World Essays Journal / 3 (2): 129-140, 2015
2015 Available online at www. worldessaysj.com

Designing an intelligent translation software by audio processing techniques

Neda Payande[1,*] Behnam Ghavami[2]

1. Computer engineering graduate student the tendency software, Javid High Education Institute
2. PhD computer, assistant professor martyr kerman Bahonar University

ABSTRACT: a lot of researches in the fields of image processing and sound processing are being done nowadays in the world. Usually, they use artificial Intelligence techniques, and different processing algorithms such as DSP, Genetic algorithms, neural networks, etc. creating an intelligent method to add ability to recognize words is objective of this research. This methodology by proper network training is able to separate and classify different audio signals. Finally, it determines some concepts for each group of sounds to learn by user. In this research, network with audio signals of numbers from zero to nine were taught in Persian language. Aim of network after training is separating input signals and finding the number corresponding to the input signal

***Keywords*:** Speech recognition Technology, language processing, data to speech conversion

INTRODUCTION

This research is going to provide an intelligent 2 way speech to speech translator in Farsi language, as it should translates sentences and words and phrases expressed by user and say them orally. It is a useful tool that facilitates communications between Iranian people who use Persian to communicate in other countries in their travels and other places, and foreigners. Additionally, it provides a good method for using English training books for users. The creation of such a system, which is called the diagnosis or speech recognition, in Persian language, has allocated several years of researches of scholars of different countries to itself. Making Farsi speech data and initial system of Persian speech recognition in intelligent center of signs were the most important success in 10 years ago.

Database to create an early warning system in the central smart signs. Recently, speech recognition systems of Guyesh Pardaz Company are the most important achievement of this technology for Persian language. Text-to- speech changing, machine translation, optical characters recognition and correction of typing errors, language models and language data are considered as requirements of artificial intelligent system. Guyesh Pardaz Company has applied the last technology of natural languages processes. Extraction of large volume of information in Farsi is result of it for the first time. Persian statistical language models, Persian grammar models, and different computing vocabulary of the Persian language are considered as data used in speech recognition systems of this company.

1. *History of Speech Recognition Technology*

In computer science and electrical engineering, speech recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT).Some SR systems use "training" (also called "enrolment") where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in more increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent".

Speech recognition applications include voice user interfaces such as voice dialing (e.g. "Call home"), call routing (e.g. "I would like to make a collect call"), domotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

The term voice recognition or speaker identification refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process.

From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the world-wide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems. These speech industry players include Microsoft, Google, IBM, Baidu (China), Apple, Amazon, Nuance, IflyTek (China), many of which have publicized the core technology in their speech recognition systems being based on deep learning.

The first system based on speech recognition technology were drawn in 1952 in

Bell Labs. It worked as limited to 10 words. It works as discrete Speech and dependent on the speaker.

Hidden Markov Model was presented in 1980 decade for the first time. This algorithm was considered as an important step to draw systems based on connected speech. Artificial Neural Networks and artificial intelligence were used to design this system. Dragon Naturally Speaking was the first product as the first continuous speech recognition system by James K.Baker Company in 1970.this system was able to recognize 160 words in one minute.

Raj Reddy was the first person to take on continuous speech recognition as a graduate student at Stanford University in the late 1960s. Previous systems required the users to make a pause after each word. Reddy's system was designed to issue spoken commands for the game of chess. Also around this time Soviet researchers invented the dynamic time warping algorithm and used it to create a recognizer capable of operating on a 200-word vocabulary. Achieving speaker independence was a major unsolved goal of researchers during this time period.

In 1971, DARPA funded five years of speech recognition research through its Speech Understanding Research program with ambitious end goals including a minimum vocabulary size of 1,000 words. BBN. IBM. Carnegie Mellon and Stanford Research Institute all participated in the program. The government funding revived speech recognition research that had been largely abandoned in the United States after John Pierce's letter. Despite the fact that CMU's Harpy system met the goals established at the outset of the program, many of the predictions turned out to be nothing more than hype disappointing DARPA administrators. This disappointment led to DARPA not continuing the funding. Several innovations happened during this time, such as the invention of beam search for use in CMU's Harpy system. The field also benefited from the discovery of several algorithms in other fields such as linear predictive coding and analysis.

During the late 1960's Leonard Baum developed the mathematics of Markov chains at the Institute for Defense Analysis. At CMU, Raj Reddy's student James Baker and his wife Janet Baker began using the Hidden Markov Model (HMM) for speech recognition. James Baker had learned about HMMs from a summer job at the Institute of Defense Analysis during his undergraduate education. The use of HMMs allowed researchers to combine different sources of knowledge, such as acoustics, language, and syntax, in a unified probabilistic model.

Under Fred Jelinek's lead, IBM created a voice activated typewriter called Tangora, which could handle a 20,000 word vocabulary by the mid-1980s. Jelinek's statistical approach put less emphasis on emulating the way the human brain processes and understands speech in favor of using statistical modeling techniques like HMMs. (Jelinek's group independently discovered the application of HMMs to speech.) This was controversial with linguists since HMMs are too simplistic to account for many common features of human languages. However, the HMM proved to be a highly useful way for modeling speech and replaced dynamic time warping to become the dominate speech recognition algorithm in the 1980s. IBM had a few competitors including Dragon Systems founded by James and Janet Baker in 1982. The 1980s also saw the introduction of the n-gram language model.

Much of the progress in the field is owed to the rapidly increasing capabilities of computers. At the end of the DARPA program in 1976, the best computer available to researchers was the PDP-10 with 4 MB ram. A few decades later, researchers had access to tens of thousands of times as much computing power. As the technology advanced and computers got faster, researchers began tackling harder problems such as larger vocabularies, speaker independence, noisy environments and conversational speech. In particular, this shifting to more difficult tasks has characterized DARPA funding of speech recognition since the 1980s. For example, progress was made on speaker independence first by training on a larger variety of speakers and then later by doing explicit speaker adaptation during decoding. Further reductions in word error rate came as researchers shifted acoustic models to be discriminative instead of using maximum likelihood models.

Another one of Raj Reddy's former students, Xuedong Huang, developed the Sphinx-II system at CMU. The Sphinx-II system was the first to do speaker-independent, large vocabulary, continuous speech recognition and it had the best performance in DARPA's 1992 evaluation. Huang went on to found the speech recognition group at Microsoft in 1993.

IBM Company worked on speech recognition project for several continuous years too. It produced Via Voive package in this field.

Microsoft Company worked in this field to produce and using this technology too. Bill Gates has emphasized on good future of using words recognition systems in his books and conferences.

The 1990s saw the first introduction of commercially successful speech recognition technologies. By this point, the vocabulary of the typical commercial speech recognition system was larger than the average human vocabulary. In 2000, Lernout & Hauspie acquired Dragon Systems and was an industry leader until an accounting scandal brought an end to the company in 2001. The L&H speech technology was bought by

Scan Soft which became Nuance in 2005. Apple originally licensed software from Nuance to provide speech recognition capability to its digital assistant Siri.

In the 2000s DARPA sponsored two speech recognition programs: Effective Affordable Reusable Speech-to-Text (EARS) in 2002 and Global Autonomous Language Exploitation (GALE). Four teams participated in the EARS program: IBM, BBN, Cambridge University and a team composed of ISCI, SRI and University of Washington. The GALE program focused on Mandarin broadcast news speech. Google's first effort at speech recognition came in 2007 with the launch of GOOG-411, a telephone based directory service. The recordings from GOOG-411 produced valuable data that helped Google improve their recognition systems. Google voice search is now supported in over 30 languages.

The use of deep learning for acoustic modeling was introduced during later part of 2009 by Geoffrey Hinton and his students at University of Toronto and by Li Deng and colleagues at Microsoft Research, initially in the collaborative work between Microsoft and University of Toronto which was subsequently expanded to include IBM and Google (hence "The shared views of four research groups" subtitle in their 2012 review paper). A Microsoft research executive called this innovation "the most dramatic change in accuracy since 1979." In contrast to the steady incremental improvements of the past few decades, the application of deep learning decreased word error rate by 30%. This innovation was quickly adopted across the field. Researchers have begun to use deep learning techniques for language modeling as well.

In the long history of speech recognition, both shallow form and deep form (e.g. recurrent nets) of artificial neural networks had been explored for many years during 80's, 90's and a few years into 2000. But these methods never won over the non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively. A number of key difficulties had been methodologically analyzed in 1990's, including gradient diminishing and weak temporal correlation structure in the neural predictive models. All these difficulties were in addition to the lack of big training data and big computing power in these early days. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around 2009-2010 that had overcome all these difficulties. Hinton et al. and Deng et al. reviewed part of this recent history about how their collaboration with each other and then with colleagues across four groups (University of Toronto, Microsoft, Google, and IBM) ignited the renaissance of neural networks and initiated deep learning research and applications in speech recognition.

The performance of speech recognition systems

Performance of speech recognition systems is the same for any application, as changing words to data and analyzing it by statistics models.

Converting speech into data

A system should pass a hard way to change speech to a text on a page or a computer command. Some vibration are made in air when a person talks. Speech recognition system receives analogue sound waves initially.

Analog-to-digital converter (ADC):

These waves convert analogue waves to digital data. Then, signal is divided into small segments as a few fractions of a second, or about plosive consonants as a few thousandths of a second. It converted these segments to known phonemes in language. Phoneme is the smallest element of a language. The program tests available phonemes with other phonemes beside it. It plots phonemes of the same context by a complex statistics model, and compare them with a large collection of known words, phrases and sentences. In next stage, user determines what user has said and gives it out as a text or picture or computer command or sound.

Speech recognition systems: classification based on performance

Speech recognition technology is classified based on 3 standards as follow:

A. Number of speakers

B. Way of talking

C. Bank of words

Speech recognition systems: classification based on output

Necessity of audio inputs is common feature of speech recognition systems. These systems are classified in 3 types as follows:

A-Speech to text systems

B-Speech o speech systems

C-Speech to commands systems

Natural Language Processing

Text-to-speech transformation, machine translation, optical character recognition, and correction of typing errors, language models and language information are required for artificial intelligence systems, such as speech recognition. Guyesh Pardaz Company has used the new technology in natural languages process to extract and apply language information. As a result of this effort, this company has extracted high amount of Persian data for the first time. Persian statistical language models, Persian grammar models and computing vocabulary for Persian language are considered as some data used in speech recognition system. It is possible to use these data in different forms in software and investigation.

Possibility to determine the level of correct pronunciation of words and phrases in educational software are considered as helpful capabilities. In addition to better instruction, it increases attraction of the software. This module can be used independently of the speaker and language or dependent on them.

Speech Quality Enhancement

Always a method is required for speech or sound quality enhancement by removing the extra scratch sounds and digitalized sounds of old audio tapes, or for recorded files in a conference. Guyesh Pardaz Company has applied modern technologies in this field, to investigate and produce a production for doing this work. It can be used both as an independent software, or is used as an independent unit in other software. For example, using this unit in speech recognition systems in noisy environments improves efficiency and accuracy of the systems.

Sound processing

Voice recognition or identifying speaker is one of issues of computer science and artificial intelligence that aims to identify a person based solely on a person's voice. Hidden Markov Models are considered as the main mathematical tools to solve this problem. Hidden Markov Models is a statistical model and determines the hidden parameters of observed parameters. The extracted parameters can be used for other analyzes. Voice pattern recognition by neural network is not investigated enough in Iran. There are few articles in this field. They introduced this topic generally. The results are quite practically. Result of it is software by MATLAB programming language. Results are presented as graphs and tables at the end of work. Different neural network methods are used in foreign articles. Sound samples are considered without any change as input data to network. It results in large networks, the length of the network training stages, high dependency of results to signal amplitude, and high sensitivity of results to noise. The presented method in this paper has decreased some of above problems due to a correction stage and data changing stage, including high dependency of network to voice tone and data that networks train by it. Therefore, the network requires a lot of data of different people, dialects, and accents to generalize performance of network.

METHODOLOGY

Stages of running this project from beginning to end are divided as follows:

1) Production of data
2) Reform of the raw data to provide it to network
3) Create a proper network
4) Network training

All of the above stages are implemented by tools and different instructions of MATLAB. First stage includes data providing. Data Acquisition Toolbox is used in this stage as follows:

- defining an analog input
- determining received input channel (sound card under the control of the operating system or..)
- Define input channel or channels (reference hardware may be multiple inputs)
- determining the frequency of sampling.
- determining the default input for sampling of the defined channels.
- Specify how to start sampling (a hardware stimulation or a software command to start).

Command to start sampling including a one-thousand loop to take 1000 signals from 0 to 9.

DISCUSSION AND RESULTS

- (Fig 1) is a signal of 1. Figure 2 shows signals of 0-9. Pattern of signal of other numbers are different, but pattern of same numbers are not match completely. There are some differences.

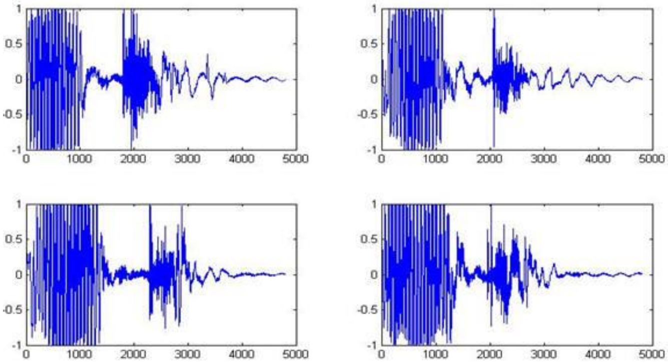


Fig. 1. Signal of 1.

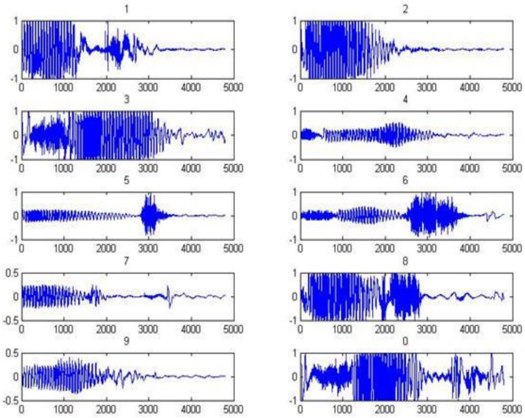


Fig .2. Signals of 0-9.

Each signal includes 4800 samples. Input signal is divided into 12 parts based on frequency signal, each one includes 400 samples. Then, a characteristic is extracted in each part that represents signal behavior. So, instead of 4800 samples, there are 12 samples results of project confirm it.

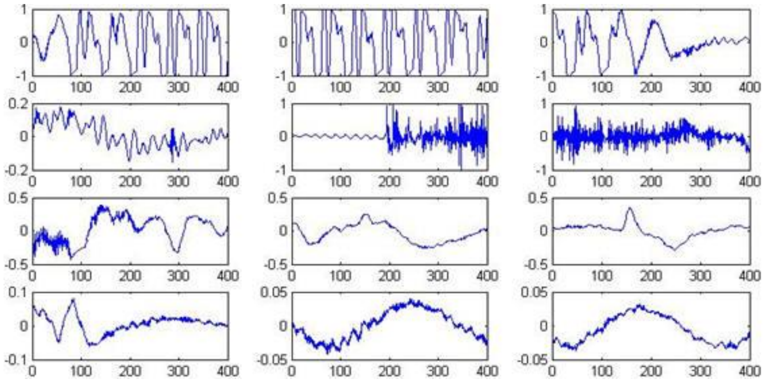


Fig. 3. Shows signal of 1 divided into 12 parts.

Fig 4 shows FFT of each section. Energy distribution is observed easily in frequency field. Half of data is enough for determining the dominant frequency.

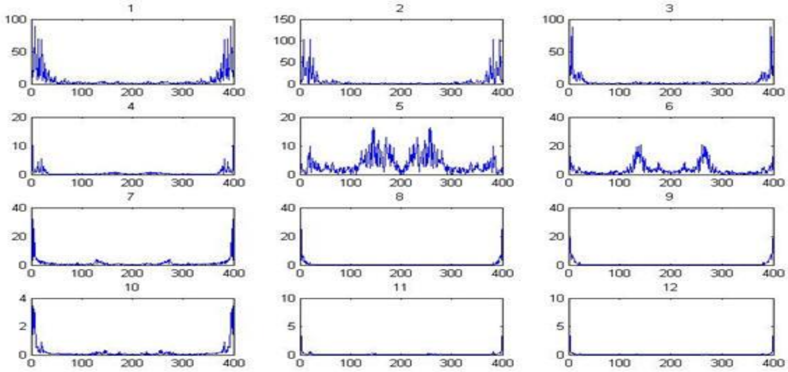


Fig. 4. shows FFT of each section

Different statistics methods are used for calculating the dominant frequency in investigations, but a specific averaging method has been used in this project.

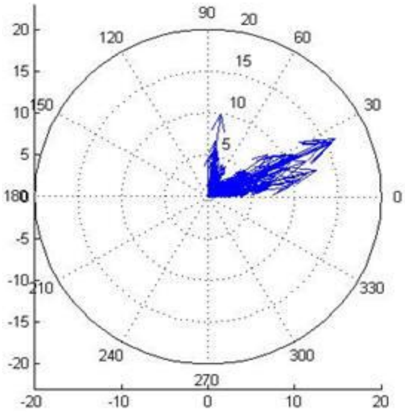


Fig. 6. The angle of the dominant frequency

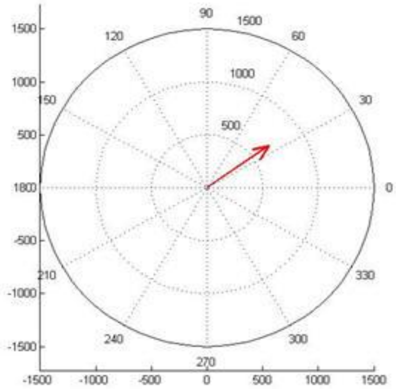


Fig. 5. Vector 1 factor

Obviously, there is a vector for each signal section. Angle of the vector represents the dominant frequency of that section.

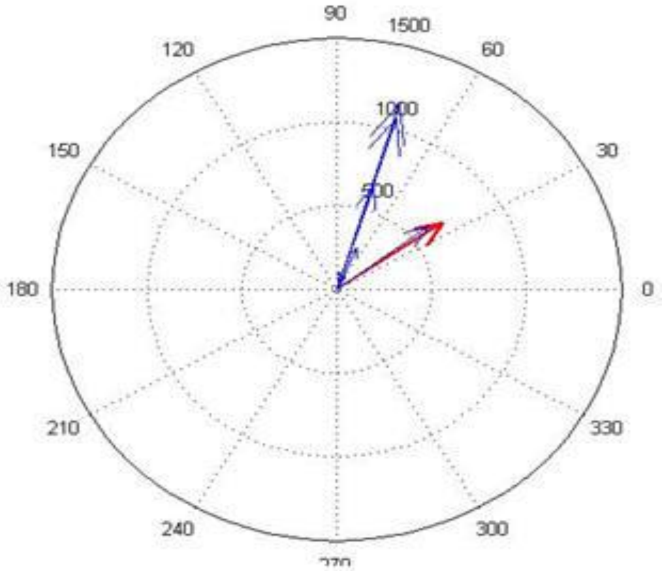


Fig .7. shows 12 vectors of signal beside each other

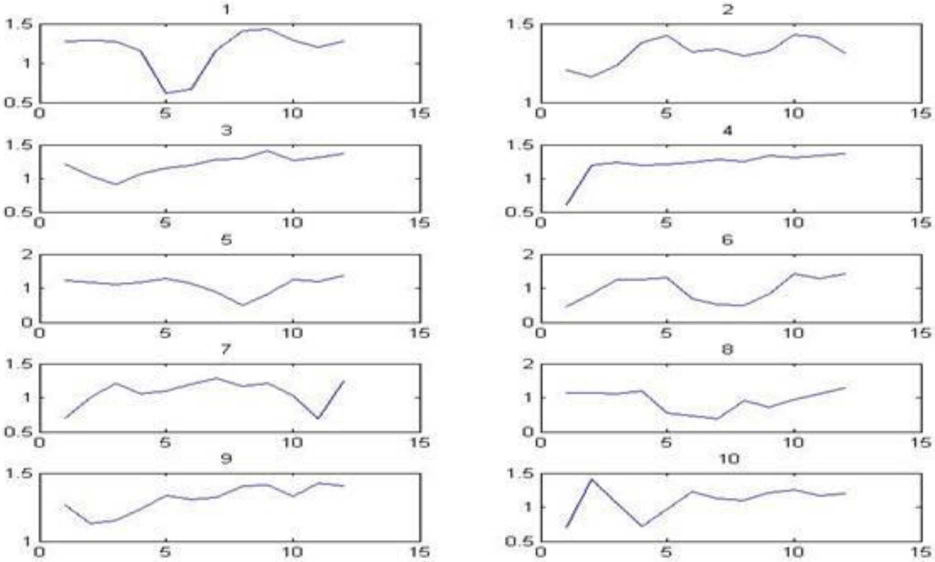
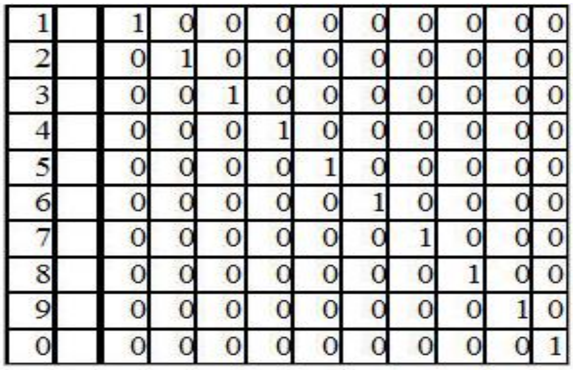


Fig. 8. Applying above algorithm on audio signals of 0, 1, 2,….9

Back-propagation two-layer competitive neural network is used in this project by LM training. 12 vectors are inputs of the network. Output of the network is targets of decimal vector. Fig 9 is a view of network. Fig 10 shows output pattern.



1		1	0	0	0	0	0	0	0	0	0
2		0	1	0	0	0	0	0	0	0	0
3		0	0	1	0	0	0	0	0	0	0
4		0	0	0	1	0	0	0	0	0	0
5		0	0	0	0	1	0	0	0	0	0
6		0	0	0	0	0	1	0	0	0	0
7		0	0	0	0	0	0	1	0	0	0
8		0	0	0	0	0	0	0	1	0	0
9		0	0	0	0	0	0	0	0	1	0
0		0	0	0	0	0	0	0	0	0	1

Fig. 9. A view of the network

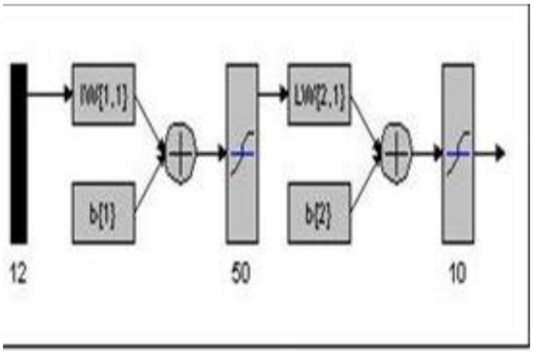


Fig. 10. Output pattern

The pattern considered for output creates a competitive position. After training to network, each output shows an amount resulted from competition of different neurons. Winner output neuron represents the highest amount, even if it is not based on the trained amount, i.e. 1. Levenberg-Marquard method is used for network training (fig 11) .it represents convergence of training. Accuracy is 0.0048 in epoch 150. Slope of the curve is closed to zero. It represents that accuracy more than 0048.0 is not possible.

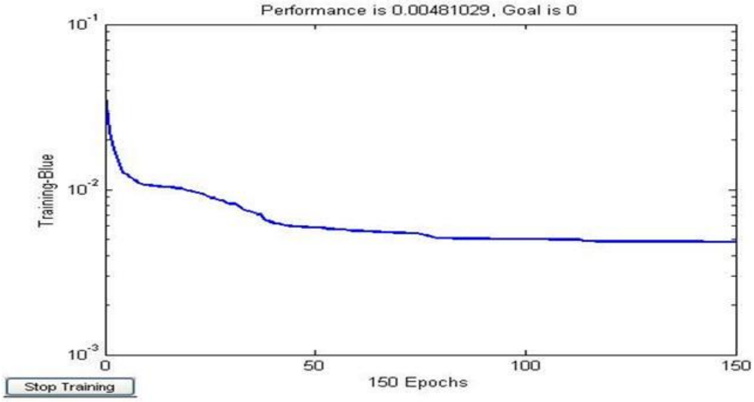


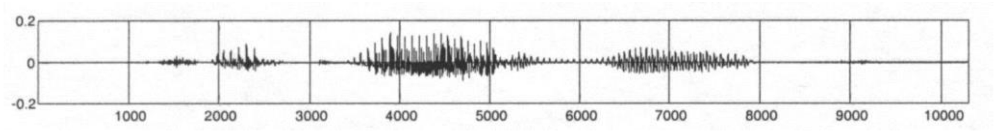
Fig. 11. represents convergence of training

2 different examinations are done to test the network.

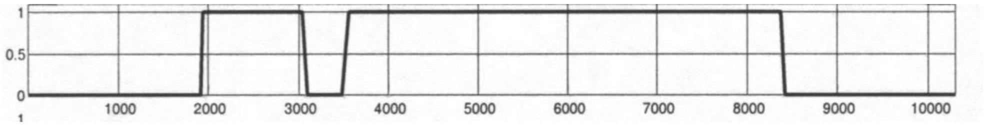
Automatic naming of different parts of speech

Identifying phonemes and determining boundaries of uttered syllables and words in sentences as follows:

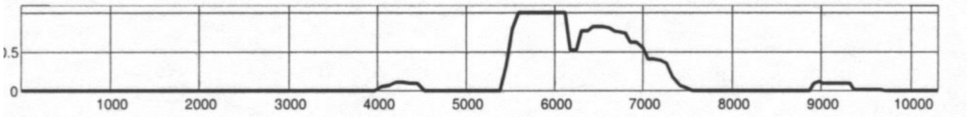
A) Identifying voiced phonemes
B) Identifying voiceless and silent phonemes
C) Naming voicing phonemes according to the place of phonemes
D) Naming of all the phonemes



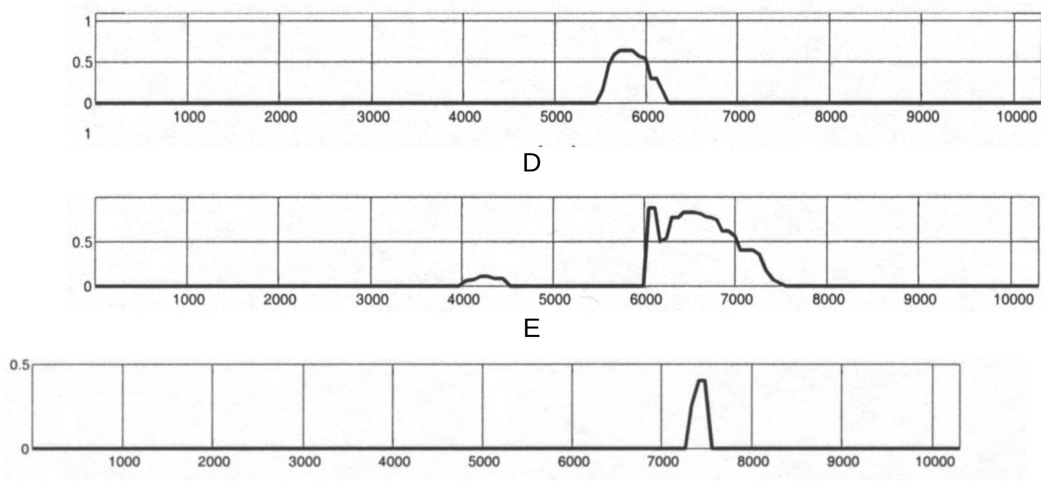
A



B



C



D

E

F

Fig. 12. Acoustic scoring of speech components: a) the part of speech, b) strong points, c) voiced scores consonant, d) nasal points, d) semi-vowel points, c) scores voiced wear.

Multi-Layer Perceptron neural network and 9 knots of input layer (including frequency data of 3 first ferments of the pieces and 2 nearby pieces) , The middle layer, respectively, with 20 and 30 knots and 29 knots in the output layer (according to the number of phonemes in Persian) were used to recognize voicing phonemes .

another MLP and 9 knots of input layer (including signal energy in frequencies more than 2KHz, signal energy in 0-5 KHz frequencies, and pass of zero rate of currency piece and 2 besides), 2 middle layer, respectively, with 20 and 30 knots, and 30 knots in external layer (according to the number of phonemes in Persian and silence) were used to recognize voiceless phonemes. Naming of voicing phonemes is done by the mentioned neural network scoring, and input text. 25 standards were applied for remove of excess MAX. Border of syllables is determined by different syllabuses styles in Persian language (CV, CVC, CVCC), by finding place of syllabuses, and considering its previous phoneme .

Step frequency model, duration, intensity and delay were determined by considering automatic naming of parts and sectioning. Therefore, instruction data is provided automatically for tone generating system. A natural language text-to-speech system was designed and implemented for Persian language as follows:

A) Natural language processing (including 2 outputs as "Phonological strings" and "syntactic information text words”)

B) Generating tone (to predict the step frequency model, energy model, duration data, pause in syllabuses and its components)

C) Synthesis of speech (by HNM method and with some corrections to change tone).

Some innovative methods were used for fast and automatic preparation of essential data to train neural network of generating tone for naming and sectioning of parts of speech. Efficiency of system was evaluated based on P.85 of ITU-T. MOS average in 6 studied criteria was 3.59. It is in modern rate of TTS for English language. Scholars work continuously to create variation in the parts of speech in synthesis data base, Semantic analysis in NLP, and the real-time system performance .

Conclusion

The frequency of correct answers is significant for signals of 0-9. According to results of this investigation, recognition of network is correct in more than 70% of items. In some cases, output of 2 or several numbers is 1.so diagnosis is wrong. Results were obtained by network training by 100 signals of each number. Error is decreased if network training is done by more data. In the best conditions, about an authorized word, one output is 1, but 9 outputs are 0. Variance and average are expected to be 0.1.

Distribution method of network output was used to diagnose unauthorized inputs. Distribution of outputs is uniform in most of cases. Sometimes, there is more than 1 Maximum simultaneously. Therefore, output variance would be small. Average recedes of 0.1. So, the error would be evident. Notably, network success in correct recognition depends on the number and variety of data.

REFERENCES

[1] Markowitz, J. A. (1996), “Using speech recognition”. New Jersey: Prentice Hall.

[2] Kinsler, L. Austin, R. Alan, B. and James, V. Sanders. (1999), “Fundamentals of Acoustics, 4th Edition”. New York: John Wiley & Sons.

[3] Towsey, M. Diederich, J. Schellhammer, I. Chalup, S. and Brugman, C. (1998), “Natural language learning by recurrent neural networks: A comparison with probabilistic approaches”. In Proceedings of the NIPS (Vol. 96).

[4] Nave, Carl R. (2003). “Department of Physics and Astronomy, Georgia State University”. How does an electric motor work.

[5] Molau, S. Pitz, M. Schluter, R. and Ney, H. (2001), “Computing mel-frequency cepstral coefficients on the power spectrum”. In Acoustics, Speech, and Signal Processing. (ICASSP'01). 2001 IEEE International Conference on (Vol. 1, pp. 73-76).

[6] Brian, C. Jay, B. and Andrew, R. (2007), “Snort IDS and IPS Toolkit”, Syngress Publishing.

[7] Krutz, R. L. (2005). “Securing SCADA systems”. John Wiley & Sons.

[8] Debar, H. (2000), “An introduction to intrusion-detection systems. Proceedings of Connect”.

[9] Abad, C. Taylor, J. Sengul, C. Yurcik, W. Zhou, Y. and Rowe, K. (2003). “Log correlation for intrusion detection: A proof of concept,”. In Computer Security Applications Conference. Proceedings. 19th Annual (pp. 255-264).

[10] Cannady, J. (1998). “Artificial neural networks for misuse detection,”. In National information systems security conference (pp. 368-81).

[11] Coolen, R. Luiijf, H.A.M. (2002). ‘‘Intrusion Detection: Generics and State-of-the-Art’’ , Research and Technology Organization (RTO) Technical Report 49.

[12] Debar, H. and Wespi, A. (2001). ‘‘Aggregation and correlation of intrusion-detection alerts’’. In Recent Advances in Intrusion Detection (pp. 85-103). Springer Berlin Heidelberg.

[13] Jazzar, M. and Jantan, A. (2008). ‘‘A novel soft computing inference engine model for intrusion detection’’. IJCSNS International Journal of Computer Science and Network Security, 8(4), pp1-9.

[14] Kendall, K. (1999). ‘‘A database of computer attacks for the evaluation of intrusion detection systems’’. Massachusetts Inst of tech Cambridge Dept of Electrical Engineering and Computer Science.

[15] Kumar, S. (1995). ‘‘Classification and detection of computer intrusions ’’ (Doctoral dissertation, Purdue University.