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 Decision Rule (BMDR) were employed and evaluated. The weighted Linear Predictive Cepstrum coefficients LPC and the ∆LPCC 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 Binbasi[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 ∆LPCC coefficients. These two representations proved to be highly relevant 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 measures 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 1.0-0.94z⁻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 ∆LPCC coefficients. The ∆LPCC 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 weighted by the reciprocal of its standard deviation obtained using 2s of training speech from each of the 17 targets.
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 Decision Rule (BMDR), is proposed: A general model is first created using training 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_1 < D_2), and if not i.e. (D_1 ≥ D_2) he/she will be rejected, where D_1 and D_2 represent the distance (distortion) between test utterance and the individual 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 16x16=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.

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 utterance 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.

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, and 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 lengths.

<table>
<thead>
<tr>
<th></th>
<th>3s</th>
<th>2s</th>
<th>1s</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>256</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>FA</td>
<td>53</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>TE</td>
<td>16</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>R</td>
<td>37</td>
<td>54</td>
<td>32</td>
</tr>
</tbody>
</table>

* 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: \( \text{TER} = (\text{FR}+\text{FA})/2 \), and EER obtained by SOM as a function of test utterance duration and codebook size.

<table>
<thead>
<tr>
<th>TEST UTTERANCE DURATION</th>
<th>3s</th>
<th>2s</th>
<th>1s</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODEBOOK SIZE</td>
<td>256</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>14.6%</td>
<td>17%</td>
<td>20.5%</td>
</tr>
<tr>
<td></td>
<td>17%</td>
<td>20.3%</td>
<td>24.3%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>TEST UTTERANCE DURATION</th>
<th>3s</th>
<th>2s</th>
<th>1s</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOSM1 BMDR</td>
<td>33.1%</td>
<td>31.5%</td>
<td>33.1%</td>
</tr>
<tr>
<td>EER</td>
<td>21.3%</td>
<td>26.7%</td>
<td>35.9%</td>
</tr>
<tr>
<td>SOSM2 BMDR</td>
<td>33.2%</td>
<td>33.2%</td>
<td>34.0%</td>
</tr>
<tr>
<td>EER</td>
<td>18.4%</td>
<td>24.8%</td>
<td>33.6%</td>
</tr>
</tbody>
</table>

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.

![Graph showing EER for SOM, LVQ3, and Second Order
Statistic Measures (SOM1 and SOM2) as a function of test
utterance duration.]

Figure 5. EER obtained for SOM, LVQ3, and the Second Order
Statistic Measures (SOM1 and SOM2) 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 SOM gives a
better performance for test durations less than 5 seconds. But
the SOM1 and SOM2 are better than the SOM and LVQ3
methods for longer test utterances. Based on figure 1, when
the test utterance duration increases, the EER decreases
strongly for SOM1 and SOM2 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 SOM1. The SOM1 gives a 2.9% 
error rate when test duration is about 15 seconds. SOM2
performs better than SOM1 for short test utterances. But for
test utterances up to 2s in lengths, SOM1 performs better.

The EER shows generally 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 decision
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 decision
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 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

for Text-Independent Speaker Identification, ESCA
Workshop on Speaker Recognition, Identification, and
Venna, "Microprocessor Implementation of a Large
Vocabulary Speech Recognizer and Phonetic Type Writer
for Finish and Japanese", Proc. European Conference on Speech
approach for automatic speaker identification", ICASSP 90,
1990.
the Auditory Image Model and Self Organizing Feature Map:
A comparison with traditional techniques", ESCA Workshop
on Speaker Recognition, Identification, and Verification,
locuteur d'un systeme de reconnaissance phonemique", PHD
[7] H. Gish, "Robust discrimination in automatic speaker
some relevant parametric representations for speaker
verification", ESCA workshop on Automatic Speaker
Recognition, Identification and Verification, pp. 185-188,
April 1994.