Research Journal of Applied Sciences, Engineering and Technology 4(1): 45-50, 2012
ISSN: 2040-7467

Submitted: September 07, 2011 Accepted: September 30, 2011 Published: January 01, 2012

Speaker Identification System using Autoregressive Model

Moh'd Rasoul Al-Hadidi
Computer Engineering Department, Engineering College, Al-Balqa Applied University,
Al-Salt 19117, Jordan

Abstract: The Autoregressive Model is used as a tool to design a recognition system that is presented by speaker identification. The main goal of this paper is to design a speaker identification system by using the autoregressive model to identify the identity of speaker according to the voice frequency, the speaker is being asked to say a certain word then it's matched to the same word stored earlier in the database. This research is based on speech recognized words using the Autoregressive Model based on a limited dictionary.

Key words: Autoregressive model, envelope detection, speaker identification

INTRODUCTION

Every year there is a new technique arises to be used in our life to make it more comfortable and easy to interact with the surrounding environment. Our voice is the most natural way that used to interact with people and machines, so we can use it to do any job and remote any machine.

Speech recognition process is the process in which a computer identifies the spoken words. It means that when you talk to your computer, it will recognize your words. Voice recognition"is the technology by which sounds, words or phrases are spoken by humans that are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned" (Rabiner and Juang, 1993).

Speaker Recognition is a process by which the speaker can be recognized. This process has two types of job: First is the Speaker Identification (SI) in which the speaker can be recognized according to the matching process between the input sample with the samples which is stored in the database of the system. The second is Speaker Verification (SV) is the process by which the system accept or reject the identity claim of a speaker (Kinnunen *et al*., 2006). Figure 1 and 2 show the structure of the speaker recognition system.

There is a main difference between speaker identification and speaker verification presented by two cases, the first is for each speaker the system provides one model, and the second case, the system provides a total of two models: one for the hypothesized speaker and one representing the hypothesis that the speech sample comes from some other speaker the background model (Grimaldi and Cummins, 2008). There are many techniques which were used in the speaker recognition, such as: the using of Hidden Markov Modeling (HMM) (Doddington *et al*., 2000), the using of Gaussian mixture models (Reynolds, 1995), and the using of the Artificial Neural Networks (ANNs) as a good solution and to yield a good performance (Clarkson *et al*., 2001; Phan *et al*., 2000) .

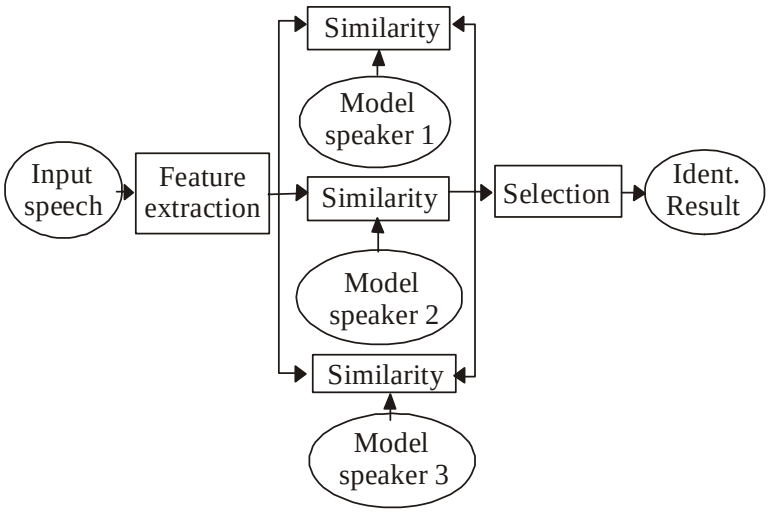
The Autoregressive model can be defined as a type of random process. A various types of natural phenomena can be modeled and predicted by using the autoregressive mode. The prediction of an output signal is based on predict an output of a system according on knowing the previous outputs. The autoregressive model is one of a group of linear prediction.

An autoregressive model is simply a model used to find an estimation of a signal based on previous input values of the signal. The actual equation for the model is shown in the Eq. (1):

$$y(t) = \sum_{i=1}^{m} a(i)y(t-i) + \in (t) \quad (1)$$

The model contains of three parts: a constant part, an error or noise part, and the autoregressive summation represents the fact that the current value of the input depends only on previous values of input just like the correlation model. The variable m represents the order of the model. The higher the order of the system the more accurate a representation it will be. Therefore, as the order of the system approaches infinity, we get almost an exact representation of our input system.

The main Significant Contribution of this research study is the using of the Autoregressive tool to design a system that identify the identity of the speaker according to the voice frequency, another tools are the Envelop detection, the Fast Fourier Transform (FFT), and finding the formants of vowels.



Fig, 1: Speaker identification (Melin *et al*., 2006)

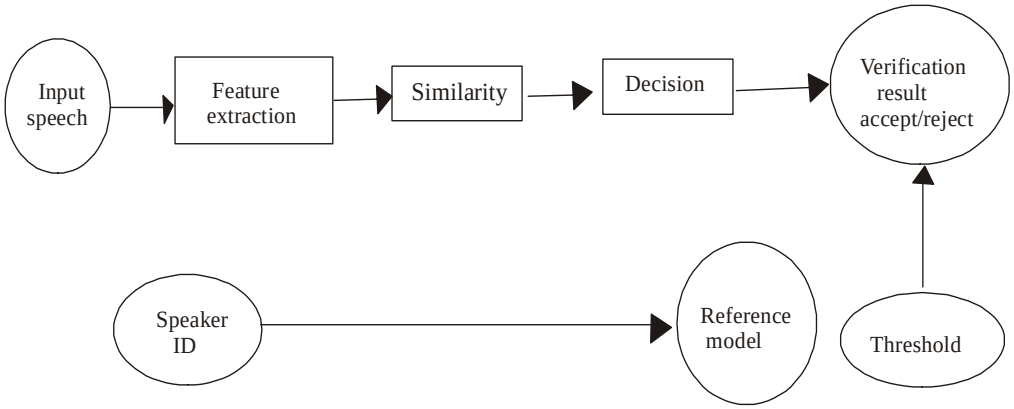


Fig. 2: Speaker verification (Melin *et al*., 2006)

The Fourier transform is a mathematical operation that is used to provide a frequency domain signal from a time domain signal.

Fourier transform is based on discovery that it is possible to take any periodic function of time f(t) and resolve it into an equivalent infinite summation of sine wave and cosine wave with frequency that start at 0 and increase in integer multiples of base frequency F = 1/T where T period of F(t) (Stremler, 1990). By using the following equations we can find the fourier series for a periodic functions:

$$f_T(t) = \sum_{n=-\infty}^{\infty} Fne^{jnowt} \quad (2)$$

$$F_n = \int_{-T/2}^{T/2} f_T(t)e^{-jnwot} \quad (3)$$

The rest of study presents some related works and research studies that related to this study, then explain the experiment of this study and the actual steps of designing the system, Finally the conclusion and some recommendations of future works are presented.

Related works: There are many researches presented in speaker identification system, the following explanation introduce some them.

In the Phan *et al.* (2000). research study they proposed a system of speaker identification system by using the Artificial Neural Network (ANN) and the wavelet transform. They present an off-line system that uses the wavelet to generate multiresolution time-frequency features that characterize the speech waveform to successfully identify a speaker in the presence of speakers. They discuss ALOPEX, which is an optimization paradigm that incorporates the features into a recognition system that used a feed forward artificial. neural network (Phan *et al*., 2000).

Yuo *et al*. (2005) proposed a robust approach for speaker identification when the speech signal is distorted by the noise and a channel distortion. The Robust features are derived by assuming that the corrupting noise is stationary and the channel effect is fixed during an utterance. The system is proposed by two steps temporal filtering procedure on the autocorrelation sequence to minimize the effect of additive and convolutional noises. The first step of this system applies a temporal filtering procedure in autocorrelation domain to remove the additive noise, and then the second step is to perform the mean subtraction on the filtered autocorrelation sequence in logarithmic spectrum domain to remove the channel effect. The additive noise in the voice signal can be a colored noise. Then the proposed robust feature is combined with the projection measure technique to gain further improvement in recognition accuracy. The results of the proposed system show that this method can significantly improve the performance of speaker identification task in noisy environment (Yuo *et al*., 2005)

Another kind of studies was proposed by Hetingl *et al*. (2006) which used the explicit lip motion information, in addition to lip intensity and geometry information, for the speaker identification and speech-reading within a unified feature selection and discrimination analysis framework, and by using two important issues:. First the useful of using explicit lip motion information, Second what are the best lip motion features for these two applications?.

The best lip motion features for speaker identification are considered to be the result in the highest discrimination of individual speakers in a population, whereas for speech-reading, the best features are those providing the highest phoneme/ word/phrase recognition rate. Several lip motion feature candidates have been considered including dense motion features within a bounding box about the lip, lip contour motion features, and combination of these with lip shape features. Furthermore, a novel two-stage, spatial, and temporal discrimination analysis is introduced to select the best lip motion features for speaker identification and speech-reading applications. The results show that the using of the Hidden Markov Model based recognition system indicate that the explicit lip motion information which is used in this system provides additional performance gains in both applications, and lip motion features prove more valuable in the case of speech reading application (Hetingl *et al*., 2006).

In 2007, Aronowitz and Burshtein proposed a speaker identification system by using Approximated Cross Entropy (ACE). They used Gaussian mixture modeling for representing both training and test sessions and to perform speaker recognition and retrieval extremely efficiently without any notable degradation in accuracy compared to classic GMM-based recognition. They presented the GMM compression algorithm. This algorithm decreases considerably the storage needed for speaker retrieval (Aronowitz and Burshtein, 2007).

In the same year, a robust speaker identification and verification research study is proposed by Wang *et al*. (2011). They introduced a robust and text-independent speaker identification/verification system. The proposed system based on a subspace-based enhancement technique and probabilistic Support Vector Machines (SVMs). First, a perceptual filter- bank is created from a psycho-acoustic model into which the subspace-based enhancement technique is incorporated. They used the prior SNR of each subband within the perceptual filter bank to make decision about the estimators gain to effectively suppress environmental background noises. Then, probabilistic SVMs identify or verify the speaker from the enhanced speech. The proposed system has been demonstrated by twenty speaker data taken from AURORA-2 database with added background noises (Wang *et al*., 2007).

In another study, Wang *et al*. (2007) introduced a Robust Speaker Recognition system Using Denoised Vocal Source and Vocal Tract Features. They proposed this system to alleviate the problem of severe degradation of speaker recognition performance under noisy environments because of inadequate and inaccurate speaker discriminative information; they proposed a method of robust feature estimation that can capture both vocal source and vocal tract-related characteristics from noisy speech utterances. And they employed a Spectral subtraction, a simple yet useful speech enhancement technique, to remove the noise specific components prior to the feature extraction process. They proposed feature estimation method which leads to robust recognition performance, especially at low signal-to-noise ratios. In the context of Gaussian mixture model-based speaker recognition with the presence of additive white Gaussian noise, the new approach produces consistent reduction of both identification error rate and equal error rate at signal-to-noise ratios ranging from 0 to 15 dB (Wang *et al*., 2011).

EXPERIMENT

Recording the voice: In this system we record the voice of the speaker(s) which will be recognized by the system. There are many methods of the recording process such as, the sound recorder which is in the accessories of the windows (start menu), the Audio recorder with any program with three inputs (N, Fs and CH). That is record N audio samples at Fs Hertz from CH number of input channels. With the WAVE recording as output, and the third method is a group of specific commands which are written in the command window in the Matlab and record the desired voice. In this system we record the voice of 60 users (30 male users and 30 female users) that will be recognized. These sounds were saved in the database to be used in the recognizing stage.

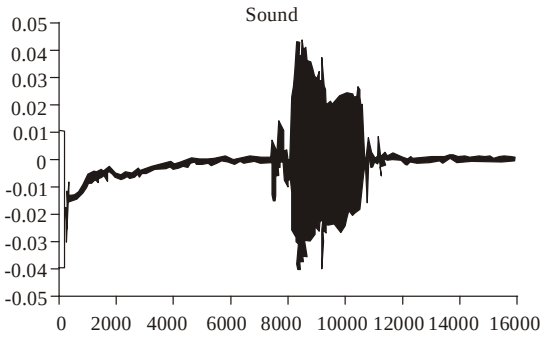


Fig. 3: Voice signal

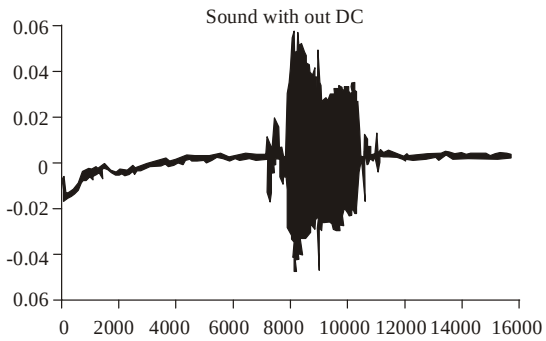


Fig. 4: Signal after removing DC component

Inserting the voice: While the recording process was done in the offline stage, the inserting of the voice was in the testing stage (online stage). The inserting process was achieved by using any type of microphone, external or internal microphone.

After plugging the microphone in the computer and recording the voice, it will look as shown in Fig. 3, the signal was obtained by using the built in function plot for the matrix returned from Waverecord function.

Preparing the signal: Preparing the signal includes several steps such as:

Step 1: Removing the DC components
Step 2: Squaring the signal to see the peak
Step 3: Set the maximum value of the signal to one

In this system the preparing step is done by removing the frequency jaggedness in the signal and leaves behind simply the magnitude of the signal. So we have a clear signal that is fairly easy to process, as shown in Fig. 4.

The second step is achieved by squaring the signal so we can examine the peaks more efficiently (Fig 5).

The third step includes normalizing the signal and then set the maximum value of it to one. This step is done to account different volumes of speakers; the signal must

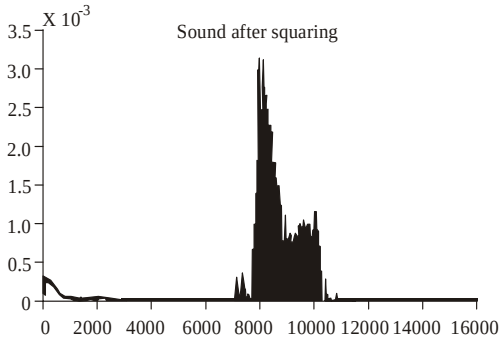


Fig. 5: Signal after squaring

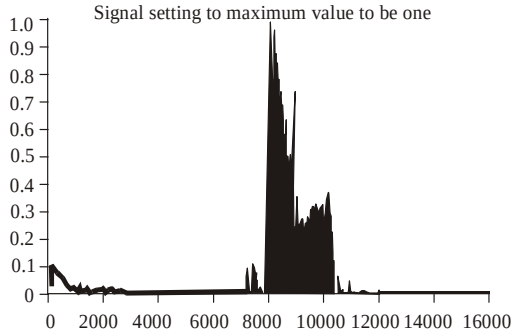


Fig. 6: Normalized signal

be normalized to the same volume before they are examined. Each signal is normalized about zero such that all of the signals will have the same relative maximum and minimum values, and so that comparing two signals with different volumes is the same as comparing the same two signals if they were to have the same value. Fig. 6 shows the normalized signal.

Fast fourier transform: The importance of the FFT appears in various digital signal processing applications, such as linear filtering, correlation analysis. Fourier transform is based on discovery that it is possible to take any periodic function of time f(t) and resolve it into an equivalent infinite summation of sine wave and cosine wave with frequency that start at 0 and increase in integer multiples of base frequency F=1/T where T period of F(t) (Stremler, 1990). The synthesize Eq. (4) that represent the Fourier series for periodic functions can be calculated by summation all Fourier coefficients values Fn - which is defined by Eq. (5)- multiplied by exponential function:

$$f_T(t) = \sum_{n=-\infty}^{\infty} Fne^{jnwot} \quad (4)$$

$$F_n = \int_{-T/2}^{T/2} f_T(t) e^{-jnwot} dt \quad (5)$$

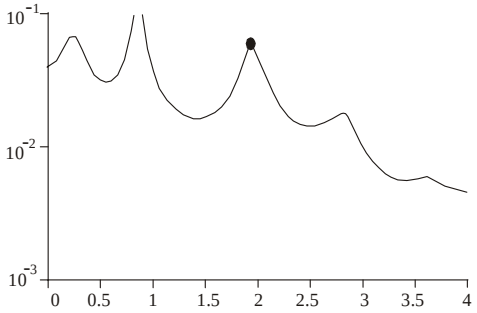


Fig. 7: Formant analysis

In this system, the Fast Fourier Transform is used to examine the signal as discrete frequency samples.

Envelope detection: In this step the filtering process and the envelope detection process achieved in our system, but we need to take into consideration to choose the right threshold voltage.

This step enables us to examine each individual peak alone, just after the signal is smoothed by the filter, we use an envelope function to detect all of the peaks related to signal, this guarantees us that if the signal passes a certain threshold amount, it will be examined and compared with the corresponding signal in the database. The analysis will not include the entire signal, but rather a formant analysis, or vowel sounds in the signal will be examined and those will be used to verify the speaker.

After applying the detection process on the signal, the obtained signal will has a shape similar to the following fig. 7. For any signal, the process of detection is repeated until all peaks are examined; for example the word boat will be examined twice to detect A and O.

In this step the varying of speed is solved; the envelope of the peak will determine which vowels are available, and the actual formants themselves will be relatively unchanged. It is difficult to handle very high speed voices, but most other can be handled effectively.

Handling formants: After the signal is broken down into frequency samples and each peak represent a certain formant or a sound vowel, the corresponding axis the actual frequency of the vowel. This data are stored in a database with the information of the speaker.

In this system we tried to ensure the reliability, so we tried different words with different vowel sounds and each time we almost had the same result but to make sure we took the average of these trials. This system is trained and learned by using the following voice samples: Car, boot and meet.

The matching process: When a speaker tries to access the system, he will be asked to say a certain word. This word passes through all the stages of recognition system; so it will be recorded, filtered, detected and analyzed. Then it will be matched with the values of formants which were stored in the database.

Table 1: The results of the system

Corpus	Accuracy (%)
Male users	93.80
Female users	95.60
All users	94.7

At the beginning we only recorded two voices so when we tried to test a speaker the result appeared in a few seconds. But when we increased the numbers of speakers stored in database, it took us more time to obtain the result which makes a sensation because the search will extend to more speakers.

The result of this system when it was applied on our corpus (male and female users) is shown in Table 1.

CONCLUSION

In this study, the autoregressive model used to recognize the identity of the speaker depending on the voice frequencies.

The proposed system is an efficient way to ensure security. The using of the autoregressive model with formants analysis make the recognizing process much faster than using the neural networks method, because it compares numbers to numbers which easier than comparing templates with templates.

The proposed system can take a large number of authorized people, it only reduce the speed of search. Finally, the accuracy of this system is 94.7%.

RECOMMENDATION

There are many directions are recommended to enhance the Speaker Identification System using autoregressive model, such as: improving the accuracy of the system by applying the wavelet transform on the voice samples to compress its size. Make a comparison study between the accuracy of this system when it is applied on male voice with the accuracy of the same system when it is applied on female system. Improve the accuracy by taking a specific duration of the signal which is eliminating the unnecessary parts of the signal.

REFERENCES

Aronowitz, H. and D. Burshtein, 2007. Efficient speaker recognition using approximated cross entropy (ace). IEEE T. Audio, Speech Language Proc., 15(1): 196-205.

Clarkson, T., C. Christodoulou, Y. Guan, D. Gorse, D. Romano-Critchley and J. Taylor, 2001. Speaker identification for security systems using reinforcement-trained pram neural network architectures. IEEE transactions on systems man and cybernetics-part c: Applications and reviews, Vol. 31.

Doddington, G., M. Przybocki, A. Martin and D.Reynolds, 2000.The speaker recognition evaluation overview, methodology, systems, results, perspective. Speech Communication, Vol. 31.

Grimaldi, M. and F. Cummins, 2008. Speaker identification using instantaneous frequencies. IEEE T. Audio, Speech Language Proc., 16(6): 1097-1111.

Hetingl, Y.Y., E. Erzin and A. Tekalp, 2006. Discriminative analysis of lip motion features for speaker identification and speech-reading. IEEE T. Image Proc., 115: 279-289.

Kinnunen, T., E. Karpov and P. Frnti, 2006. Real-time speaker identification and verification. IEEE Trans. Audio Speech Language Proc., 14: 277-288.

Melin, P., J. Urias, D. Solano, M. Soto, M. Lopez and O. Castillo, 2006. Voice recognition with neural networks, type-2 fuzzy logic and genetic algorithms. Eng. Lett., 13: 108-116.

Phan, F., E. Mitheli-Tzanakoul and S. Sideman, 2000. Speaker identification using neural networks and wavelets. IEEE Eng. Med. Biol., 19(1).

Rabiner, L. and B. Juang, 1993. Fundamentals of Speech Recognition. Simon and Schuster Company, PTR Prentice-Hall, Inc.

Reynolds, D., 1995. Speaker identification and verification using gaussian mixture speaker model. Speech Communication, 17: 91-108.

Stremler, F., 1990. Introduction to Communication Systems. Addison-Wesley Publishing company Inc.

Wang, J., C. Yang, J. Wang and H. Lee, 2007. Robust speaker identification and verification. IEEE Computational Intelligence Magazine, 12: 52-59.

Wang, N., P. Ching, N. Zheng and T. Lee, 2011. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE T. Audio, Speech Language Proc., 19: 196-205.

Yuo, K., T. Hwang and H. Wang, 2005. Combination of autocorrelation-based features and projection measure technique for speaker identification. IEEE T. Speech Audio Proc., Vol. 13.