NEURAL NET APPROACHES TO SPEAKER VERIFICATION: COMPARISON WITH SECOND ORDER STATISTIC MEASURES

M. Mehdi Homayounpour(*,**), Gérard Chollet(+)
(*)CNRS/URA 1027, 19 rue des Bernardins, 75005, Paris, France
(**) Amirkabir University of technology, Hafez Street, Tehran, Iran
(+)IDIAP, C.P. 609 , 1920, Martigny, Switzerland

ABSTRACT

The non-supervised Self Organizing Map of Kohonen (SOM), the supervised Learning Vector Quantization algorithm (LVQ3) [1], and a method based on Second-Order Statistical Measures (SOSM) [2] were adapted, evaluated and compared for speaker verification on 57 speakers of a POLYPHONE-like data base. SOM and LVQ3 were trained by codebooks with 32 and 256 codes and two statistical measures; one without weighting (SOSM1) and another with weighting (SOSM2) were implemented . As decision criterion, the Equal Error Rate (EER) and Best Match Decison Rule (BMDR) were employed and evaluated. The weighted Linear Predictive Cepstrum coefficients LPCC and the ΔLPCC were used jointly as two kinds of spectral speech representations in a single vector as distinctive features. LVQ3 demonstrates a performance advantage over SOM. This is due to the fact that LVQ3 allows the long-term fine-tuning of an interested target codebook using speech data from a client and other speakers, whereas SOM only uses data from the client. SOSM performs better than SOM and LVQ3 for long test utterances, while for short test utterances LVQ is the best method among the methods studied here.

1. INTRODUCTION

Neural network clustering algorithms have been employed for a large number of applications such as speech recognition and pattern recognition. It is possible that the representation of knowledge be in the particular form of a feature map that is geometrically organized. Kohonen showed that a set of interconnected adaptive units has the ability to change its responses in such a way that it will adapt to represent the characteristics of the input signal. It is the same as the classification problem in classical pattern recognition such as vector quantization algorithm, where the feature vector space is to be partitioned into a set of non-overlapping regions, and where each region is represented by a reference vector. In this study, the Self Organizing Map neural network clustering technique as a non-supervised classifier was compared to a supervised clustering technique (LVQ3). LVQ has been used by Kohonen in a system of phonetic recognition of Finnish and Japanese[3], and by Bennani et al. [4] and T. R. Anderson et al. [5] for speaker identification. Anderson et al. used a two-stage approach to speaker recognition. In the first stage a classification into broad categories of vowel phonemes is done. The second stage uses one or more of those categories for speaker identification. Speaker-dependent codebooks are created using a SOM technique and tuned by a LVQ3 algorithm. Bennani obtained a performance of 97% using Mel Frequency Cepstrum Coefficients on a data base of 10 speakers. The second order statistical measures were proposed by Grenier[6], Gish[7] and were recently developed by Bimbot[2]. These measures are easy to implement and are computationally efficient. They are expressed as a function of the eigenvalues of a covariance matrix related to a reference covariance matrix. These measures have frequently been used for speaker Identification [2].

The parametric representations used in this study are based on Linear Predictive spectra. They comprise the LPCC, and ΔLPCC coefficients. These two representations proved to be highly relevent in speaker verification tasks when they are weighted by the reciprocal of their variability [8].

The following section provides a description of the data base used and the processing done on this data base. A description of the two decision criteria is given in section 3. Supervised and non-supervised training procedures and the SOSM mesures used are described in section 4. Experimental results are given in section 5, and the last section provide a conclusion

2. DATA BASE AND PROCESSING

Recorded utterances of 57 speakers (36 males and 21 females) of a POLYPHONE like data base were used in this study. These speakers were recorded in several sessions over more than 3 months. This data base is a telephone data base recorded over local and long distance telephone lines using different types of handsets. Recordings took place from the speaker's office or from his/her home. The Signal-to-noise ratio was generally better than 15 dB. Recordings were done using a DIALSYS 4 PC-board at 8kHz in logarithmic A law by the IDIAP research center in Switzerland. Recorded utterances were transformed to 8kHz, 16 bits linear form and pre-emphasized by a first order filter with the transfer function of $1\text{-}0.94z^{-1}$ and subsequently multiplied by a Hamming window function. Each analysis frame spanned 30ms and was shifted by 10 ms. A vector of length 22 was retained which comprised 12 LPCC and 10 ΔLPCC coefficients. The ΔLPCC coefficients as the first-order orthogonal polynomial coefficients represent the slope of the time-function of each coefficient in the cepstral vector. The transitional feature window length used here corresponds to a window width of 90 ms. Of the 57 speakers that participated in our speaker verification experiments, 17 speakers (10 males and 7 females) were considered as targets and participated in a training phase, and 40 others (26 males and 14 females) played the role of imposters. A cepstral mean removal channel compensation technique was used to remove any fixed frequency-response distortion introduced by the transmission system. Normalized cepstral coefficients were obtained simply by subtracting from the cepstral coefficients their averages over the duration of the entire telephone call. Each coefficient in the feature vector was weighed by the reciprocal of its standard deviation obtained using 2s of training speech from each of the 17 targets.

0-7803-2431-5/95 $4.00 © 1995 IEEE

3. DECISION CRITERIA

Two decision criteria were used to evaluate the three methods for speaker verification under study. EER correspond to the intersection point of False Rejection (FR) and False Acceptance (FA) error rate curves. This criterion has largely been used to show the performance of speaker verification systems. Another criterion, named the Best Match Decison Rule (BMDR) is proposed: A general model is first created using trainig data from a number of clients or all clients in the training data base. Individual models for each client are created using their training data. In a verification test, a test utterance is compared with both the general model and the individual model of the claimed speaker. If the best match is to the speaker's individual model, then he/she will be accepted $(D_i<D_g)$, and if not i.e. $(D_i \geq D_g)$ he/she will be rejected, where D_i and D_g represent the distance(distortion) between test utterance and the indivual model of client i and between the test utterance and the general model, respectively. The number of times that a target was rejected and the number of times that a non-target speaker was accepted divided by the total number of intra- and inter-speaker verification tests gives the FR and FA error rates, respectively.

4. TRAINING AND VERIFICATION PROCEDURES

An Euclidean distance measure was used in this experiment to calculate the quantization distortion. In our experiments with the SOM method, a general prototype map with x and y dimensions of 16, i.e. a map of 16*16=256 cells, was obtained by using 2 seconds of speech obtained from 17 target speakers. It serves as the initial condition for every target speaker. The topological structure of this map was hexagonal and the neighborhood function type was a step function. Two learning phases were conducted to obtain the general prototype map. In the first one the number of training steps was 5000 with an initial radius of 16 and a learning rate of alpha=0.05. In the second phase the number of training steps was 12000 and the initial radius was 4 with alpha=0.02.

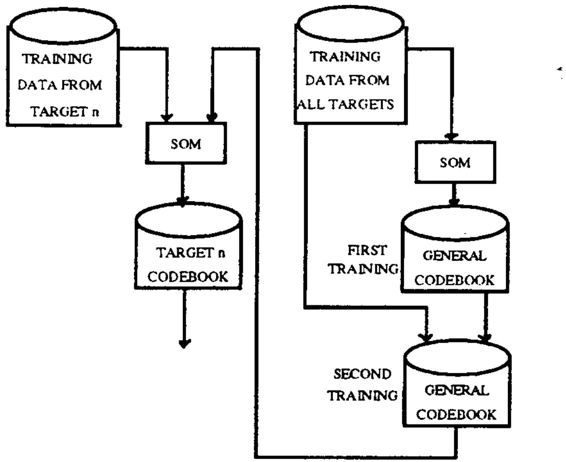

Figure 1. SOM training phase for speaker verification

The initial radius decreased linearly to 1 during learning. To obtain an individual map for each speaker, another training procedure was conducted. This procedure was to start at each time with the general prototype map and to train it by 32s of speech from a target in order to obtain an individual map of this speaker. The number of training steps was again 12000 and the initial radius was 4 with alpha=0.02. In this way 17 individual maps, one map for each target, were constructed.

In the verification phase, the vectors x(t) from a short parameterized speech interval from the speaker under verification are compared to both the general prototype and the speaker's individual map. As a result, two accumulated distortions were calculated: one between x(t) and the general prototype map (D_g), and the other between x(t) and the individual map (D_i). The decision made was whether the speaker's voice matches, with minimum distortion, the claimed speaker's individual SOM or the general prototype SOM, using BMDR criterion. The Equal Error Rate (EER) criterion was also used to evaluate the performance of the SOM algorithm. A similar experiment was conducted with a map of 4*8=32 cells.

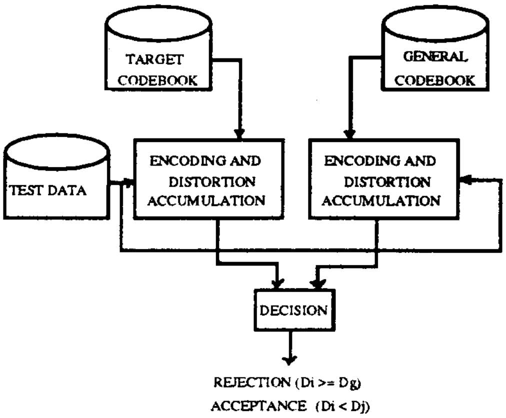

Figure 2. SOM verification phase using BMDR criterien

In order to apply LVQ3 to speaker verification, the following method was proposed: for each target a set of training feature vectors corresponding to 32 seconds of speech data was extracted. This data comprises 16 seconds of speech from one target speaker and 1 second from each of the 16 other target speakers in the target population. We gave a label of 1 to those vectors belonging to the target speaker and a label of zero to those of the other speakers. A total of 256 entries was entered in a codebook for the target speaker, with 128 vectors from the target speaker and 128 vectors from other speakers. These codebook vectors were extracted from the training set and were supposed to fall inside the class borders, which were tested automatically by a knn (k-nearest-neighbors) classifier using k=7. Training of this initial target codebook using the LVQ3 algorithm was done using the training feature vectors of this speaker. The following configuration for LVQ3 parameters was chosen: alpha=0.02, window width = 0.2, epsilon = 0.2, and the number of training steps = 10000.

In the verification phase, the feature vector of a test utterence was compared to all vectors in the codebook and the label of codebook-vector with the smallest distance to this feature vector was considered. This procedure was repeated for all feature vectors in the test utterance. A verification score was obtained which is equal to the number of testing vectors classified with the label 1. A speaker was accepted if his verification score was higher than a decision threshold, otherwise he/she was rejected.

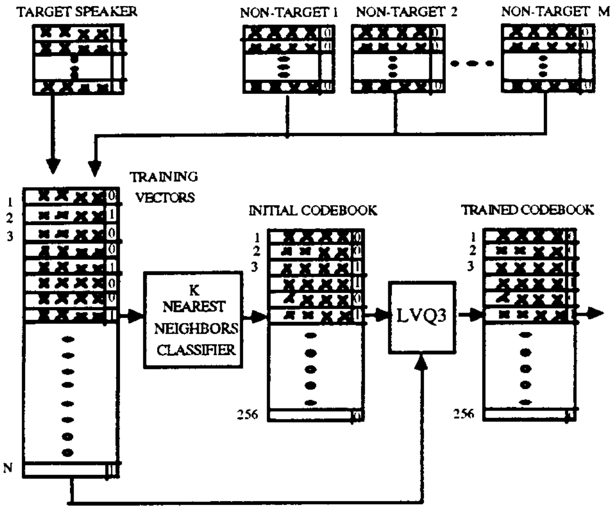

Figure 3. LVQ3 training phase for speaker verification

The principle of LVQ3 training for speaker verification doesn't alow the creation of a general model of speakers to be used by BMDR. So only EER criteria can be used to obtain the verification error rate. A similar experiment with a codebook size of 32 was also conducted.

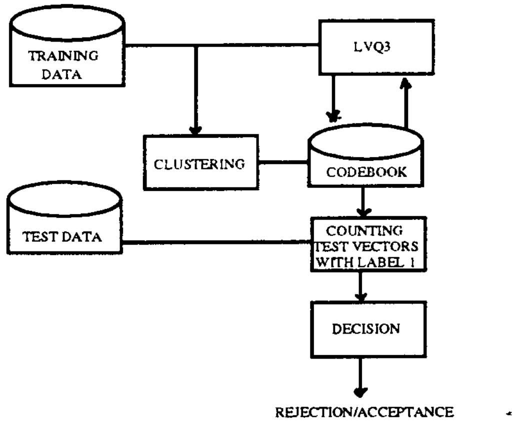

Figure 4. LVQ speaker verification phase

Two following Second Order Statistic Mesures (SOSM1 et SOSM2) were implemented:

- SOSM1: an arithmetic-harmonic sphericity distance measure between a test covariance matrix Y and a reference covariance matrix X is defined as:

$$\mu_{AH}(X, Y) = \log \left[\frac{A}{H}\right]$$

- SOSM2: the weighted symetric sphericity mesure proposed by [2] as defined as:

$$\mu_{SPH\ sym}(X, Y) = \rho_{mn} \cdot \log(\mathrm{tr}(YX\text{-}1)) + \rho_{nm} \cdot \log(\mathrm{tr}(XY\text{-}1))$$

$$-\frac{1}{m} \cdot (\rho_{mn} - \rho_{nm}) \cdot \log\left[\frac{\det(Y)}{\det(X)}\right] - \log(m)$$

$$\text{with: } \rho_{mn} = \frac{m}{m+n} \quad \text{and} \quad \rho_{nm} = \frac{n}{m+n}$$

where m represent the number of training vectors and n the number of test vectors.

A General and for each client an individual covariance matrix were obtained using the same size of training speech material used already for training of SOM. Similar experiments with SOM, LVQ3, and second order statistic measures were conducted for test utterances with 1s, 2s, 3s, 6s, 9s, 12s, and 15s duration.

5. EXPERIMENTAL RESULTS

Tables 1, 2, et 3 show the performance of SOM, LVQ3, and SOSM as a function of test utterance durations of 1s, 2s, and 3s for different codebook sizes and decision criteria. Error rates were obtained by conducting 1989 intra-speaker and 1790 inter-speaker verification tests. Figure 5 shows the EER of the three methods for test utterances up to 15s in lenghts.

		3s			2s			1s		
		FR	FA	TER	FR	FA	TER	FR	FA	TER
Map Size	256 +	11 %	33 %	22 %	14 %	34 %	24 %	16 %	37 %	26 %
	32 +	5%	61 %	33 %	7%	59 %	33 %	10 %	54 %	32 %
	256 *			20 %			22 %			26 %

* Error rate obtained by the Equal Error Rate criterion.
+ Error rate obtained by the "Best Match Decision Rule"

Table 1. FR, FA, Total Error Rate: TER = (FR+FA)/2, and EER obtained by SOM as a function of test utterance duration and codebook size.

		TEST UTTERANCE DURATION		
		3s	2s	1s
CODEBOOK SIZE	256	14.6 %	17 %	20.5 %
	32	17.0%	20.3 %	24.3 %

TABLE 2. The EER obtained by LVQ3 as a function of test utterance duration and codebook size.

		TEST UTTERANCE DURATION		
		3s	2s	1s
SOSM1	BMDR	33.1%	31.5 %	33.1 %
	EER	21.3 %	26.7 %	35.9 %
SOSM2	BMDR	33,2%	33,2 %	34.0 %
	EER	18,4 %	24,8 %	33,6 %

Table 3. Speaker verification errors obtained by EER and BMDR criteria obtained by the SOSM1 and SOSM2 as a function of utterance duration.

6. DISCUSSION AND CONCLUSION

Tables 1 and 2 show a higher performance for LVQ3 in relation to SOM. A codebook size of 256 comparing to a codebook size of 32, gives a better performance for both LVQ and SOM. It should be noticed that in our experiments with the SOM, 32s of speech material were used to train each target model, while for LVQ3 we used only 16s of speech material from the target speaker. The rest of the necessary training data was obtained from a very short duration of speech (only 1s) from each of the other speakers in the target population. Therefore in LVQ3, a shorter utterance recorded by each target speaker would be sufficient to train a target model. In order to create a single speaker's model, LVQ3 employs possible data from all speakers in the population. This model can capture the differences between that speaker and other speakers and so it contains information that allows it to verify the target to whom this model belongs and to reject impostors.

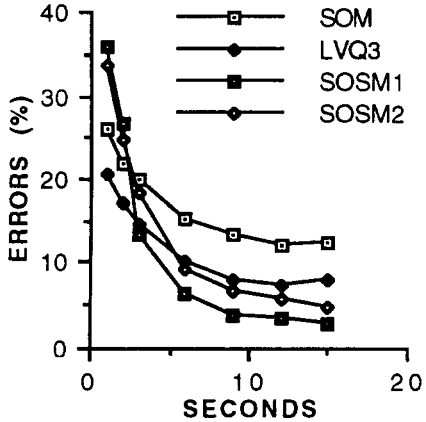

Figure 5. EER obtained for SOM, LVQ3, and the Second Order Statistic Measures (SOSM1 and SOSM2) as a function of test utterance duration.

A model trained by the SOM as a non-supervised technique uses only a similarity measure and uses data from only one target without taking into consideration the differences between this speaker and other speakers. Figure 1 shows that the LVQ3 comparing to SOM and SOSM gives a better performance for test durations less than 5 seconds. But the SOMS1 and SOSM2 are better than the SOM and LVQ3 methods for longer test utterances. Based on figure 1, when the test utterance duration increases, the EER deacreases strongly for SOSM1 and SOSM2 compared to SOM and LVQ3. This result may show that short test utterances are not sufficient for calculating a covariance matrix which represents a speaker well. As can be seen from figure 1, when the test utterance duration is longer than a certain duration, the error rate doesn't change significantly. For example, for test utterance durations greater than 9 seconds, there is no important improvement in the performance of the three methods especially for the SOSM1. The SOSM1 gives a 2.9% error rate when test duration is about 15 seconds. SOSM2 performs better than SOSM1 for short test utterances. But for test utterances up to 2s in lengths, SOSM1 performs better.

The EER shows generaly smaller error rates than the BMDR for the SOM algorithm. Using the EER criterion, the distance (distortion) between the test utterance of a given speaker and its reference model is compared to the decision threshold of this speaker and the result is used as the bases for accepting or rejecting him/her. Using the BMDR criteria, there is no *a priori* procedure for determining the decison threshold, and the decision threshold for a given speaker is the distance obtained by comparing a test utterance produced by this speaker to the general model. So the decison threshold is determined as a function of the test utterance. A comparison between the error rates obtained by EER and BMDR (tables 1 and 3), shows that the error rate obtained by BMDR doesn't change significantly as a function of test utterance duration. The FR error rate provided by BMDR is usually much less than the FA error rate. This may be explained in the following way: when a test utterance which belongs to a target speaker is compared to both the target's reference model and a general model, the distortion (distance) between the test utterance and the target's reference model must be inferior to the distortion (distance) between this utterance and a the general model. This is because an utterance produced by a speaker is usually more similar to its reference model than to any other model. However, when the test utterance belongs to an impostor, it is less probable that the the distortion (distance) between this utterance and the general model be necessarily inferior to the distortion (distance) between this utterance and the reference model of the target. A small FR error rate may be more useful for some applications where a client should be rejected as seldom as possible.

ACKNOWLEDGEMENT

This work was supported by the CNRS/URA 1027, PRESCOM, and the Ministry of Culture and Higher Education of Iran, which provided the financial support and the necessary research materials. The authors would like to thank the IDIAP research center which provided the POLYPHONE-like data base used in this study.

REFERENCES

[1] T. Kohonen, "The Self Organizing Map", Proceding IEEE, Vol 78, pp. 1464-1480, september 1990.

[2] F. Bimbot, L. Mathan, "Second-order Statistical Measures for Text-Independent Speaker Identification, ESCA Workshop on speaker Recognition, Identification, and Verification", pp. 51-54, Avril 1994.

[3] T. Kohonen, K. Torkkola, M. Shozakai, J. Kangas and O. Ventä, "Microprocessor Implementation of a Large Vocabulary Speech Recognizer and Phonetic Type Writer for Finish and Japanese", Proc. European Conference on Speech Technology (Edinburgh, 1987) pp. 377-380, 1987.

[4]Y. Bennani, F. Fogelman, P. Gallinari, "A connectionist approach for automatic speaker identification", ICASSP 90, 1990.

[5] T. R. Anderson, R; Patterson, " Speaker Recognition with the Auditory Image Model and Self Organizing Feature Map: A comparison with traditional techniques", ESCA Workshop on speaker Recognition, Identification, and Verification", pp. 153-156, Avril 1994.

[6] Y. Grenier, "Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique", PHD ENST-E-77005, 1977.

[7] H. Gish, "Robust discrimination in automatic speaker identification", ICASSP 90, pp. 289-292, 1990.

[8] M. M. Homayounpour, G. Chollet, "A comparison of some relevant parametric representations for speaker verification", ESCA workshop on Automatic Speaker Recognition, Identification and Verification, pp. 185-188, April 1994.