SPEAKER RECOGNITION WITH ARTIFICIAL NEURAL NETWORKS AND MEL-FREQUENCY CEPSTRAL COEFFICIENTS CORRELATIONS

Roberto A. B. Sória, Euvaldo F. Cabral Jr.
University of São Paulo - DEE/EPUSP
Laboratory of Communication and Signals - LCS
CAIXA POSTAL 8174, São Paulo, SP, 01065-970, Brazil
e-mail: rsoria@lcs.usp.br, euvaldo@lcs.usp.br

ABSTRACT

The problem addressed in this paper is related to the fact that classical statistical approach for speaker recognition yields satisfactory results but at the expense of long length training and test utterances. An attempt to reduce the length of speaker samples is of great importance in the field of speaker recognition since the statistical approach, due to its limitations, is usually precluded from use in real-time applications. A novel method of text-independent speaker recognition which uses only the correlations among MFCCs, computed over selected speech segments of very-short length (approximately 120ms) is proposed. Three different neural networks - the Multi-Layer Perceptron (MLP), the Steinbuch's *Learnmatrix* (SLM) and the Self-Organizing Feature Finder (SOFF) - are evaluated in a speaker recognition task. The ability of dimensionality reduction of the SOFF paradigm is also discussed.

1. INTRODUCTION

Many researchers have tried speaker recognition using statistical features [1]. Previous works in the field of speaker recognition using statistical approach and correlation of features attempted to use the short-time power spectrum of speech at different frequencies, or features involving the pitch, log area ratios and correlations. It has been shown that the short-time power spectrum of speech, at different frequencies, present a significant degree of correlation and that the spectral correlations can be very useful for text-independent speaker recognition [2], but a stable evaluation requires averaging over long utterances. The contribution of this work lies in the proposal of a novel method of parametric analysis for text-independent speaker recognition based on the correlation of Mel-Frequency Cepstral Coefficients [3], computed over of very-short length selected speech segments collected from simple utterances. This new approach came from the observation that, depending on the selected segments, some MFCCs are found to be highly correlated and that those correlations can be used for speaker recognition. A statistical analysis showed that the correlated coefficients present low intra-speaker variances and high inter-speaker variances. That means that the correlations vary according to the speaker. That fact led to many experiments which resulted in a high recognition rate for a small speaker population. The performance of what is named here, Mel-Frequency Cepstral Coefficients Correlations (MFC^3) was evaluated on a speaker identification and verification task, using a Multi-Layer Perceptron (MLP) and a Binary *Learnmatrix* network as classifiers. The MLP was found to classify better than the binary network, but both performed in a quite robust way. The Self-Organizing Feature Finder (SOFF) was used to reduce the dimension of the space of features. The use of correlation among MFCCs combined with neural networks produced very high recognition rates. The MFC^3 appear to be a powerful tool for talker recognition.

2. SPEECH CORPUS AND FEATURE EXTRACTION

The speech corpus is a subset of the POLIDATA database, created by the *SANNgroup* (*Speech and Artificial Neural Networks research group*) at the Polytechnic School of the University of São Paulo. The selected subset is composed by a set of ten speakers. Nine utterances of the phonetically balanced portuguese sentence "Amanhã ligo de novo" ("tomorrow I will call again")have been used in these experiments; ten speakers uttered nine times the

sentence. Speech data was digitally sampled at 11kHz, using a 16 bits A/D converter. Each utterance was endpointed and divided into speech segments with duration of approximately 120 ms. A shift of 12ms, between them, was used to detect exactly which of them yields the higher recognition rate. Five utterances were used in the training phase and four in the test phase. The pre-processing produced a set of conventional Mel-Frequency Cepstral Coefficients (MFCCs). The speech signal, inside a speech segment, was Hamming windowed and preemphasized with a first-order digital filtering. The 12 first MFCCs were computed for sequential frames of 23.2 ms of duration, 1ms apart. Once the MFCCs were obtained, the correlation matrix was calculated for each utterance.

3. ASSESSMENT OF THE MFC^3 METHOD

As a first experiment, the recognition rate, for a set of five speakers, was evaluated along the chosen sentence, using a MLP as classifier. This heuristic search has pointed out to speech segments that maximizes the recognition rate. Since the correlation matrix was 12×12, the symmetrical matrix had 144 elements, of which 66 were useful. The mean vector of those 66 elements could be used to train the neural net. Many experiments were carried out in order to choose the number of training vectors for each speaker, but it was realized that the mean vector (the mean matrix put in a vector form) alone was quite robust to represent the speaker and training the net. A second experiment which goal was to compare the classification through the neural network with a simple binary associative network was carried out. Steinbuch's *Learnmatrix* [4] was used as classifier and trained only with the mean vector coded in 1-bit. Steinbuch's network is a binary matrix which accepts a binary vector as input, produces a binary vector as output, and is distinguished by its use of binary weights. It was invented in 1958 by Karl Steinbuch whose goal was to produce a network that could form associations between pairs of binary patterns. Learning takes place via a Boolean form of Hebbian learning. The *Learnmatrix* presents itself as a very important associative network due to its clever analysis capacity and simplicity. The recognition rate along the sentence, for the two first experiments, is presented in figure 1.

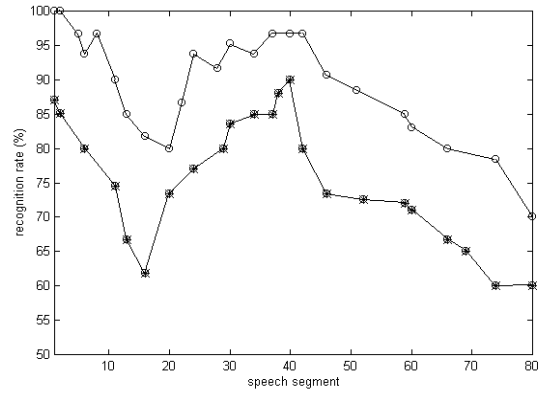

figure 1 - Trajectory of the average recognition rate along the sentence, for the MLP (o), and the Learnmatrix(*)

Those experiments allowed the selection of the speech segments which are more suitable to perform a speaker recognition task with high recognition rate. As illustrated in figure 1, the speech segments numbered 1, 2 and 40 are quite suitable for that task. It should be noted that the two shapes are quite similar, which means that the *Learnmatrix* is an important tool for previous analysis of the speech data. Some experiments were also carried out in order to combine different speech segments. The resulting recognition appeared to be much more robust. A 2-bits coding using the segments 1 and 40 was performed for comparison and the results are presented in table I. In a third experiment, the MFC^3 were evaluated for a set of ten speakers, using the first speech segment. The number of MFCCs varied from 12 to 15. That means that the number of independent elements of the correlation matrix, i.e. MFC^3, varied from 66 to 105. A great advantage of this method is that a quite small neural net could be used as a classifier. The MLP was optimized in terms of the number of hidden neurons and neural function and trained with one mean vector per speaker so that the net's training was very fast. The experimental results, presented in table II, show that for a set of ten speakers, the identification rate was 100% when the correlation matrix was computed for 15 MFCCs, of which 105 independent MFC^3 were used to represent each speaker. Table III presents some results for the classical approach using the MFCCs as features. In this case the eight first MFCCs were computed along

Table I -Identification results for 1 bit and 2 bits coding using the SLM, in comparison with the MLP

	training set		test set	
Speech segment number	01	40	01	40
Learnmatrix - 1-bit Coding	83.4%	86.7%	86.7%	93.4%
Learnmatrix - 2-bits Coding	96.7%	93.4%	93.4%	93.4%
MLP	100%	100%	100%	93.4%

Table II - Identification results using the MLP as classifier for a set of ten speakers.

number of MFCCs	number of MFC^3	number of hidden neurons	identification Rate (%)	
			training set	test set
12	66	15	95	100
12	66	20	95	96.7
13	78	15	98.4	93.4
13	78	20	96.7	96.7
14	91	30	98.4	96.7
14	91	35	98.4	100
15	105	40	98.4	100
15	105	30	100	100

Table III-Identification results using the MFCCs as features and the MLP as classifier, for a set of ten speakers

number of MFCCs	dimension of the input space	non-optimized number of hidden neurons	identification rate (%)
8	1200	60	86.7

the sentence and then normalized using the Linear Time Warping algorithm (LTW) with 150 points. In this approach, 72600 synapses were formed while only 3450 synapses were needed in the MFC^3 approach, for the best identification rate. Some experiments were also carried out to evaluate the performance of that process in a speaker verification task. The MFC^3 mean vector for one speaker was labeled as "one" and the mean vectors for the remaining speakers were labeled as "zero". One MLP was then trained for each speaker. If the output exceeded a threshold, the speaker was accepted, otherwise, it was rejected. The threshold was experimentally obtained so that it could yield simultaneously a minimum false acceptance rate and false rejection rate. The system achieved a 99% verification rate. A last experiment evaluated the Self-Organizing Feature Finder (SOFF) [5], with two layers, presented in figure 2. In the recognition phase, each feature detector (FD), which is a vector of templates, computes the match (normalized dot product - equation 1) between its resident feature and the input vector. This means that whenever an input vector is applied to any layer of the SOFF, each feature detector outputs a number that expresses the similarity between both features. In learning mode, the SOFF creates a set of feature detectors which are tuned to correspond to some class of features.

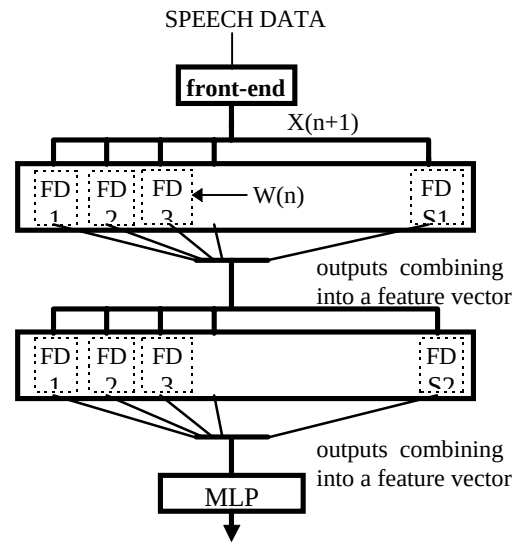

figure 2 - SOFF recognizer. Speech data is converted into a feature vector which is transformed to different feature representations by the SOFF layers.

The learning is accomplished by locating the maximum output in each layer, say $\eta_{max}(n)$, defined by $\eta_{max}(n) = \max\{\eta(n)\}$, where $\eta(n)$ is given by:

$$\eta(n) = \frac{W^T(n) \cdot X(n+1)}{\|X(n+1)\| \cdot \|W(n)\|} \qquad (1)$$

Table IV - Identification results using the SOFF recognizer for a set of ten speakers.

number of MFCCs	input layer	output layer	number of hidden	identification rate for
12	66	48	40	100
12	66	18	18	96.3
12	66	18	15	96.3
12	66	18	10	88.9
12	66	10	10	88.9

$\eta(n)$ represents the match between a feature *W(n)* and the input *X(n+1)*. The parameters HIGH and LOW are used to classify an input into three categories with respect to a given feature detector - detection, learning, or null. If $\eta_{max}(n) >$ HIGH (detection region), then the corresponding feature detector remains unmodified. Else, if $\eta_{max}(n) >$ LOW (learning region), then a learning rule (equation 2) is applied to the components of the feature detector.

$$W_{max}(n+1) = f_1 \cdot X(n+1) + f_2 \cdot W_{max}(n) \qquad (2)$$

with f_1, f_2, functions of $\eta_{max}(n)$. Else, if $\eta_{max}(n) <$ LOW (null region), a new feature detector is created since the input pattern is not part of the known features. The training set of MFC^3 was then presented to the first layer of the SOFF. Ten feature detectors were initialized with the mean vector of the MFC^3 of each speaker. The thresholds HIGH and LOW were varied so that the SOFF ability of creating feature detectors could be evaluated in terms of recognition performance. The value LOW controls the number of FD that are actually created.The higher the LOW threshold value, the higher the number of FD created. The number of FD of the second layer corresponds to the dimension of the output vector of the net. Once the SOFF was trained, each test vector was applied to the network and the output, with reduced dimensionality, was applied to a MLP to make the decision. Some results of speaker identification using a two layers SOFF for a set of ten speakers are presented in table IV. It should be noted that for 48 FD in the output layer, the SOFF produced a 100% identification rate. However, the more useful result was the one for 18 FD, in the output layer, which produced an identification rate of 96.3%.

4. CONCLUSION

This paper introduced a novel method of speaker recognition using statistical features based on the cross correlation of MFCCs, called here *Mel-Frequency Cepstral Coefficients Correlations* (MFC^3). An advantage of the MFC^3 approach is that there is no need for long time utterances. Segments of only 100ms to 150ms duration are sufficient. That means that to characterize a certain speaker, it is enough to analyze the phonetic contents of the sentence and select a segment that is appropriate to the correct recognition. Final results reached an identification rate of 100% for a set of ten speakers and a verification rate of 99%, using the MLP as classifier. The good results obtained with the *Learnmatrix* using few bits of coding over very short speech segments, showed that the MFC^3 can be a powerful tool in a real time speaker recognizer. In addition, the Self-Organizing Feature Finder (SOFF) presented very good results for closed-set speaker identification and it appears to be quite useful for dimensionality reduction, representing a reduction of 72.7% with respect to the original dimension of the feature space, achieving a recognition rate of 96.3% for a set of ten speakers.

5. REFERENCES

[1] FURUI S., “Comparison of Speaker Recognition Methods Using Statistical Features and Dynamic Features”, IEEE Transaction on ASSP, Vol. 29, No. 3, pp. 342-350, June 1981.

[2] ATAL B., S., “Automatic Recognition of Speakers from Their Voices”, Proc. of the IEEE Vol. 64, No. 4, pp. 460-475, April 1976.

[3] MERMELSTEIN P., DAVIS, B., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Trans. on ASSP, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[4] STEINBUCH K., “Die Learnmatrix - The Beginning of Associative Memories”, Advanced Neural Computers, R. Eckmiller, Elsevier Science Publishers, B.V. (North-Holland), pp. 21-29, 1990.

[5] LERNER S. Z., DELLER J. R. Jr., “Speech Recognition by Self-Organizing Feature Finder”, International Journal of Neural Systems, Vol. 2, Nos. 1 & 2, pp. 55-78, 1991.