6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEAKER IDENTIFICATION USING DISCRIMINATIVE FEATURES SELECTION

Bogdan Sabac *Inge Gavat* *Zica Valsan*

Polytechnic University of Bucharest
7 Spinis St, Bl 45, Ap37, cod 75314, Bucharest, Romania
Phone: +(004-01)-6295544, e-mail: {sbogdan, inge, zica}@helix.elia.pub.ro

ABSTRACT

A new method of text-dependent speaker identification using discriminative feature selection is proposed in this paper. The characteristics of the proposed method are as follows: feature parameters extraction, vector quantization with the growing neural gas (GNG) algorithm, model building using gaussian distributions and discriminative feature selection (DFS) according to the uniqueness of personal features. The speaker identification algorithm is evaluated on a database that includes 25 speakers each of them recorded in 12 different sessions. All speakers spoke the same phrase for 10 times in each recording session. The test results showed that both FRR (False Rejection Rate) and FAR (False Acceptance Rate) were about 1[%].

1. INTRODUCTION

With many practical applications like, banking transactions over a telephone network, telephone shopping, database access services, information services, voice-mail, security control for confidential information areas and remote access to computers the speaker identification technology is an important area in paralinguistic speech analysis. In the early works of speaker identification, feature parameters are not always considered whether they have sufficient information for verifying the identity of individuals. In the case of identifying a person in our everyday life, it is considered that we extract the person's personal features from various aspects and identify the person by integrating those extracted features. Therefore, the same concept of feature selection & integration process will be useful in speaker identification. To realize this idea, we propose a new speaker identification scheme based on a VQ approach using the GNG algorithm, gaussian distributions and DFS.

The outline of this paper is as follows. Section 2 discusses the theoretical aspects of the GNG algorithm. In section 3 are presented the speaker modeling technique and the discriminative feature selection process. The speaker identification setup is provided by the section 4, in 5-th section are presented the experimental results and conclusions form section 6 close the paper.

2. GROWING NEURAL GAS

The VQ model proposed consists of a set A of formal units. Every unit $c \in A$ has associated a n-dimensional representative vector $w_c \in R^n$. The set W of all representative vectors is the current codebook. The GNG model has a graph like structure [4]. The vertices of the graph are the neurons and the edges denote neighborhood relations. The general idea of our method is to construct the codebook incrementally by interpolating new codebook vectors from existing ones. Interpolating is always done among topologically close neighboring units. After each interpolation the current codebook is adapted with a fixed number of vectors from the original data [9].

Furthermore, at each adaptation step a local information is accumulated at the winning unit bmu:

$$\Delta E_{bmu} = (error\ term) \qquad (1)$$

The particular choice of the above error term depends on the application. For vector quantization one would, e.g., choose $\Delta E_{bmu} = \|w_{bmu} - x\|^2$ whereas for entropy maximization an appropriate term is $\Delta E_{bmu} = 1$. Abstractly speaking, the error term should be a measure which is to be reduced and which is likely to be reduced in a particular area of the input space by insertion of new units in exactly this area. Since the neurons are slightly moving around, more recent signals should be weighted stronger than previous ones. This is achieved by decreasing all error term variables by a certain fraction after each adaptation step. The accumulated error information is used to determine (after a fixed number of adaptation steps) where to insert new units in the network.

Insertion of a new cell take place if after a number of adaptation steps the maximum accumulated error exceeds a insertion threshold [7], [8]. The new cell r is inserted between direct neighboring cells f and q with f having the largest accumulated error over the insertion threshold and q being a direct neighbor of f with the maximum accumulated error.

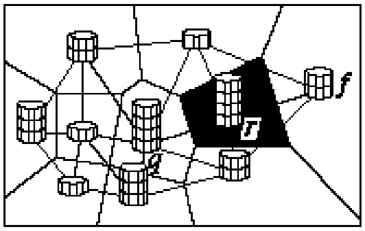

Figure 1: Insertion of a new cell r.

When an insertion is done the error information is locally re-distributed, increasing the probability that the next insertion will be somewhere else. The local error variables act as a kind of memory which lasts over several adaptation/insertion cycles and indicates where much error has occurred. Deletion of a neuron takes place if after a preset number of adaptation steps that neuron have not been a bmu. By insertion and deletion of neurons the structure is modified. The result are problem-specific network structures potentially consisting of several separate sub networks [10].

3. SPEAKER MODELING AND DFS

The parameters used for speaker identification are the first 12 mel frequency cepstral coefficients (MFCC) and the first 12 delta mel frequency cepstral coefficients (DMFCC) [2]. With the extracted feature vectors from a speaker, using the GNG algorithm two codebooks are constructed for that speaker. After the centroids are set for each of them we compute the mean and variance for each dimension obtaining in this way a non weighted gaussian mixture. The Gaussian distributions that model the acoustic features for each speaker are developed using the maximum likelihood (ML) estimation procedure. The mathematical expressions for the ML estimates for the means and variances for a particular speaker model are shown below.

$$\overline{\mu}_j = \frac{1}{n_j}\cdot\sum_{k=1}^{n_j} x_{j,k} \qquad (2)$$

$$\overline{\sigma}_j^2 = \frac{1}{n_j-1}\cdot\sum_{k=1}^{n_j}\left(x_{j,k}-\overline{\mu}_j\right)^2 \qquad (3)$$

$$P_w(j) = \frac{n_j}{n} \qquad (4)$$

Where:

j = the j-th broad class;
n_j = the number of tokens for class j;
n = total number of tokens
$x_{j,k}$ = the k-th data token for class j;
$\overline{\mu}_j$ = the ML estimate of the mean for class j;
$\overline{\sigma}_j^2$ = the ML estimate of the variance for class j;

The discriminative feature selection process is based on the idea that for a particular speaker model (codebook) we can detect the level of uniqueness of personal features for each of the gaussian distributions and weight them accordingly, instead of using the well known GMM [6] weighting technique where the codebooks vectors are weighted according to the relative number of frames assigned to them (4). The gaussian distribution weighting coefficients are calculated according to the uniqueness of personal features in each codebook as described in (5):

$$P_w(i) = \frac{Sc_m^P(i)\cdot H_n^P(i) - Sc_m^I(i)\cdot H_n^I(i)}{Sc_m^P(i)\cdot H_n^P(i)} \qquad (5)$$

if $P_w(i) < p$ then $P_w(i) = 0$, with $0 < p < 1$ where:

$P_w(i)$ represents the weight associated to the i-th gaussian distribution from the current codebook and represents the level of uniqueness of personal features *stored* in the gaussian distribution

$Sc_m^{P,I}(i)$ represents the medium (m) score (Sc) obtained at the i-th gaussian distribution when trough the current codebook are passed the codebook training vectors (P) or the impostor training vectors (I).

$H_n^{P,I}$ represents the normalized (n) histogram (H) of the bmu when trough the current codebook are passed the codebook training vectors (P) or the impostor training vectors (I).

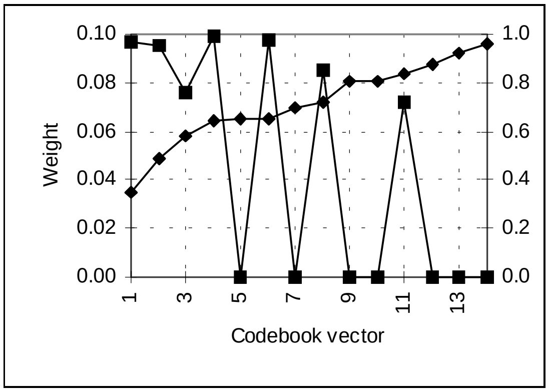

Figure 2: GMM◆ versus DFS■ weighting 42.7% of the frames are actually used for speaker identification

We can see that by using the DFS weighting procedure we obtain totally different weights; and also we can calculate the amount of information that is used by the speaker identification procedure. Figure 3 presents the amount of "useful" information "generated" by each speaker in the database.

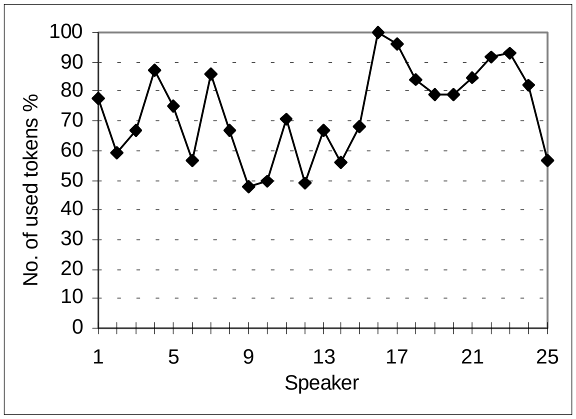

Figure 3: No. of "useful" frames (%) used in the identification process from each speaker

4. SPEAKER IDENTIFICATION SETUP

Speech signal was sampled at 11 kHz with a 16 bit digitizer. The speech signals are analyzed with a 20 ms Hamming window shifted every 10 ms in order to extract the following parameters from each frame of speech discarding low energy speech frames: 12 mel frequency cepstral coefficients (MFCC), 12 delta mel frequency cepstral coefficients (DMFCC) calculated as polynomial expansion coefficients over speech segments of five frames in length.

With the extracted feature vectors from a speaker, using the growing neural gas algorithm two codebooks are constructed and centroids weighted according to the DFS rule for that speaker. This process is repeated for all speakers in the population.

The linear opinion pool has been considered in our speaker identification experiment for the combination of features, namely cepstrum and delta cepstrum, in order to take the identification/rejection decision.

5. EXPERIMENTAL RESULTS

The algorithm is evaluated on a database that includes 25 speakers each of them recorded in 12 different sessions. 20 speakers are selected as registered speakers and 5 as impostors speakers with the aim of testing the identification system in 'open set' mode. All speakers spoke the same phrase: "My voice is my passport" for ten times in each session. All speakers where male between 21 and 23 years old. The codebooks where constructed using the first 2 pronunciations of the recording sessions 1 to 5 for each speaker.

The cohort size for the registred speakers have been set to 5 and as cohort speakers pool we have used the same set of 25 speakers.

Speaker identification experiments where performed using the fixed thresholds established in the test phase on the utterances form recording sessions 6 to 12 and also using the cohort comparisons technique. The threshold for the VQ accumulated distortion is varied from the point of 0% false acceptance to 0 % false rejection in order to have the operating curves shown in figure 4 and figure 5.

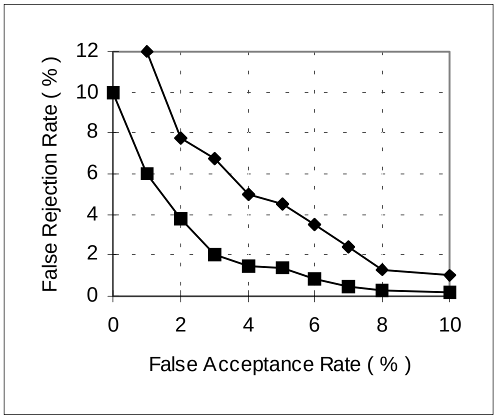

Figure 4: Open set speaker identification 26 speakers. Performance comparisons for GNG◆ (FRR = FAR = 5%) and GNG-DFS■ (FRR = FAR = 2.8 %) for a codebook size 32, fixed threshold.

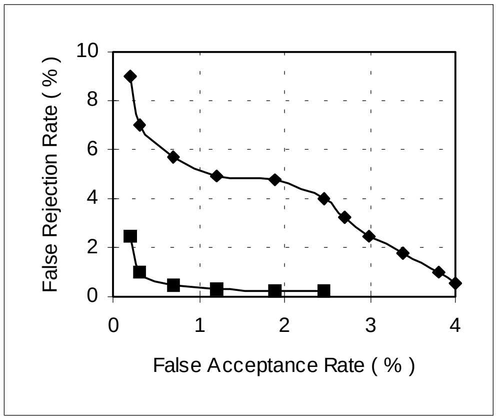

Figure 5: Open set speaker identification 25 speakers. Performance comparisons for GNG◆ (FRR = FAR = 2.9 %) and GNG-DFS■ (FRR = FAR = 0.5 %) for a codebook size 32, cohort.

In figure 4 and 5 it can be seen that the GNG + DFS algorithm improves the speaker recognition performance and also that the cohort normalized scores yields better results in both cases.

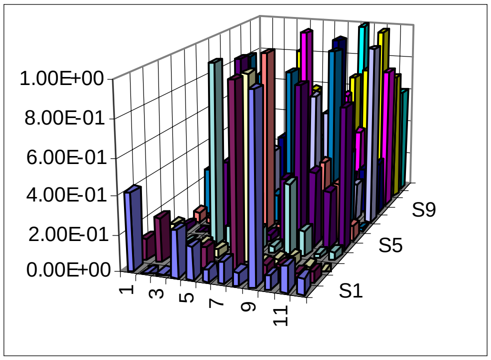

Figure 6: (a) GNG - Confusion Matrix

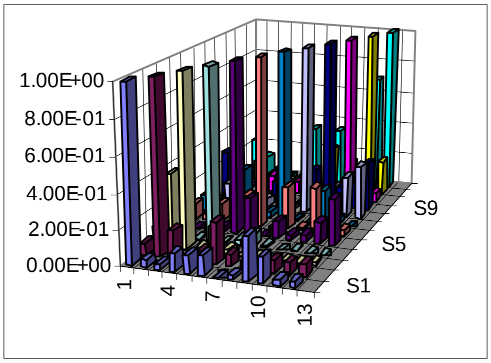

Figure 6: (b) GNG+DFS - Confusion Matrix

Figure 6: Confusion matrix from a live test with 12 out of 25 speakers. For a codebook of size 8. The scores on the diagonal are the "true speaker" scores. The rest are "impostor" scores. The difference shows how well the system separates them. The GNG + DFS method improves the speaker identification performance (b).

Table 1: Speaker identification success rate as a percent for GNG and GNG + DFS methods for different codebook size.

Codebook size	8	16	32	64
Overall Speaker Identification Performance (fixed threshold)				
GNG	67%	89%	95%	96.2%
GNG + DFS	92%	96%	97.2%	98.7%
Overall Speaker Identification Performance (cohort)				
GNG	76%	92%	97.1%	97.6%
GNG + DFS	94%	98%	99.5%	99.7%

6. CONCLUSIONS

We presented a vector quantization method which incrementally builds up a codebook through interpolation. In the proposed method, feature parameters are classified into several categories, by using the new clustering algorithm, and weighted properly according to the uniqueness of personal features. The system is evaluated for a combination of two feature sets, namely MFCC and DMFCC, employing the linear opinion pool criterion. The database for making the codebooks consists of 25 male Romanian speakers. The test results showed that both FRR (False Rejection Rate) and FAR (False Acceptance Rate) were under 1% at the minimum of (FRR+FAR) under the condition of a codebook of size 32 produced by the proposed algorithm.

7. REFERENCES

1. Kohonen, T., Self-Organized Formation of Topologically Correct Feature Maps, Biological Cybernetics, 43, pp. 59-69, 1982.

2. Furui S., On the role of dynamic characteristics of speech spectra for syllable perception, Fall Meeting of Acoust. Soc. Japan, 1-1-2: October, 1984.

3. B. Fritzke, Kohonen Feature Maps and Growing Cell Structures - a Performance Comparisons, Advance in Neural Information Processing Systems 5, L. Giles, S. Hanson, J. Cowan, eds., Morgan Kaufmann Publishers (San Mateo, CA), 1993.

4. M. Kunze, J. Steffers, Growing Cell Structures and Neural Gas - Incremental neural Networks, Proc. of the 4th AIHEP Workshop, Pisa, World Scientific, 1995.

5. Bogdan Sabac and I Gavat, "Speaker Verification with Growing Cell Structures", EUROSPEECH'99, Budapest, Hungary, Vol. 2, pp. 1011-1014, 5-9 September, 1999.

6. Guido Kolano, Peter Regel-Brietzmann, "Combination of Vector Quantization and Gaussian Mixture Models for Speaker Verification with Sparse Training Data", EUROSPEECH'99, Budapest, Hungary, Vol. 3, pp. 1203-1206, 5-9 September, 1999.

7. Bogdan Sabac, Gavat I., "Vector Quantization with Growing Cell Structures Applied in Speaker Verification", EUFIT'99, Aachen, Germany, pp. 314-317, September 13-16, 1999.

8. Bogdan Sabac, I. Gavat, "Speaker Identification Using Discriminative Centroids Weighting - A Growing Cells Structure Approach", Text, Speech and Dialog (TSD'99) International Workshop, Plzen, Czech Republic, pp. 171-174, September 13--17, 1999.

9. Bogdan Sabac, I. Gavat, "Speaker Identification Using Discriminative Features Selection - A Growing Neural Gas Approach", International Workshop Speech and Computer (SPECOM'99), Moscow, Russia, pp. 212-215, 4-7 October 1999.

10. Bogdan Sabac , Gavat I., "The Growing Neural Gas Algorithm Applied in Speaker Identification", to appear in ICSPAT'99, Orlando, Florida, USA, Nov. 1-4,1999.