TEXT INDEPENDENT SPEAKER RECOGNITION USING THE MEL FREQUENCY CEPSTRAL COEFFICIENTS AND A NEURAL NETWORK CLASSIFIER

HASSEN SEDDIK, AMEL RAHMOUNI* and MOUNIR SAYADI

CEREP, ESSTT, AV. TAHA HUSSEIN, 1008, TUNIS, TUNISIA
* ECOLE NATIONALE DES SCIENCES INFORMATIQUES, 2010, MANOUBA, TUNISIA
Emails: seddik_hassene@yahoo.fr, Amel Rahmouni@esstt.rnu.tn, mounir.sayadi@ipeim.rnu.tn

ABSTRACT

Modern speaker recognition applications require high accuracy at low complexity and easy calculation. In this paper, we propose a new method of text independent speaker recognition based on the use of the mean of the Mel Frequency Cepstral Coefficients (MFCC) as a Speaker Model. These MFCC are extracted from the speaker phonemes in the pre-segmented speech sentences. A multi-layer neural network trained with the back propagation algorithm is proposed to classify these discriminative models. A study is carried out in order to view these models efficiency. Several experiments are made and show that the proposed method gives a high speaker recognition rate. Furthermore, throw these experiments, a technique is proposed to improve this recognition rate by an appropriate phonemes database selection.

1. INTRODUCTION

Since many years, the two most common and successful approaches for speaker recognition, independently of the pronounced text, are based on modeling the speech by Gaussian Mixture Models, and Hidden Markov Models [14] [12]. These methods are attractive for their phonetic discrimination capacity [7]. In the other hand, the ear model quality was, since along time, put to evidence and attracted the interest of studies for its properties in nervous coding, and voice hearing methods [11]. The acoustics analyses based on the MFCC, which represent the ear model [1], has proved good results in speaker recognition especially when a high number of coefficient is used [7]. Furthermore, it's considered the most successful speaker recognition system when confronted to different variations such as: prosodi, intonation, noise [11]. It executes also the task of filtering, modeling, processing, decoding, phonemes or words and languages distinction. Basing on the features acoustic decoding, the speaker is identified by mean of a neural-psycologic function with a cerebral distinctive process [11].

Because of the capacity of the ear model to operate in a separated mode, we can some times recognize a person from he's voice without understanding what he is speaking about [5]. Basing on these facts, in the present work, a new method is proposed to improve the speaker recognition rate and to give more accuracy to the characterization process. This approach, is performed using a phonetic decomposition of the incoming speech.

We use the MFCC [10] extracted from the speaker phonemes as a discriminative features. The text independent speaker recognition is done by classifying these features by A multi-layer neural network. In order to determine the optimum neural network and to achieve the best recognition, several experiments are carried out.

2. DATABASE COMPOSITION AND PHONEMES EXTRACTION

The voice signal is presented as different sentences for a defined number of speakers composing our references belonging to the TIMIT Data-Base. A signal pre-processing is applied. It consists on a pre-emphasis filter to equalize the accurate, always more weak than the graves [1]. A Hamming Window is applied on each bloc in order to decrease the edge effects due to the windows cutting. A Fast Fourier Transform is applied on the treated signal and smoothed by a series of triangular filters distributed on a Mel Scale. The MFCC are then calculated. The scale Mel is given by $M=\frac{1000}{\log 2}\log\left(1+\frac{f}{1000}\right)$ [3], where f notes the frequency.

Twenty speakers compose the database. Every speaker has different recorded sentences. The sentences are segmented in phonemes in order to compose both training and test databases. Many phonemes can be repeated several times in the same sentence.

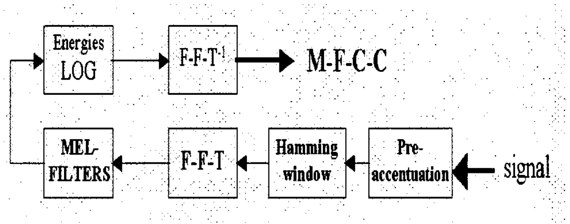

Figure 1: MFCC extraction.

0-7803-8379-6/04/$20.00 ©2004 IEEE.

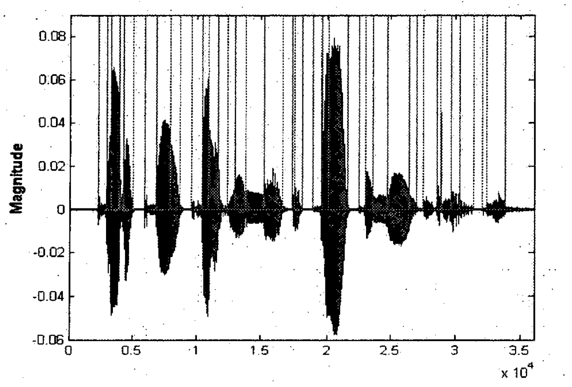

Figure 2: The phonemes limits on a segmented sentence.

Figure 2 gives an example of phonemes limits on a segmented sentence. The phoneme model can characterize a voice speaker and then discriminates two different speakers.

Firstly, a phonemes database is created using 10 sentences for every speaker: 3 learning sentences (Table 1) and 7 sentences used for test. To build a learning database, a set of 48 kinds of phonems are extracted for the three learning senteces. For each phonem kind, we select 5 examples. A learning data-base of 4800 phonems (240 for each one of the 20 spearkers) is then otained.To test the neural network after the traing phase, a second database is created by the phonemes extracted from the 7 sentences kept for test. For each phonem kind, 10 examples are used. A test database of 9600 phonems is then otained.

For every speaker, the phonemes are collected and sampled at fs=16 KHz. They are then filtered using a filterbank containing p filters given by p=floor(3*log(fs)) i.e. 29. The log power outputs of the filter bank were transformed into twelve MFCC values [2].

Every phoneme gives one or more rows of MFCC coefficients, depending on its length. All the phonemes are used with their different lengths without any normalization.[4] These coefficients are arranged successively in a matrix of size $L.C$, with "L" caracterizes the number of lines that are equal to the number of phonemes frames i.e. MFCC extracted vectors, and "C" characterizes the number of columns or number of MFCC extracted i.e. twelve. The mean of the columns matrix gives twelve parameters vectors representing the model that we call Speaker Model MFCC (SMMFCC). Basing on the fact that the cerebral speaker recognition process is based on the processed voice signal features in the ear model, the arranged mean of the matrix columns will impose a loss of information about the evolution of the MFCC in the phoneme, but it will preserve a mean magnitude of these indexed coefficients (from 1 to 12). These twelve coefficients will be used as inputs of the neural network used for classification.

Sentence 1	"She had your dark suit in greasy wash water all year"
Sentence 2	"Don't ask me to carry an oily rag like that"
Sentence 3	"A sailboat may have a bone in her teeth one minute and lie becalmed the next"

Table.1 : List of the three learning sentences.

SH	**IX**	**HV**	**EH**	**DCL**	**JH**
IH	**D**	**AH**	**KCL**	**K**	**S**
UX	**Q**	**UH**	**EN**	**GCL**	**G**
R	**W**	**AO**	**EPI**	**DX**	**AXR**
L	**Y**	**N**	**AE**	**CH**	**M**
OY	**AX**	**DH**	**TCL**	**IY**	**V**
F	**T**	**K**	**PCL**	**OW**	**HH**
AXH	**EY**	**BCL**	**B**	**AY**	**AA**

Table.2 : The 48 phonemes extracted from the three learning sentences.

Also, We found that, as seen in Figure 3, for similar phonemes extracted from different sentences and spoken by the same person, the correlation of the SMMFCC coefficients is high and the models represented by these coefficients are too close in looks.

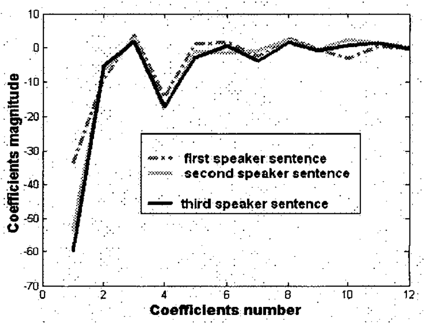

Figure 3 : The look of 3 phonemes, (SH phoneme) extracted from different sentences pronounced by the same speaker.

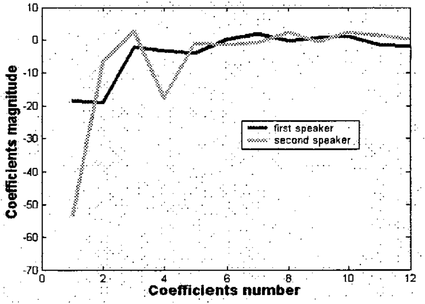

Figure 4: The look of the same phonemes (SH), extracted from two different speakers speech.

In the other hand, if we represent the same phonemes (SH) extracted from the speech of different speakers, we find that the they are different in looks (Figure 4). Then, this phoneme models characterizes the phoneme itself and the voice speaker in the same time, because of it's similarity for the same speaker and its difference from speaker to other. The correlation of the same speakers phonemes, and the difference in look in the case of the same phonemes for different speakers are then exploited to discriminate the speaker.

3. RESULTS OF CLASSIFICATION WITH A NEURAL NETWORK

To recognize a speaker, the SMMFCC extracted from its phonemes are used as input vectors to a multilayer neural network. The network is trained using the gradient descent back-propagation algorithm [9] with training database (4800 phonems) and tested with the test database (9600 phonemes). The network weights were updated on each presentation of a feature vector. The set of training examples is changed at each iteration and their order is randomly chosen. For each speaker, we define the Recognition Rate (RR) as the ratio of the number of positive tests to the total number of tests. In order to determine the optimum neural network to achieve a maximal Recognition Rate, we carried out several experiments using various architectures, that is: various training coefficients and various numbers of neurons in each layer [6]. We used two hidden layers. The initial random values of the weights were set between –1 and 1. A smoothed threshold function given by:
$f(y)$= [1-exp($-ay$)]/ [1+exp($-ay$)] [9] is applied to the output of each neuron. "a" notes the sigmoid threshold.

Four experiments are carried out using two different neural network kinds. The first three experiments are based on training the network with different phoneme's database sizes. Every speaker is treated by an individual neural network. The task of this network is to decide if the phoneme's SMMFCC belongs to this speaker or not.

In the test phase, the phonemes composing the test sentences pronounced by each speaker are introduced successively to the corresponding network in order to identify the phoneme learned belonging to the desired speaker, and check the network behavior when confronted with a never learned phonemes. This allows to carry out a technique for a best phonemes database selection.

In these experiments, because of the high number of features introduces (240 phonemes for each one of the 20 speakers), with 12 coefficients in every phoneme, a binary output can't be used to characterizes both phonemes and speakers.

The neural network output layer will be composed by only one decision neuron (output = 0.5) with a fixed threshold band. All the outputs into the threshold band are considered identified speakers. Those out of the fixed threshold band are considered false. In the fourth experiment, we will test the results of this method when only vowels are used. The neural network output layer is composed by a binary decision.

Experiment 1: Small phonemes database

We choose 5 of the phonemes kinds, in the database shown in Table 2 (SH-IX-HV-EH-DCL). For each phoneme, 5 examples are selected for training. The 12 SMMFCC are extracted and introduces to the neural network structured as 12 neurons in the first layer, 45 neurons in the second layer and one neuron in the third layer. The speaker is considered identified if the network output is in the band width $0.5 \pm 15\%$, that means if the output of the phonemes tested are above 0.58 the speaker is considered as different.

In an indicative way, to view the network behavior, instead of introducing only the same learned phonemes for test, all the phonemes features (SMMFCC) composing our seven test sentences previously segmented in the TIMIT data-base are introduced after localizing the learned phonemes position, and viewing their outputs.

The Recognition Rate of the tested phonemes is 98.57 %. Moreover, in the same test, 10 phonemes that were never introduced to the network for training were identified as learned phonemes. Then, we introduce a Confusion Rate (CR) to characterize the false recognized phonemes. In this case, the CR is about 11%.

Experiment 2: Medium phonemes data-base

The phonemes chosen are in number of 10 as the following: (SH-IX-HV-EH-DCL-IH-QH-KCL-K-S), five examples of each phonemes, are introduced and the same structure of the last network is preserved. The recognition rate will be about 97.05%.

We found that the network confuses between the TCL phoneme never learned) and the DCL, KCL phonemes, learned with five examples for each one . The confusion rate in is CR= 35.15 %.

Furthermore, when the choice of the training phonemes is not correctly made, a confusion in the network decision would be noted. When the near phonemes models are avoided in the training phase, and the network is tested with only the phonemes kind learned, the recognition rate will increase up to 100% and the confusion rate will decrease to 1%.

Experiment 3 : Large phonemes database:

All the 48 phonemes kinds SMMFCC of the learning database are used for the training with five examples for each phoneme.

We obtain a recognition rate equal to 87.23 %, and a confusion rate CR = 42.934 %.

We conclude that there is no need to characterize the speaker by a large number of phonemes. This does not improve the Recognition Rate comparing to a rigorously selected set of phonemes.

Experiment 4: Speaker recognition using only vowel phonemes :

Basing on the fact that vowels in signal processing are rich of energy and have a high frequencies,[1] we try in this last experiment to view the efficiency of the vowels phonemes in speaker recognition.

From the three training sentences, we extract the only vowels phonemes, to form the database. A set of 11 vowel kinds is extracted. Five examples for each vowel kind are used for learning, the desired outputs are as shown in the Table 3.

Vowel phonemes		**Desired output**			
1	**IX**	-0.5	-0.5	-0.5	-0.5
2	**EH**	-0.5	-0.5	-0.5	+0.5
3	**IH**	-0.5	-0.5	+0.5	-0.5
4	**AH**	-0.5	-0.5	+0.5	+0.5
5	**AO**	-0.5	+0.5	-0.5	-0.5
6	**AE**	-0.5	+0.5	-0.5	+0.5
7	**IY**	-0.5	+0.5	+0.5	-0.5
8	**OZ**	-0.5	+0.5	+0.5	+0.5
9	**EY**	+0.5	-0.5	-0.5	-0.5
10	**AY**	+0.5	-0.5	-0.5	+0.5
11	**AA**	+0.5	-0.5	+0.5	-0.5

Table.3 : Vowel phonemes extracted, and their desired outputs

The test phase is done by extracting the vowel phonemes parameters from the test sentences, and compared with those learned by the network.

We carried out several experiments using various neural network architectures. The best obtained Recognition Rate is equal to 77 % and is given by a neural network having 12 inputs, 45 neurons in the first layer and 4 output neurons. The sigmoid threshold is 0.1.

4. CONCLUSION:

This paper proposes a new approach to characterize the speaker using the model of he's phonemes speech. This method has proved its efficiency in speaker identification and different speakers discrimination.

The experimental results has proved that using a small or medium phonemes database provides an excellent recognition rate, that we can improve taking care of the following:

- Phonemes that are near in the look or pronunciation must not be introduced in the learning step because of the confusion that the network can make in the test phase.
- The introduce of all the speakers sentences (segmented into phonemes) in the test step aims only to show the confusion that can make the network when we introduce a too near phonemes.
- The appropriate choice of the neural network structure, and the re-injecting of the weights in the network inputs, provide more success for the classification method, and decrease the learning error rate.
- The robustness in recognition decision can be increased or decreased, in the first three experiments, by the adjust of the error threshold band.
- The approach based on vowels has not provided a high recognition rate, in spite of the medium phonemes database used, and the different number of example introduced for learning.

Finally, the proposed method proved it's efficiency by giving a high recognition rate when compared with other methods [8]. In the other hand, because of it's capacity to characterize the phoneme and the speaker, it's clear that this method can be used in both speech and speaker recognition

5. REFERENCES :

[1] R Boite, M.Kunt, "Traitement de la parole", EPFL-Ecublens, Lausanne, Suisse.

[2] J. Caelen, "Le traitement du signal voccal", http:// www-geod.imag.fr/jcaelen/rapport.htm.

[3] Calliope, "The speech and it's automatic processing", Edition Masson, 1989, Vol 2-4.

[4] S. B. Davis, P. Mermelstein, "Comparision of parametric representation for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. ASSP, vol.28, no. 4, August 1980, pp.357-366.

[5] James L. Flanagan, "Speech Analysis Synthesis and Perception", second expended edition, Heidelberg, New York.

[6] J. A. Freeman and D. M. Skapura, "Neurals networks algorithms application and programming techniques", Addisson-wesley 1991.

[7] B. Jacob, "Automatic speech recognition", Doctorat, Paul Sabatier university, Toulouse, September 2003.

[8] B. N. Li and James.N.K.LIU, "A comparative study of speech segmentation and features extraction on the recognition of different dialects", proc. IEEE SMC 1999, vol. 1.

[9] R.P.Lippmann "An introduction to computing with Neural Networks " IEEE ASSP Mag., vol.4. no. 2, pp. 4-22, April 1987.

[10] H. Matsumoto, "Evaluation of mel-lpc cepstrum in a long vocabulary continuous speech recognition", Proc. IEEE ICASSP 2001.

[11] D. Meuwly, "Speaker recognition in forenstic science", Doctorate, criminals and scientific police institut, Lusiane university, December 1996.

[12] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Mixture Models", Digital Signal Processing, vol. 10, pp 181-202, 2000.

[13] S. A Selouani, "A Hybrid HMM Autoregressive Time Delay Neural Network Automatic Speech Recognition System", INRS-Telecommunication, Université du Québec, Canada.

[14] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T.F. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," Proc. of IEEE ICASSP, May 2002, vol. 1, p. 677-680.