Investigation and Implementation of a Method for Extracting and Recognizing of the Digitalized Voice Sequence

Vladan Vučković[1] , Milena Stanković[2]

Abstract - **This paper analyses theoretical and implementation details of the efficient method for extracting and recognizing of the isolated speech sequence. In theory, there are many methods and approaches covering this domain. We have extracted and presented some of them, especially *Dynamic Time-Warping* and *Hidden Markov Models* approach. Implementation details connected with the machine isolated speech recognition in real-time are also concerned. The main principles and algorithms are implemented in author's experimental voice command processor – AREX.**

Keywords - **Digital speech processing, Machine speech sequence recognizing, Acoustic structure of speech, Digital voice spectrogram.**

I. Introduction

The machine recognition of the isolated speech sequence in real time is one of the most propulsive hi-tech development directions today. There are numerous applications and algorithms [1]. The main and common goal is to create and human-friendly voice guided interface for computers. This simple goal generates a huge amount of problems in realization, so we could say that there is not yet such a system today – the keyboard or mouse are still main input peripheral devices for computers. Theoretically, human speech contains many characteristics that are specific to each individual, many of which are independent of its semantic level. In speech recognition, they are generally considered as a source of degradation or noise. But, the biometric characteristic of speech could help listeners to recognize the speaker identity very quickly even over the week communication lines, like phone. The machine system-recognizing speakers rather than speech have been the subject of much research over the last two decades, and commercial systems are already in use.

Speaker recognition is a generic term for the classification of a speaker's identity from an acoustic signal [2]. For speaker identification, the speaker is classified as being one of a finite set of speakers. From the other hand, in the case of speech recognition, this will require the comparison of speech characteristics with a set of references for each potential speaker. For the case of speaker verification, the speaker is classified as having the system identity or not [3]. In this case, the goal is to automatically accept or reject an identity that is claimed by the speaker. As for the case of speech recognition, speaker recognition could implement different kind of applications of Hidden Markov Model (HMM) technology [4],[5]. The resulting approaches for these two application areas are very similar. In speaker recognition, each speaker is represented by one or several specific HMMs.

However, the main goal of this paper is speech recognition, independent from a concrete speaker. In the following sections, we pointed out, that there is one alternation for isolated speech recognition that is called *Dynamic Time-Warping* method [6],[7]. Of course, there are many other experimental approaches like artificial neural networks or chaotic fractal modeling, but we will focus on mentioned approaches. Digital voice signal separation and isolation is one of the most challenging problems in auditory perception. Good solution for many signal separation problems is necessary to improve the accuracy of automatic speech recognition systems in practical applications.

As technology for automatic speech recognition is transferred from research level into practical applications, the need to ensure robust recognition in a wide variety of acoustical environments becomes very important. Also, we could add that algorithms designed to handle the unknown additive noise and unknown linear filtering are numerous; today's applications also have good performance in many more difficult environments like: speech in high noise, with low signal-to-noise ratios (SNRs), in the presence of background speech or music etc. But, much investigation in this field still needs to be preformed.

The aim of this paper is to investigate some standard and advanced possibilities for real-time speech recognition. Some theoretical approaches like HMM, dynamic-time warping, support vector machines (SVM) [8],[9], gender separation [10] are briefly presented. Based on some of those approaches, in the second part of the paper the hardware/software of the author's application AREX for the real time digitalization, segmentation and recognition of the isolated voice sequence are presented [11].

II. The Basic Methods for Analyze of the Digitalized Speech Sequence

The basic methods for automatic recording and analyzing of the acoustic structure of vowels will be presented in this Section.

[1,2] Vladan Vuckovic and Milena Stankovic are with the Faculty of Electronic Engineering, Aleksandra Medvedeva 14, University of Nis, Serbia, Emails: vld@elfak.ni.ac.yu ; mstankovic@elfak.ni.ac.yu

A. Hidden Markov Model (HMM)

The Markov's stochastic process of the first degree is characterized with causal probability and could be represented with the next relation [5]:

$$p[x(k)|x(k-1),x(k-2),......,x(0)] = p[x(k)|x(k-1)] \qquad (1)$$

where x(n), n=0..k represents samples of the stochastic signal. The previous relation shows that probability that some sample gets value x(k) depends only from the last sample x(k-1) but not on all previous samples. Based on that relation, the Markov's signal x(k) could be defined with next recurrent relation:

$$x(k+1) = ax(k)+bv(k) \qquad (2)$$

where labels a and b are constants and v signal of the white (Gauss) noise. The causality of the previous state enables the definition of the finite state machine representing the HMM. So, HMM represents double stochastic process that generates sequence of symbols fitting the input sample. In that way, using the HMMs, the real speech could be represented as the series of the stochastic processes. Each of these processes could be defined as the one state in HMM. The change from one state to another is characterized with causal probability.

The next figure (Fig.1.) represents a paths through the one HMM. The speech recognition process starts for state 1, finally with state 6, where machine decides which signal is recognized. The hidden Markov models could be divided into the three categorizes: continual HMM (CHMM), Discrete HMM (DHMM) and half-continual HMM (SCHMM).

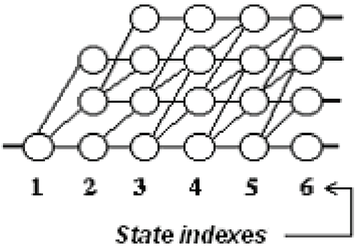


Fig. 1. The paths through the HMM (*trellis diagram*)

The method of recognition of the speech signal has a list of following steps [4]:

- In the phase of pre-processing of the speech signal, the serious of the spectral components are generated. This is an input vector X with length L.
- Next phase is to determine probability that input vector X is a part of some pre-defined models in HMM. For that reason the *Baum-Welth* algorithm [4] is used, this algorithms determines the optimal path through the HMM, based on input signal.
- For the HMM learning or upgrading purposes, the *Viterbi method* [5] is used.

B. Acoustic parameters

In speech recognition, the main goal of the acoustic processing module is to extract features that are invariant to the speaker and channel characteristics, and are representative of the lexical content. Speaker recognition requires the extraction of speaker characteristic features, which may be independent of the particular words that were spoken. There are many characteristics that could be used for those purposes. Such characteristics include the main properties of the spectral envelope – for instance, the average formant positions over many vowels (F1,F2,F3) or the average range of fundamental frequency (F0).

There are many methods for extracting such a features. The first method is *spectrogram* based on the usage of the *spectrograph* - device composed of electronic amplifiers and mechanical drawers, which are able to draw the harmonic structure of the pronounced word. In modern devices the new digital technique is used and programs *(Fourier transformations)* calculate specific spectral components [1],[11].

The results and diagrams are printed using standard peripheral devices - printers and plotters. The second important method, which will be mentioned, is *phonetography*. The *phonetogram* is two-dimensional representation of the specific acoustic parameters of the pronounced vowel. The horizontal axis contents the values of the fundamental frequency F0, and the vertical axis contains values of the sound intensity (SPL - *Sound-Pressure Level*). Each pixel in this array could have different intensity that is represented by appropriate color and density. The generation of the phonetogram is parallel with their pronouncing. Besides fundamental frequency and SPL, the system computes jitters in F0 as the measurement of the signal disequabling, the SPL difference among 0-1.5 kHz and 1.5-5kHz scopes as the measurement of the "sharpness" of the signal, and a quantity above 5 kHz as the criterion of the noise presence in the signal. *Bloothooft* has developed the device for the automatic generation of the phonetogram [1]. The central computer used in the original version is PDP 11/10 replaced with modern PC nowadays. The next picture presents the original diagram of the device (Fig.2.):

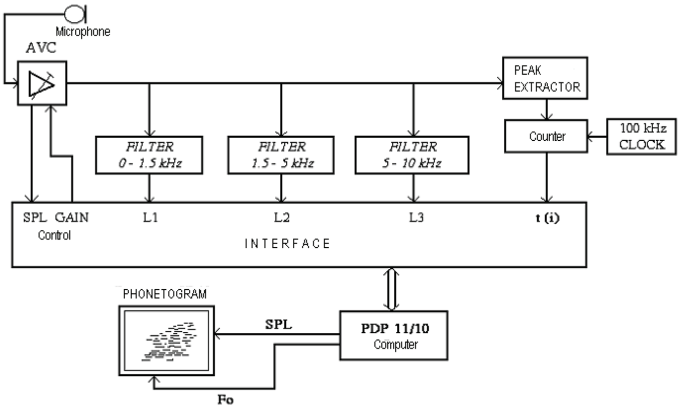


Fig. 2. Diagram of the automatic digital device for the phonetogram generation

The device runs in the similar way as the digital spectrograph. The signal from microphone is conducted through the automatic *gain level control unit* (AGC) to the system of parallel-connected analogue filters which function is to mark off the peaks and frequent scopes from the speech signals. The filters are connected with main computer via interface processing the received data and generating the phonetogram. The phonetogram may be used for medical purposes as well as the base for realization of the automatic voice recognition algorithm. The phonetogram reflects the picture attained from the speech signal transformed into the two-dimensional graphical representation so the recognizing of the speech adds up to the pattern recognition. The standard algorithms (for instance pattern matching) could be used for recognizing purposes. The usage of phonetogram as the basic data structure is the new interesting approach in the speech recognizing research array.

C. Improving Speech Recognition using Gender Separation

Most characteristics of speech are highly speaker dependent, and probability distributions in HMM suitable for a certain speaker may not be suitable for other speakers. The good examples of speaker-dependent parameters are between male and female vocal tracts, age group, differences in regional accents etc. In basic approaches of training independent speaker models these parameters are not considered, they try to be focused on common features. In training phase, these systems are tuned to the statistically include statistics over many speakers. Male and female speakers can be trained to improve the recognition performance given enough training data separate models. There are some good practical systems, which are improved from adding gender-dependent parameters [10].

D. An Support Vector Machines Model for Isolated Word Speech Recognition

There are a few classic acoustic-modeling approaches for speech recognition; e.g. HMM, neural networks, stochastic segment models... The recognition rates of the systems that are based on these classic models have been limited due to computational complexity or model limitations. Support vector machines (SVM) are one of the most successful modeling strategies in pattern classification problems [8]. This success is connected to SVM capability to make a good balance between learning and generalization characteristics [9].

There are a few main additional ideas that outstands SVM from other linear discriminator pattern recognizers. SVM contains just the training samples near the boundary to represent the classification borders; it employs not all the training samples. The parameters of the boundary are mainly estimated by the training samples whose classes are not obvious. Hence, the classifier focuses on potentially misclassified samples. The recognition error rate is optimized by maximizing the margin distance among classes in the estimation phase in contrast to probability density estimating procedures (e.g. HMM, MLP, etc. [9]), which normally minimize the mean square error through all samples. Selection of the boundaries to maximize the margin between two classes makes the generalization capability of the system optimal base on a given the known training samples. The SVM classifier is also able to handle nonlinear boundaries in complex feature spaces that are another advantage of them.

E. Human Audio Perception Model for Signal Separations

The human auditory system uses a number of well-defined cues to separate and isolate individual sound sources in a complex acoustical environment [2]. The use of these cues to achieve acoustic sources grouping and signal separation and solution should be very useful in improving the accuracy of automatic speech recognition in very difficult environments (background music and noise). This researching area is very benefit and has become a goal of several research groups in computational auditory scene analysis. Unfortunately, the analyses of the human perception is widely interdisciplinary approach, connected with behavior medical investigations, so we are still quite a far from the first concrete implementations in automatic voice separating [2].

III. The Application For Automatic Voice Sequence Sampling and Recognizing

The authors have developed the application named A.D.S. v2.0 for the some medical research of the stress influence to the parameters of the human speech [11]. As the part of that program system, the separation procedure is developed. The basic idea is similar with using of the phonetogram although the different set of features is used. In our application we use primary two characteristic simultaneously: the energetic and frequency. As the improvement of this basic application, the AREX application is developed [11]. The AREX has recognizing feature added. As the empirical results show, this approach implicates much faster execution compared to the previous methods. The success ratio is relatively high (above 95%) in respect of simplicity and running speed of the program. Of course, there is potentional for further improvements of the algorithm but the main conception is successfully sustained. After that determination, the wave period could be automatically recognized consulting the wave shape database using the standard pattern-matching algorithm.

A. The Realization of the Hardware

The hardware subsystem that supports AREX application for isolated speech sequence recognition consists of personal computer equipped minimally with 1.8Ghz CPU. The solution for automatic sequence determination is based on amplifying and filtering of the input signal. The analogue signal is parallel conducted though *Sound Blaster* and impulse recognizer using the *Busy line* of LPT printer port (Fig. 3). In that way, the speech signal generates two features; one is sampled signal (16-bit, 22kHz) and the other is serious of impulses through the stabilizer 7805, which acts like voltage limitator [11].

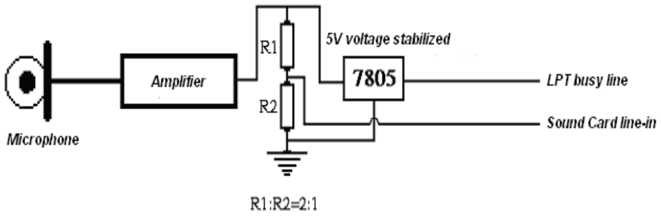


Fig.3. Support Hardware for AREX

The application scans LPT 500 times per second, and calculates changes in frequency. If these changes are rapid or exceeds pre-defined threshold, the interval is determined.

B. Dynamic Time Warping

Dynamic Time Warping has much simpler theoretically basis compared to HMM, so it is very well suited for the systems with low processing power [6]. This method is used for one dimension signal recognizing which is exactly the speech signal. The procedure tries to shrink or exceed some parts of the input signal in order to fold the original and database pattern. The variations in the input voice signal are unavoidable so method is very good in their handling.

The process is illustrated in Fig. 4.:

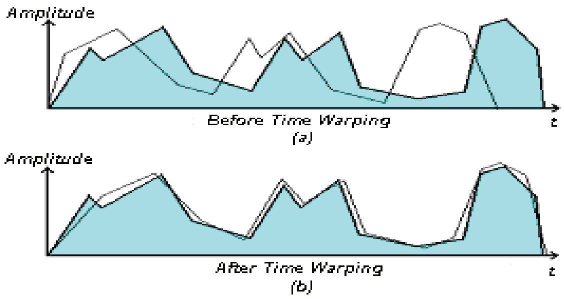
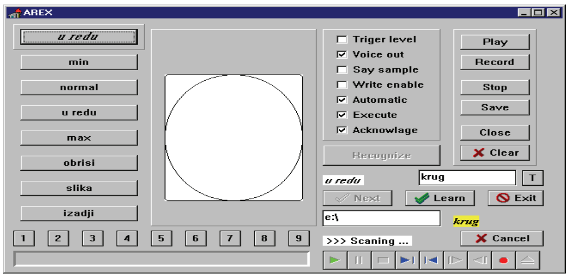


Fig. 4. Dynamic Time Warping

The Fig.4. (a) shows the original and database signal before the processing. and the Fig.4. (b) after it. After the dynamic time warping, a pattern matching recognition is employed. This method is very fast and is the basic algorithm in AREX [7].

C. AREX – Voice Recognition Application

The program AREX is the application for machine recognition of the isolated speech sequences and execution of the corresponding pre-programmed commands (Fig.5.):



☐ Triger level
☑ Voice out
☐ Say sample
☐ Write enable
☑ Automatic
☑ Execute
☑ Acknowlage

u redu

krug

e:\

krug

>>> Scaning ...

Fig.5. AREX for Windows

The application performs like voice driven command processor. It is speaker independent recognizer. Using the hardware subsystem, the application is very fast and reliable. The separation method is based on rotating arrays and recognizing on time-warping method with large amount of digital pre-filtering [7],[11].

IV. Conclusion

This paper describes the theoretical bases of the efficient methods for extracting (separating) and recognizing of the digitalized voice sequence. In this propulsive researching domain, there are numerous theoretical approaches and algorithms. Some of them are briefly inspected in the paper. The main intention of this paper is to give the concrete empirical contribution to the research field of the fast voice sequence analyzing and automatic recognizing on PC machines. As the practical part, the AREX application, based on time-warping method is developed. The supporting hardware and software are also included. The application proved to be very good bases for the future research of the fast recognition algorithms.

References

[1] L. R. Rabiner, R. W. Schafer "*Digital Processing of Speech Signals*", Bell Laboratories, Prentice-Hall, Inc. , U.S.A. , 1978.

[2] Richard D. Peacocke, Daryl H. Grat "An Introduction to Speech and Speaker Recognition" , IEEE Computer, Vol. 23, No. 8, pp. 26-34, August 1990

[3] Rabiner, L., On the use of autocorrelation analysis for pitch detection. IEEE Trans.ASSP, 25(1): pp. 24 - 33., 1977.

[4] X.D.Huang "Phoneme Classification Using Semicontinous Hidden Markov Models" , IEEE Transactions on signal processing, Vol. 40, No. 5, pp. 1062-1067, May 1992.

[5] Biing-Hwang Juang, Kuldip K. Paliwal "Hidden Markov Models with First-Order Equalization for Noisy Speech Recognition" , IEEE Transactions on signal processing, Vol. 40, No. 9, pp. 2136-2143, September 1992.

[6] James W. Pitton, Kuansan Wang, Biing-Hwang Juang "Time-Frequency Analyses and Auditory Modeling for Automatic Recognition of Speech" , Invited Paper, Proceedings of the IEEE, Vol. 85, No.9, pp. 1199-1215, September 1996.

[7] Vučković Vladan "Dynamic Time-Warping Method forIsolated Speech Sequence Recognition", V International Conference on Telecommunication in Modern Cable, Satellite and Broadcasting Services - TELSIX 2001, Proceedings of Papers, Volume 1, pg. 257-260. , Niš, 19-21. September 2001.

[8] C. J. C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition". Knowledge Discovery and Data Mining, 2(2), 1998.

[9] A. Ganapathiraju, "Support Vector Machines for Speech Recognition", Ph.D.Thesis, Mississippi State University, MS State, MS, USA, 2002.

[10] Abdulla, W.H. and Kasabov, N.K. Improving speech recognition performance through gender separation. In: Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES2001), pp. 218-222., 2001.

[11] Vučković Vladan "Digital Processing and Machine Recognizing of the Isolated Speech Sequences". MasterTheses, The Faculty of Electronic Engineering, May1997. (in Serbian).