. 35-41.

[11] Dr. H. B. Kekre, Ms. Vaishali Kulkarni, "Speaker Identification by using Vector Quantization", International Journal of Engineering Science and Technology, Vol. 2(5), 2010, 1325-1331. http://www.ijest.info/docs/IJEST10-02-05-134.pdf

[12] Sarma, M. P.; Sarma, K. K., ―Assamese Numeral Speech Recognition using Multiple Features and Cooperative LVQ – Architectures‖, International Journal of Electrical and Electronics 5:1, 2011.

[13] Sivaraman. G.; Samudravijaya, K., ―Hindi Speech Recognition and Online Speaker Adaptation‖, International Conference on Technology Systems and Management: ICTSM-2011, IJCA.

[14] Nadungodage, T.; Weerasinghe, R., ―Continuous Sinhala Speech Recognizer‖, Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011.

[15] DR. H. B. Kekre, Vaishali Kulkarni, "Performance Comparison of Speaker Recognition using Vector Quantization by LBG and KFCG," in International Journal of Computer Applications (0975-8887) Volume 3 -No. 10, July 2010. http://www.ijcaonline.org/volume3/number10/pxc3871086.pdf

[16] http://www.fzhjkj.com/en/yysb.aspx?gclid=CMeeo-m487kCFSTKtAodzE4A7g