Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition

Reda Adjoudj
University of Sidi-Bel-Abbes
59 PUBLICATIONS **191** CITATIONS
SEE PROFILE

Aoued Boukelif
University of Sidi-Bel-Abbes
78 PUBLICATIONS **310** CITATIONS
SEE PROFILE

Artificial Neural Network & Mel-Frequency Cepstrum Coefficients-Based Speaker Recognition

Adjoudj Réda 1*, Boukelif Aoued 2**

- *Evolutionary Engineering and Distributed Information Systems Laboratory,EEDIS,
Computer Science Department, University of Sidi Bel-Abbès, Algeria
Phone/Fax: (213)-48-57 77 50 ,*
AdjReda@yahoo.fr

** *Digital Signal processing laboratory, Electronic Department,
University of Sidi Bel-Abbès, Algeria
Phone:(213)-48-57 82 81 ,*
aboukelif@yahoo.fr

Abstract: Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

This document demonstrates how a speaker recognition system can be designed by artificial neural network using Mel-Frequency Cepstrum Coefficients of voice signal. Note that the training process did not consist of a single call to a training function. Instead, the network was trained several times on various input ideal and noisy signals coded by Mel-Frequency Cepstrum Coefficients, the signals which contents voices.

In this case training a network on different sets of noisy signals forced the network to learn how to deal with noise, a common problem in the real world.

Key words: Artificial Neural Network, Biometric, Mel-Frequency Cepstrum Coefficients, Pattern Recognition, signal processing.

1 Introduction

Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker(Ostendorf & al., 1996),(Minh & al., 2003),(Carrillo, 2003).

Speaker recognition methods can also be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, on the other hand, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc(Carrillo, 2003).

All technologies of speaker recognition , identification and verification, text-independent and text-dependent, each have its own advantages and disadvantages and may requires different treatments and techniques. The choice of which technology to use is application-specific (Carrillo, 2003),(Hosom, & al., 1999),(Fokou, & al., 2002),(Patel & al., 1999).

1.1 How speaker recognition works

At the highest level, all speaker recognition systems contain two main modules (Ostendorf & al., 1996),(Minh & al., 2003),(Cornaz, & al., 2003): feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers. We will discuss each module in detail in later sections.

Although voice authentication appears to be an easy authentication method in both how it is implemented and how it is used, there are some user influences that must be addressed (Carrillo, 2003),(Davis & al., 1980):

- Colds. If the user has a cold which affects his or her voice that will have an effect on the acceptance of the voice-scanning device. Any major difference in the sound of the voice may cause the voice-scanning device to react in a negative way, causing the system to reject the user.

- Expression and volume. If a person is trying to speak with expressions on their face (i.e. smiling at the same time) their voice will sound different. The user of the device must also be able to speak loudly and clearly in order to obtain accurate results.
- Misspoken or misread prompted phrases. If the user is required to authenticate by speaking a prompted phrase and they mispronounce the phrase, they will be rejected by the system.
- Previous user activity may have an impact on the outcome of the voice scanning device. For example, if the user is out of breath and is unable to speak well.
- Background noises will interfere with the user who is trying to authenticate to the device. The environment in which the user is authenticating to the device must be free of any major background noise.

2. Proposed design

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called pattern recognition. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called patterns and in our case are sequences of acoustic vectors called matrix codes or voiceprints that are extracted from an input speech using the techniques described in the later section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching.

This document demonstrates how speaker recognition can be done with a backpropagation artificial neural network as matching process, but above the acoustic characteristics of the voice signal was converted into a matrix code, also called voiceprint, by using Mel-Frequency Cepstrum Coefficients (MFCC-function)(Minh & al., 2003),(Cornaz, & al., 2003).

3. Problem statement

An artificial neural network is to be designed and trained to recognize the voice signal code of the database that is actually used (Howard & al., 1998),(Adjoudj, & al., 2004a),(Adjoudj, & al., 2004b),(Adjoudj, & al., 2004c). An imaging system that converts each voice signal in voiceprint or voice signal matrix code by using a Mel-Frequency Cepstrum Coefficients (MFCC-function) centered in the system's field of acoustic is available. The result is that each voice signal is represented as a matrix of 160 reals values. (voiceprint or voice signal matrix code size ~ 20 x 8).

3.1 Creating a voice signal code

The signal of a voice is first processed by software that convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing.

The speech signal is a slowly timed varying signal (it is called quasi-stationary). An example of speech signal is shown in Figure 2. When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal.

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Gaussian mixture models (GMM)(Mami,2003), Mel-Frequency Cepstrum Coefficients (MFCC), and others[14].

MFCC is perhaps the best known and most popular, and these will be used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech(to obtain voiceprint or voice signal matrix code). This is expressed in the mel-frequency scale, which is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The process of computing MFCCs is described in more detail in (Minh & al., 2003),(Cornaz, & al., 2003).

A block diagram of the structure of an MFCC processor is given in Figure 1. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to mentioned variations.

The voice signal matrix is immediately encrypted to eliminate the possibility of identity theft and to maximize security. For example, here is the voice signal from database(Cornaz, & al., 2003) and the voiceprint matrix of this voice signal (figure 2) :

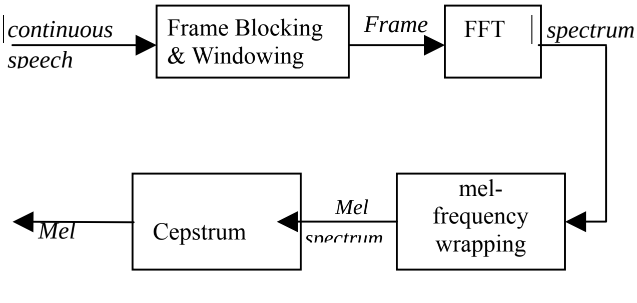

Figure 1. *Block diagram of the MFCC processor*

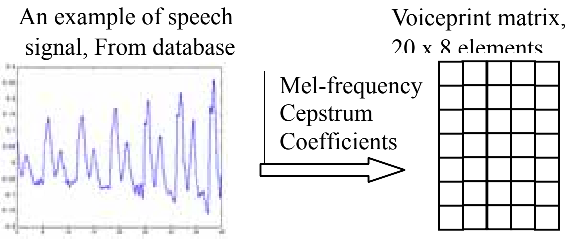
An example of speech signal, From database
Voiceprint matrix, 20 x 8 elements
Mel-frequency Cepstrum Coefficients

Figure 2. *Voiceprint*

Perfect classification of N ideal input voiceprint matrix of voice signal is required, and reasonably accurate classification of speech waveform (N is equivalent to a number of distinguish class of speaker in each database). The N 160-element input voiceprint matrix of voice signals are defined as a matrix of input matrixes (voiceprint matrix size ~ 20 x 8). The target vectors are also defined with a variable called targets. Each target vector is a N-element vector with a 1 in the position of the voiceprint it represents, and 0's everywhere else. For example, the voiceprint number one is to be represented by a 1 in the first element (as this example is the first voiceprint of the database), and 0's in elements two through N.

4. Voiceprint recognition

In less than a few seconds, even on a database of millions of records, the voiceprint generated from a voice signal is compared to previously enrolled ones to see if it matches any of them. The decision threshold is automatically adjusted for the size of the search database to ensure that no false matches occur even when huge numbers of voiceprints are being compared with the live one. Some of the bits in a voiceprint signify if some data is corrupted (e.g. the speaker has a cold), so that it does not influence the process, and only valid data is compared. Decision thresholds take account of the amount of acoustic voice signal data, and the matching operation compensates for any tilt of the speech waveform. A key advantage of speaker recognition is its ability to perform identification using a one-to-all search of a database, with no limitation on the number of voiceprint records and no requirement for a user first to claim an identity, for example with a card. For our method we use a artificial neural network for matching and perform recognition using a one-to-all search of a database, which is described in more detail next.

4.1 Neural network

The network will receive the 160 real values as a 160-element input voiceprint matrix of voice signal (voiceprint matrix size ~ 20 x 8, see figure3). It will then be required to identify the speaker by responding with a N-element output vector (for more detail about N see above). The N elements of the output vector each represent a speaker(see figure3).

To operate correctly the network should respond with a 1 in the position of the speaker being presented to the network. All other values in the output vector should be 0.

4.2. Architecture of neural network

The neural network needs 160 inputs and N neurons in its output layer to identify the speaker. The network is a two-layer tang-sigmoid/tang-sigmoid network like use in (Howard & al., 1998),(Adjoudj, & al., 2004a),(Adjoudj, & al., 2004b),(Adjoudj, & al., 2004c). The tang-sigmoid transfer function was picked because its output range (0 to 1) is perfect for learning to output Boolean values (see figure3).

The hidden layer has 150 neurons. If the network has trouble learning, then neurons can be added to this layer (Howard & al., 1998),(Adjoudj, & al., 2004a),(Adjoudj, & al., 2004b),(Adjoudj, & al., 2004c). The network is trained to output a 1 in the correct position of the output vector and to fill the rest of the output vector with 0's. However, noisy input speech signals may result in the network not creating perfect 1's and 0's. After the network has been trained the output will be passed through the competitive transfer function . This function makes sure that the output corresponding to the speaker most like the noisy input speech signal takes on a value of 1 and all others have a value of 0. The result of this post-processing is the output that is actually used (Howard & al., 1998),(Adjoudj, & al., 2004a),(Adjoudj, & al., 2004b),(Adjoudj, & al., 2004c).

4.3 Training

To create a neural network that can handle input speech signal (voiceprint) it is best to train the network on ideal speech signals. To do this the network will first be trained on ideal speech wave forms (voiceprint) until it has a low sum-squared error.

Then the network will be trained on 10 sets of ideal and noisy speech signals. The network is trained on two copies of the noise-free database at the same time as it is trained on noisy speech signals. The two copies of the noise-free database are used to maintain the network's ability to classify ideal input speech signals.

Unfortunately, after the training described above the network may have learned to classify some difficult noisy speech signals of speaker at the expense of properly classifying a speaker (voiceprint of speaker).

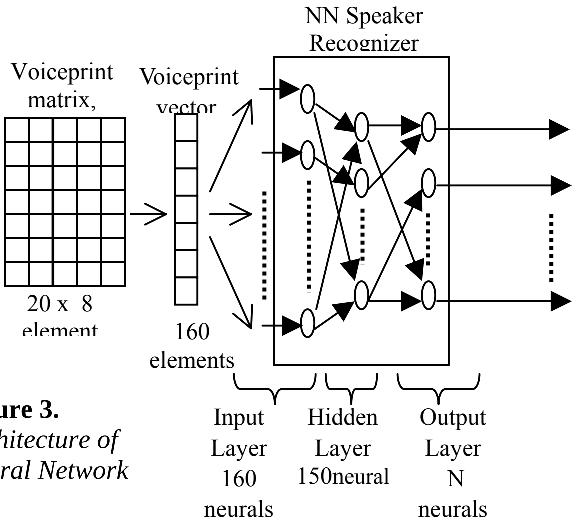

Figure 3. *Architecture of Neural Network*

Therefore, the network will again be trained on just ideal speech signals of speaker. This ensures that the network will respond perfectly when presented with an ideal speech signal. All training is done using backpropagation with both adaptive learning rate and momentum.

4.3.1. Training with artificial neural network "System 1 ". The speaker recognition system is initially trained with artificial neural network for a maximum of 5000 epochs or until the network sum-squared error falls below 0.001 (see a figure 4).

4.3.2. Training without artificial neural network "System 2 ". Now, we created a new speaker recognition system, that trained without artificial neural network but we use directly a well-know algorithm, namely LBG algorithm (Minh & al., 2003),(Cornaz, & al., 2003),[15], for clustering a set of training vectors or signals into a set of codebook or voiceprint vectors, and finaly, we compare the results of the two speaker recognition systems ("System1" & "System2").

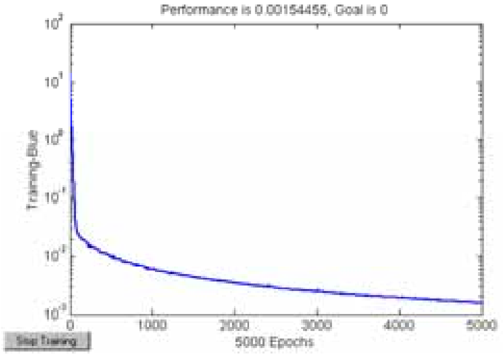

Figure 4. Training of neural net, "system1".

5. Experiment and Test

To evaluate the performance of the proposed method, we collected a large number of speech signal of different speakers at different moments using an appropriate sensor to form our own Database. The database includes 294 speech signals (N=294) from 142 different subjects. The speech signals are acquired during different sessions and with different kinds of noise, which provides a challenge to our system.

To test and compare the two systems, an speech signal with noise can be created and presented to the two systems, and a two other databases of speech signals can be downloaded and presented to the our system as data-sets, which the first one named CMU Speaker Database (Patel,1996), To the best of our knowledge, this is the largest speech signal of speaker database available in the public domain. The subjects consist of 203 members of the CMU university (N=203, for more detail about N see above).the second one named ASR Database(Minh & al., 2003), with a 60 subjects from Audio Visual Communications Laboratory Swiss Federal Institute of Technology (then in our case N=60).

Finally, these three databases of voice can be presented to the two systems as data-sets. Table1 shows the results of recognition rate and performance of a proposed two systems (see table 1 for more examples of databases with different kind of speaker's voice).

6. System performance

The reliability of the pattern recognition system is measured by testing the system with hundreds of input voices of speakers with varying quantities of noise i-e. synthetic noise. We test the two systems at various noise levels and then graphs the percentage of each system errors versus noise. Noise with mean of 0 and standard deviation from 0 to 100 are added to input voices. At each noise level 100 presentations of different noisy versions of each voice of speaker are made and the two system's outputs are calculated. The number of erroneous classifications are then added and percentages are obtained (see figure 5).

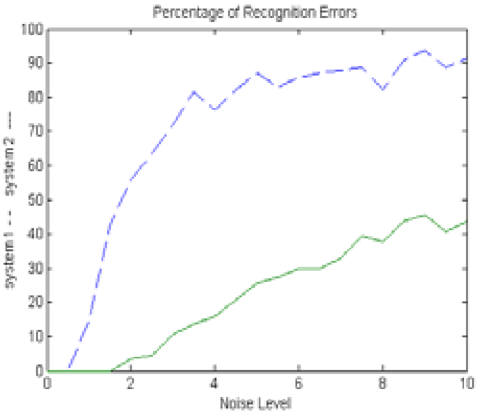

Figure 5. The reliability of the two systems .

Speaker Recognition System	Matching Algorithm As “Pattern recognition Module”	Time of Training	Recognition Rate (RR)					
			ASR Database “N = 60”		Our own database “N = 142”		CMU Data-set “N = 203”	
			Train Set	Test Set	Train Set	Test Set	Train Set	Test Set
“ System 1”	Artificial Neural Network	40 min 39 sec	95.55%	70.91%	96.83%	78.57%	98.92%	90.66%
“ System2 ”	LBG algorithm “ vector quantizer design Algorithm“	2 hours 15 min 45 sec	94.33%	80.01%	92.87%	85.71%	92.95%	69.33%

Table 1. *Recognition results on the speaker's voice database (of a three data-sets) on a PC with 1.7 GHz CPU. RR: Recognition Rate.*

The dashed line (blue dashed line) on the graph shows the reliability for the first system ‘’system1’’ (with artificial neural network as pattern recognition) trained with and without noise. The reliability of the second system ‘’system2’’ when it had only been trained without artificial neural network but it use directly a LBG algorithm as pattern recognition is shown with a solid line (green line) .

Thus, training the two systems on noisy input signal waveforms of speaker greatly reduced their errors when they had to classify or to recognize noisy signal waveforms of speaker. The two systems did not make any errors for voices with noise of mean 0.00 or 0.50 . When noise of mean is greater than 0.50 was added to the voices first system began to make errors but the second system began to make errors until 1.50.

If a higher accuracy is needed the two systems could be trained for a longer time or retrained with more neural in their hidden layers respectively for the first system.

Finally, the two systems could be trained on input speech signals with greater amounts of noise if greater reliability were needed for higher levels of noise.

7. Conclusion

Speaker recognition is challenging problems and there is still a lot of work that needs to be done in this area. Over the past ten years, speaker recognition has received substantial attention from researchers in biometrics, pattern recognition, signal processing, and cognitive psychology communities. This common interest in speaker recognition technology among researchers working in diverse fields is motivated both by the remarkable ability to recognize people and by the increased attention being devoted to security applications.

Applications of speaker recognition can be found in security, multimedia, and entertainment domains. We have demonstrated how a speaker recognition system can be designed by artificial neural network using Mel-Frequency Cepstrum Coefficients matrix of voice as inputs (for capturing the phonetically important characteristics of speech, for optimising the size of signal voice and for saving training time of neural network). Note that the training process did not consist of a single call to a training function. Instead, the network was trained several times on various input ideal and noisy signals of voices. In this case training a network on different sets of noisy signals forced the network to learn how to deal with noise, a common problem in the real world.

8. References

(Adjoudj, & al., 2004a) R. Adjoudj, A. Boukelif, “Artificial Neural Network & Multilevel 2-D Wavelet Decomposition Code-Based Iris Recognition", *the International Conference on Computing, Communications and Control Technologies,* Paper No. T645XC, CCCT 2004, Austin (Texas), USA, on August 14-17, 2004.

(Adjoudj, & al., 2004b) R. Adjoudj, A. Boukelif, " Detection et reconnaissance des visages basée sur les reseaux de neurones artificiels", *Communication Science & Technologie Journal,* Ecole Normal Supérieur de l'Enseignement Technique d'Oran, Algeria , 2004.

(Adjoudj, & al., 2004c) R. Adjoudj, A. Boukelif, “Artificial neural network-based face recognition” , *International Symposium on Control , Communications and Signal Processing,* Paper No 1107, ISCCSP 2004 , Hammamet, Tunisia, March 2004.

(Adjoudj, & al., 2003) R. Adjoudj, A. Boukelif, " Detection and recognition of the faces based on the artificial neural networks", *RIST Journal*, Vol, 13 n°02, pp.93-108 , Algeria , 2003.

(Boite & al., 1999) R.Boite, H.Boulard, T.Dutoit, J.Hancq & H.Leich, « *Traitement de la parole* » Book, Collection Electricité, Presses Polytechniques et Universitaires Romandes, 1999.

(Carrillo, 2003)C.M. Carrillo, ‘’continuous biometric authentication for authorized aircraft personnel ”, master thesis, a proposed design, naval postgraduate school monterey, california, June 2003.

(Cornaz, & al., 2003) C.Cornaz, U. Hunkeler & V.Velisavljevic “An automatic speaker recognition system”, Digital Signal Processing Laboratory, Ecole Polytechnique Federale de Lausanne, Switzerland, February 2003.

(Davis & al., 1980)S. Davis & P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing 28(4), Mel-frequency cepstral coefficients (MFCC)*:357-366, 1980.

(Fokou, & al., 2002)A.Fokou, S.Henaff, M.Lagacherie & P. Rouget, “ Modest-encoding AlgoRithm with Vocal Identification”, MARVIN project, EPITA, France, April 2002.

(Hosom, & al., 1999)J.P Hosom, R.Cole & M.Fanty, “Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding”, NSF-Graduate Research Traineeships project, Center for Spoken Language Understanding (cslu), Oregon Graduate Institute of Science and Technology , USA, July 1999.

(Howard & al., 1998) Howard Demuth, Mark Beale, “*Neural Network Toolbox User’s Guide*” For Use with MATLAB® by The Math Works, Inc.1998.

(Linde & al., 1980)Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, *IEEE Transactions on Communications, Vol. 28*, pp.84-95, 1980.

(Mami,2003)Y. Mami, "Reconnaissance de locuteurs par localisation dans espace de locuteurs de références", PhD thesis, at « Ecole nationale supérieure des télécommunications », Paris, France, October 21,2003.

(Minh & al., 2003) Minh N. Do “An Automatic Speaker Recognition System”, Audio Visual Communications Laboratory Swiss Federal Institute of Technology, Lausanne, Switzerland, February 2003.

(Ostendorf & al., 1996)M.Ostendorf,V. Digilakis & O.A. Kimball, "From HMM's to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," IEEE Transactions on Speech and Audio Processing 5: Demonstrates the existence of a continuum of possible approaches between "segmental" and "frame-based" speech recognition, 360-378, 1996.

(Patel & al., 1999)A.D. Patel, A.Lfqvist, and W.Naito, “*The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem*”, Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, USA, August 1999.

(Patel,1996)A.D.Patel,” A Biological Study of the Relationship between Language and Music”, Ph.D. thesis, CMU University, USA, 1996.