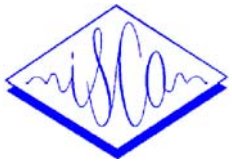

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# SPEAKER VERIFICATION WITH GROWING CELL STRUCTURES

*Bogdan Sabac and Inge Gavat*
Polytechnic University of Bucharest
Aleea Faurei, 8/11, cod 78409, Bucharest, Romania
sbogdan@helix.elia.pub.ro, inge@helix.elia.pub.ro

10.21437/Eurospeech.1999-247

## ABSTRACT

We present a new self-organizing neural network which performs unsupervised learning and can be used for vector quantization. The main advantage over existing approaches, e.g., the Kohonen feature map, is the ability of the model to automatically find a suitable network structure and size. This is achieved through a controlled growth process which also includes occasional removal of units. The algorithm is evaluated on a database that includes 25 speakers each of them recorded in 12 diffrent sesions. The overall performance was 99.5%. That is, in 99.5% of the trials, the right speaker was correctly accepted and the impostor speaker correctly rejected.

Keywords: speaker verification, vector quantization, growing cell structures, confusion matrix.

## 1. INTRODUCTION

A speaker-recognition system attempts to recognize a speaker by his/her voice. The idea is to identify the inherent differences in the articulatory organs (the structure of the vocal tract, the size of the nasal cavity, and vocal cord characteristics) and the manner of speaking.

Two general approaches have been considered for constructing classifiers for speaker recognition systems. These can be categorized as those which use unsupervised training algorithms and those which use supervised training algorithms. Unsupervised training algorithms utilize unlabeled training data. Hence, the algorithm only considers the data for the speaker to be modeled. Speakers models based on supervised training capture the differences of the target speaker to other speakers (interspeaker variability), whereas models based on unsupervised training use a self-similarity measure (intraspeaker variability).

The vector quantization (VQ) method is based on an unsupervised training algorithm, i.e., the class label is not used. In this case, clustering is used to group the training data into its individual modes or classes. The VQ classifier can be used for speaker recognition as follows. Given the extracted feature vectors from a speaker, a codebook is constructed for that speaker. This process is repeated for all speakers in the population. For speaker verification, the test vectors are only applied to the model for the speaker to be verified. The accumulated minimum distance is computed and normalized to the number of testing vectors. This normalized distance is compared with a threshold to decide if the speaker will be rejected or accepted.

A new neural VQ classifier is introduced and evaluated for speaker verification. The new classifier is based on the growing cell structures (GCS) principle [9]. The main advantage over existing VQ approaches, e.g., the Kohonen feature map, is the ability of our model to automatically find a suitable network structure and size.

The outline of this paper is as follows. Section 2 discusses the theoretical aspects of the GCS method considered. In section 3 is presented the speaker verification setup. The experimental results are provided by the section 4, and the 5-th section presents our conclusions together with future directions.

## 2. UNSUPERVISED GROWING CELL STRUCTURES

The model we propose consists of a set A of formal units. Every unit $c \in A$ has an associated n-dimensional representative vector $w_c \in R^n$. The set W of all representative vectors is the current codebook. The GCS model has a structure consisting of hypertetrahedrons (or simplices) of a dimensionality chosen in advance [8]. A k-dimensional hypertetrahedron is special among all k-dimensional polyhedrons since it is the most simple one, having only k + 1 vertices. Examples of hypertetrahedrons for $k \in \{1, 2, 3\}$ are lines, triangles, and tetrahedrons. Some typical structures

for different values of k are shown in Fig. 1.

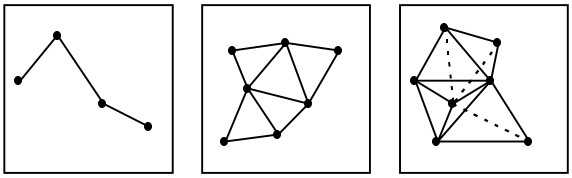


Fig.1. Cell structures of different dimensionality k.

The vertices of the hypertetrahedrons 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. Interpolation is always done among topologically neighboring units.

## 2.1 Network architecture and dynamics

The model is initialized with exactly one hypertetrahedron. During a self-organization process described further below new cells will be added to the network and superfluous cells will be removed. Every modification of the network, however, is performed such that afterwards the network consists solely of k-dimensional simplices again. After each interpolation the current codebook is adapted with a fixed number of vectors from the original data.

In principle the adaptation of the synaptic vectors in our model is done as earlier proposed by Kohonen [5]:

- Determine the best-matching unit for the current input signal.
- Increase matching at the best matching unit and its topological neighbors.

In Kohonen's model the strength of the adaptation is decreasing according to a cooling schedule. Moreover, the topological neighborhood inside which significant changes are made is chosen large at the beginning and decreases then, too. The growing cell structures model follows the same basic strategy. There are, however, two important differences [5]:

- the adaptation strength is constant over time. Specifically are used constant adaptation parameters $\varepsilon_f$ for the best matching unit and $\varepsilon_n$ for the neighboring cells, respectively.
- only the best-matching unit and its direct topological neighbors are adapted.

These choices eliminate the need to define a cooling schedule for any of the model parameters.

An adaptation step in our model can be formulated as follows:

1. Choose an input signal x from the training lot.
2. Locate the best matching unit bmu.
3. Increase matching for bmu and its direct topological neighbors

$$\Delta w_{bmu} = \varepsilon_b (x - w_{bmu}) \quad (1)$$

$$\Delta w_i = \varepsilon_n (x - w_i) \qquad (\forall\, i \in N_c) \quad (2)$$

The symbols $\varepsilon_b$ *and* $\varepsilon_n$ are adaptation constants with $\varepsilon_b \geq \varepsilon_n$. $N_c$ denotes the set of direct topological neighbors of a cell c.

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

$$\Delta E_{bmu} = (\textit{error term}) \quad (3)$$

The particular choice of the above error term depends on the application. For vector quantization one would, e.g., choose $\Delta E_{bmu} = \left\| w_{bmu} - x \right\|^2$ 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 cells 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. 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. 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. The exact re-connection procedure is actually simple enough to be described in one sentence: Let the edge which is split lead from q to a unit f then the new unit should be connected with q and with f and with all common neighbors of q and f. This is valid for an arbitrary dimension k.

Deletion of a neuron takes place if after a preset number of adaptation steps that neuron have not been a best matching unit. By insertion and deletion of neurons the structure is modified. The

result are problem-specific network structures potentially consisting of several separate sub networks.

## 3. SPEAKER VERIFICATION SETUP

Speech signal was sampled at 16 kHz with a 8 bit digitizer. The speech signals are analyzed with a 30 ms Hamming window shifted every 15 ms in order to extract the following parameters from each frame of speech discarding low energy speech frames:

(a) 20 mel frequency cepstral coefficients (MFCC). The 0.1-5 kHz frequency range was divided into 64 overlapping equal bands distributed on the Mel scale.

If we denote the output energy of the k-th. filter by $\tilde{Y}(k)$, the mel-warped cepstrum $c_{mel}(n)$ is obtained by taking the shifted discrete cosine transform (DCT) of the Mel-frequency scale (3):

$$c_{mel}(n) = \sum_{k=1}^{N_{bc}} \log\left(\tilde{Y}(k)\right) \cdot \cos\left(n \cdot \left(k - \frac{1}{2}\right) \cdot \frac{\pi}{N_{bc}}\right) \qquad (4)$$

where:

$n$=1,2,…,$L$, is the desired length of the cepstrum

$k$=1,2,…,$N_{bc}$, is the number of filters.

Because the higher order coefficients have less discriminating power and the lower order coefficients are more susceptible to channel variation, as an engineering compromise, we use a band pass liftering window. The window used is:

$$w(n) = \begin{cases} 1 + \frac{1}{2} \cdot \sin\left(\frac{n\pi}{L}\right), ...n = 1,2,...,L \\ 0, ...otherwise \end{cases} \qquad (5)$$

(b) 20 delta mel frequency cepstral coefficients (DMFCC) calculated as polynomial expansion coefficients over speech segments of five frames in length. Since the spectral transition play an important role in human perception as demonstrated by Furui [4] the introduction of such features along with the static ones will improve the recognition performance of the system.

With the extracted feature vectors from a speaker, using the growing cell structures algorithm two codebooks are constructed for that speaker. This process is repeated for all speakers in the population. After the centroids are set for each of them we compute the variance for each dimension. The linear opinion pool has been considered in our speaker verification system for the combination of features, namely cepstrum and delta cepstrum features, in order to take the verification/rejection decision.

The likelihood speaker score is computed according to equation (6) with $\alpha = 0.4$:

$$score = \alpha \cdot score_{MFCC} + (1-\alpha) \cdot score_{DMFCC}$$

$$= \frac{\alpha}{N_{MFCC}} \sum_{i=1}^{N_{MFCC}} \exp\left(-\sum_{j=1}^{20}\left(\frac{\mu_j - x_j}{\sigma_j}\right)^2\right) + \qquad (6)$$

$$\frac{1-\alpha}{N_{DMFCC}} \sum_{i=1}^{N_{DMFCC}} \exp\left(-\sum_{j=1}^{20}\left(\frac{\mu_j - x_j}{\sigma_j}\right)^2\right)$$

where: N = number of vectors MFCC or DMFCC

$\mu_j$ = mean for j dimension

$\sigma_j$ = variance for j dimension

$x_j$ = the j component of the input vector

## 4. EXPERIMENTAL RESULTS

The algorithm is evaluated on a database containing 25 speakers each of them recorded in 12 different sessions. 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. In the test phrase, all pronunciations from the recording sessions 1 to 5 where used to compute the true speaker rejection rate and impostor acceptance rate. 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 curve shown in Fig.2.

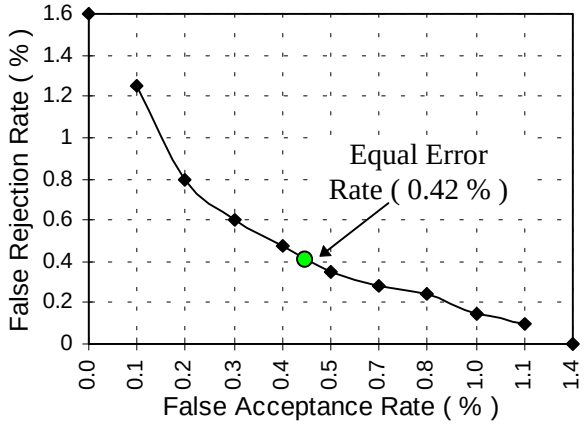


Fig. 2. Speaker verification operating curve.

Speaker verification experiments where performed using the fixed thresholds established in the test phase on the utterances form recording sessions 6 to 12. The overall performance of the system on this data set was around 99.5%.

In Fig.3. is presented a confusion matrix showing the results of our speaker verification experiment:

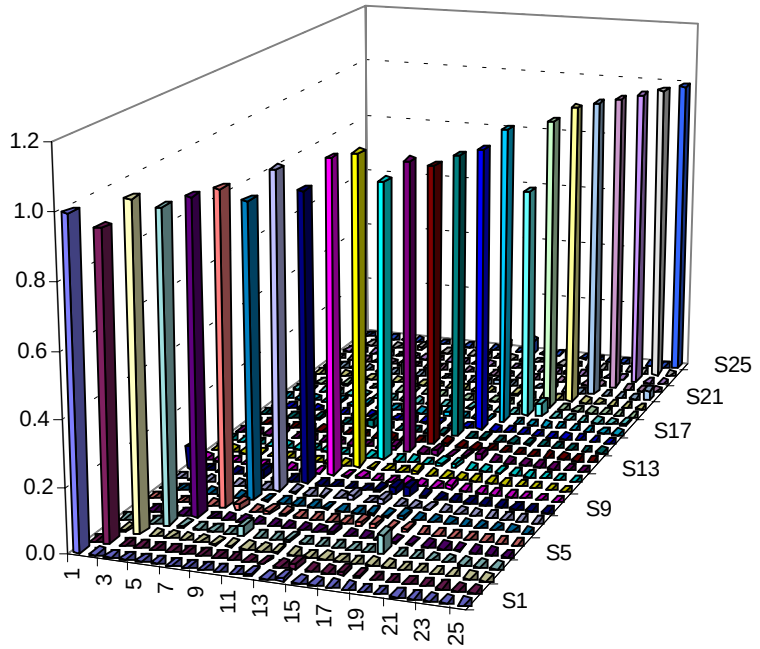


Fig. 3 Confusion matrix from the speaker verification experiment with 25 speakers. The scores on the diagonal are the "true speaker" scores. The rest are "impostor" scores. The difference shows how well the system separates them.

## 5. CONCLUSIONS

We presented a vector quantization method which incrementally builds up a codebook through interpolation. The performance of the GCS speaker verification system was evaluated for a combination of two feature sets namely MFCC and DMFCC, employing the linear opinion pool criterion, giving an overall performance of 99.5%.
In the future our attention will be focused in implementing a more powerful incremental algorithm namely the growing neural gas and also to make the pass to the next stage: text dependent speaker identification including the cohort comparison technique.

## 6. REFERENCES


[1] R.-R. Ramachandran, (1996), Robust Speaker Recognition, *IEEE Signal Processing Magazine*, Sept., pp. 58-71.

[2] T. Matsui and S. Furui (1995), Speaker Recognition Technology, *NTT Rewiev*, Vol. 7, No. 2, March, pp. 40-48.

[3] K.-R. Farell, R.-J. Mammone, K.-T. Assaleh, (1994), Speaker Recognition Using Neural Networks and Conventional Classifiers, *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, Part II, January, pp. 194-205.

[4] S. Furui (1995), Speech Recognition - Past, Present and Future - , *NTT Rewiev*, Vol. 7, No. 2, March, pp. 13-18.

[5] B. Fritzke (1993), 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).

[6] B. Fritzke (1996), Growing Self Organizing Networks - Why ? , *European Symposium on Artificial Neural Networks*, D-Facto Publishers, pp. 61-72.

[7] B. Fritzke (1993), Vector Quantization with a Growing and Splitting Elastic Net, *Proc. of the International Conference on Artificial Neural Networks*, Amsterdam, Nederlands, Sept. 13-16.

[8] B. Fritzke (1994), Growing Cell Structures - A self-organizing Network for Unsupervised and Supervised Learning, *N.N.*, 7(9):1441-1460.

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

[10] S. Segarceanu (1997), Text-dependent Speaker Identification based on VQ method, *Proc. of the second Workshop SPECOM'97*, Cluj-Napoca, Romania, 27-30 Oct., pp. 131-136.