International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 3, No. 2, 2016, pp. 6-12.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Persian Language telephone and Microphone Speaker identification using neural networks

Majid Ranjkesh, Mirmansour Ziabari

MSc, Islamic Azad University, Rash Branch, Iran.

PhD, Islamic Azad University, Rash Branch, Iran.

Abstract

Speaker identification has attracted much attention as an identity identification method in security systems. Identity identification based on telephone conversations may be very useful in guaranteeing personal information security. Nowadays, virtual neural network algorithms are highly used in machine learning. These networks are trained by labeled data and they perform classification. In this study, speaker identification systems based on Mel Frequency Cepstral Coefficients are designed for Persian language speakers. The results of two tests have been reported: Cepstral Coefficients and neural networks are used in the first test to detect microphone speaker; the data base used in this test is the authenticated data of Farsdat, and we used 7 speakers in them. The second test performs the telephone speaker identification for 10 Persian language speakers using lDC data; this test used the Persian language speakers' part of this database. The best results in microphone mode are obtained having 15 minutes training data and the proposed system achieved the identification rate of 100 percent. The best identification result with 3 minutes of training data achieved 82 percent.

Keywords: Speaker identification, virtual neural network, MFCC.

1. Introduction

Speaker identification is detecting one's identity by his voice. Like many human communication means, speech is easily available, it can be simply recorded by means of several electronic devices, and it can be transmitted through communication channels. Due to several applications of speech communication in internet such as Google talk, Skype and Google speech search, its importance is obvious. Moreover, this issue may be attractive and interesting for the users using them. Speaker identification techniques have been utilized easily in three branches of security, monitoring, and forensic [1].

The main application of security systems based on speaker identification is their usage in electronic business, where we need to authenticate the identity if the person we are talking to via telephone. On the other hand, telephone-based services are usually multi user, and the necessity of finding specific-subject or specific-speaker conversations highlights the role of speaker identification systems, due to the highly populated multi-user internet environments and existence of a large archive of different users' speeches.

Speaker identification may be categorized to speaker identification [2] and speaker verification [3]. In speaker identification, each speaker is equivalent to a class. In this situation, speaker identification is a multi-class classification, in which an unknown test speech is attributed to a specific class. Speaker verification is verifying the identity of the speaker claiming to have a specific identity. This is in fact a binary classification in which the claimant is accepted or rejected based on statistical correlation of the test speech and the target speaker's speech. A speaker identification system's framework is presented in Fig.(1). In the following, we discuss each section in brief.

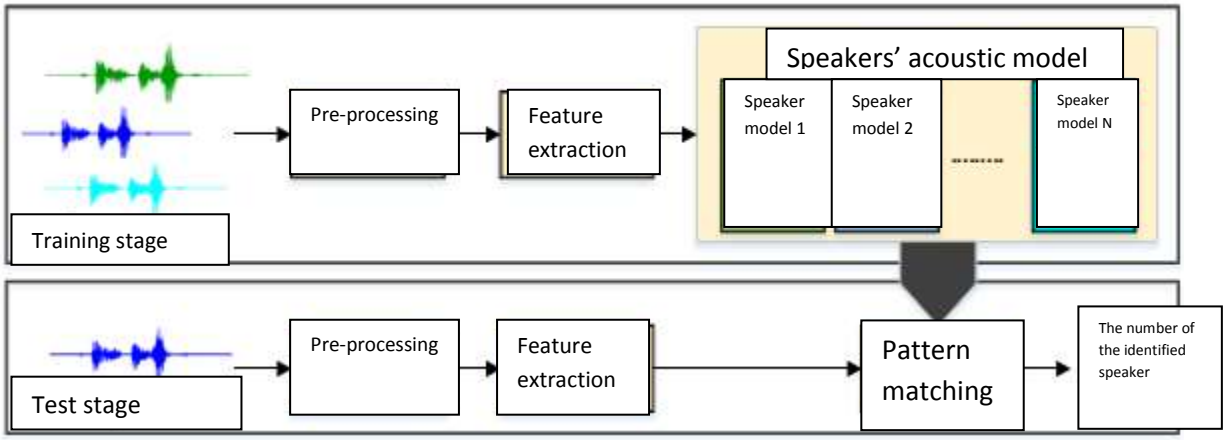


Fig1. Block diagram of a speaker identification system

The first step is pre-processing. This stage is like recording a voice signal for further processing. The recorded digital signal is usually pre-emphasized using a high-pass filter; this would amplify the high frequency components which have been attenuated through human speech generation mechanism. In some speaker identification applications, voiced activity detection is used to separate the speech part of a voice signal.

The next step is feature extraction. This part is the most important section in a speech processing system. Here, a voice signal stream is transmitted to a set of parameters in which the speaker's distinction features are emphasized and this removes the extra information. A large number of extracted parameters for speaker identification can be classified in classes of spectral, resource, prosodic and high-level features. Proper feature selection for speaker identification is done based on the specific circumstances. A proper feature is expected to have high inter-speaker diversity, low intra-speaker variations, robustness against noise and channel variations, immunity to circumstance

and speaker health, and it should be easily separable. Apart from this, features should have small dimensions,in order not to need large training data. Nevertheless, short time spectral features (like MFCC) [4] are more useful in speaker identification, due to their high identification accuracy and immediate extraction, but they are at the risk of destruction by noise [5].

Next step is building an independent acoustic model for each speaker. In this step, different modeling methods are used to model the distribution of extracted features from each speaker. Two stages of feature extraction and modeling constitute the system training stage, in which the speaker is introduced to the system. The aim of this part of the system is building a specific pattern for each speaker. Speaker models are divided into two classes of generative and distinctive. Generative models estimate each speaker's distribution separately, as HMM[7], VQ[8], GMM[8], while distinctive models are trained using the difference between different speakers, like svm[9], NN[10]. Some studies work on merging these two methods to improve the results[11,12].

After training, the system needs pattern matching to evaluate a test voice file. In this step, an unknown speech (test input) is classified based on statistical correlation with one of the classes. Pattern matching and classification step is the evaluation and test step in which the speaker identification system is evaluated based on classification efficiency. Pattern matching depends on the nature of the speaker's acoustic models. In generative statistical models, matching is examined based on Log-likelihood, which can be a simple distance for parametric mode, as Euclidean distance in VQ. For distinctive models, this can be equivalent to distance from the border of two classes, like SVM, and it would be the difference between the real class and the estimated class, like NN. The decision is made based on its given rank and the test signal is attributed to the class having the highest rank. In this study, Cepstral coefficients by Mel weighting are used to extract features and virtual neural networks are used for classification. In the following, we discuss the procedure of each processing.

2. Feature extraction

Speech signal of each speaker is passed through a voiced activity detection system. In this paper, we used Sohn detector [13]. This detector works based on the energy of each speech frame, and an example of this system's performance is shown in Fig.2.

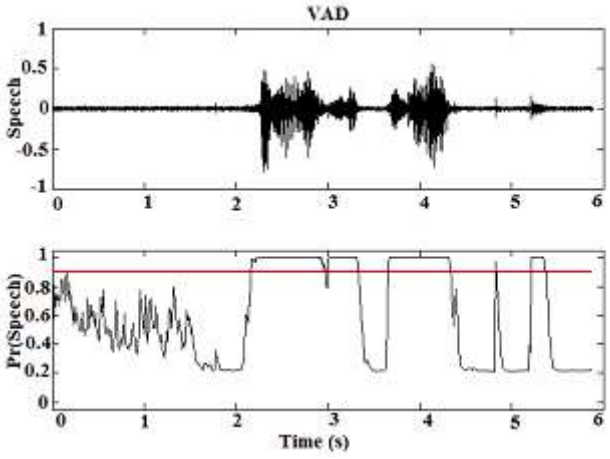


Fig.2- Performing detection on number "Five", pronounced by the speaker F02.

This would omit the silence parts of each voice file, and reduce the amount of required processing as a result. Moreover, this would improve the accuracy of the system by just evaluating the speech parts. Cepstralcoefficients (MFCC) are features highly used in speech identification and speaker identification systems. This method has been

first introduced in 1980; from that time, it has been state of the art. In this paper, we implement these coefficients in the proposed speech identification system. The feature extraction method is illustrated.

At first, we pre-emphasize the signal. Pre-emphasizing amplifies the higher frequencies by passing the signal through a high-pass filter. The voice signal is always variable, but we can assume the statistical information of this signal stationary in a short time interval(10ms-30ms); this is done by framing the signal in 20ms intervals. After that, we use a Hamming window to remove the side lobes made by the rectangular window. We take the Fourier transform of the signal multiplied to the Hamming window and separate its amplitude for further processing. After this step, there are still a lot of extra information which are not useful in identification. To solve this problem, we use a characteristic of human ear. The cochlear cannot distinguish near frequencies which are in a frequency band; as a result, we assume a group of frequencies having the same value and calculate the energy in a frequency band. This is done by Mel filter. In this filter bank, the first filter is extremely narrow and shows the value of energy around the frequency of 0 Hz. By increasing the frequency, filters become wider the frequency diversity reduces, and we only consider energy in each frequency band in our calculations (Fig.3).

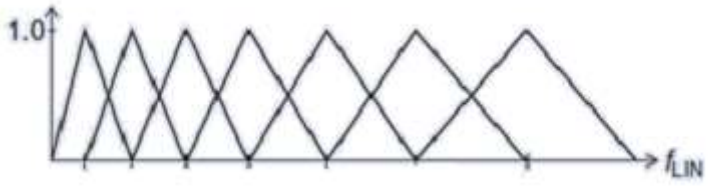


Fig.3- Mel filters

The weighted signal is transformed from frequency domain to Cepstral domain by applying logarithm. Finally, DCT makes the energies decorrelate, but only 12 DCT coefficients are kept. Large coefficients in DCT shows large variations in filter bank energies, and this information would be removed since it would reduce the efficiency of the identification system and by reducing it, we are improving the system. In this work, we use the energy of each frame as another feature. MFCC features show each frame's spectral power curve, but speech signal in transient domain may contain beneficial information as well. Under these circumstances, the first derivative and the second derivative of each coefficient are considered as features.

3. Virtual neural network

Neural network is an information processing system having similar characteristics to the natural neural network of human brain. In fact, it is the extended mathematical model of human brain based on neural biology. In neural networks, information processing is done in simple, numerous components, called Neurons. Information is transferred between network Neurons via their connection or linkage. Each connection or linkage has its own weight, which is multiplied to transmitted signals from that connection in conventional neural networks. Each neuron applies an activation function on its inputs to generate its output signal. This function usually has to be differentiable.

4. Evaluation criterion

In a multi-class classification system using virtual neural network, each output shows a class$i \in \{1, \dots, K\}$. the ith output shows the probability P(i|o), which is the probability that the test vector O belongs to the ith class. The efficiency of speaker identification systems is based on the identification rate criterion, which states the proportion of correct identifications of the system to the total identifications.

5. Implementation and results

One of the most important tools for designing and building intelligent speaker identification systems is providing an extensive set of speech data to train voice and sound acoustic models of speakers. In this study, we used two data sets that we explain in the following.

5.1. Farsdat

Farsdat data is a set of phrases and sentences stated by Persian language speakers across the country.

This database is known as a standard Persian language speech database inside and outside the country, and it is used for training intelligent speech identification systems. In this study, we use the voice of 7 male speakers of this database to train and evaluate the proposed system. The sampling frequency of this database is 22050Hz.

5.2. LDC

Telephone speech, which has been used in this study, has been collected from LDC dataset. This dataset includes multiple languages, including Persian. We will work on 10 speakers including 5 men and 5 women. The sampling frequency in this dataset is 8000Hz.

Table 1- speaker identification in different training data amount situations in microphone mode.

	Identification percent				
Training data amount (in minutes)	0.5	2	5	10	15
Speaker 1	76.5%	94.2%	96.4%	99.3%	100%
Speaker 2	78.1%	94.6%	97.2	100%	100%
Speaker 3	83.6%	97.1%	99.3%	100%	100%
Speaker 4	67.9%	87%	94.4%	99.8%	100%
Speaker 5	75.4%	91%	93.4%	99.5%	100%
Speaker 6	82.2%	98.3%	99.5%	100%	100%
Speaker 7	72.9%	96.4%	98.1%	100%	100%
Average	76.65%	94.08%	96.9%	99.8%	100%

6. Implementation and results

First, the signal is passed through the voiced activity detection and the purified signal is divided into 20ms frames. From each frame of the signal, we extract 39 features which enter the neural network for classification, using the method previously explained. First, we consider microphone mode tests. For modeling, a three layer neural network has been used; the first layer consists of 117 neurons, the output of 117 neurons are given to 1000 neurons and finally in the last layer, the output of these 1000 neurons are given to the number of classes, which are the speakers. Training data amount for each speaker in different situations have been evaluated and the results are presented in table 1. As obvious, we may obtain an acceptable efficiency percentage by having 5 minutes of training data for each speaker.

Telephone speech encounters more problems in speaker identification due to the limitations applied by communication network. Another limitation we may encounter is having less voice signals for tests, in comparison to microphone mode. As a result, we perform this test at 3 situations of 1, 2 and 3 minutes training data. In order for the system to be reliable in terms of implementation, a "silence" output is considered as well, in which the speaker does not say anything. The neural network structure is the same as the microphone mode. The results of speaker identification in different situations of training data are presented in table 2.

Table 2- results of speaker identification in different training data amount situations in microphone mode.

	Identification percentage		
Training data amount(in minutes)	1	2	3
Speaker 1	65%	74%	87%
Speaker 2	57%	69%	76%
Speaker 3	64%	82%	94%
Speaker 4	48%	53%	68%
Speaker 5	73%	86%	95%
Speaker 6	55%	63%	73%
Speaker 7	68%	74%	81%
Speaker 8	71%	79%	89%
Speaker 9	54%	67%	76%
Speaker 10	49%	58%	64%
Silence	97%	100%	100%
Average	63.72%	73.18%	82.09%

As obvious from the table (2), the system works properly for silence mode; the reason is the great difference between signal coefficients of silence and speech. When using 3 minutes of training signal for each speaker, we obtained better percentage, and we expect to have more acceptable percentage by using more training data amounts.

7. Conclusion

In this paper, the goal was to present a proper system to identify the speaker in microphone and telephone modes by using Cepstral coefficients with Mel weighting and virtual neural system. Tests were performed for different situations of training data to obtain the minimum required data for microphone and telephone modes, for the system to have proper performance. Finally, the proper functioning of the neural network in speaker identification was reported by studying the results.

References

[1] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag.*, no. 95–103, 2009.

[2] S. V. Chougule, and M. S. Chavan, "Effectiveness of LSP features for text independent speaker identification," *Computing, Communication and Security (ICCCS), 2015 International Conference on*, pp. 1-4, 2015.

[3] W. Rao, and M.-W. Mak, "Sparse kernel machines with empirical kernel maps for PLDA speaker verification," *Computer Speech & Language*, 2016.

[4] S. B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 28, no. 4, pp. 357-366, 1980.

[5] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *Speech and Audio Processing, IEEE Transactions on,* vol. 2, no. 4, pp. 639-643, 1994.

[6] T. Kinnunen, J. Saastamoinen, V. Hautamäki, M. Vinni, and P. Fränti, "Comparative evaluation of maximum a posteriori vector quantization and Gaussian mixture models in speaker verification," *Pattern Recognition Letters,* vol. 30, no. 4, pp. 341-347, 2009.

[7] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech communication,* vol. 41, no. 4, pp. 603-623, 2003.

[8] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, and A. Rastrow, "Subspace Gaussian mixture models for speech recognition," *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4330-4333, 2010.

[9] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication,* vol. 66, pp. 130-153, 2015.

[10] K. Daqrouq, and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing,* vol. 27, pp. 231-239, 2015.

[11] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language,* vol. 20, no. 2, pp. 210-229, 2006.

[12] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE,* vol. 13, no. 5, pp. 308-311, 2006.

[13] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE,* vol. 6, no. 1, pp. 1-3, 1999.