STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LI**, Number 1, 2006

BROAD PHONETIC CLASSES EXPRESSING SPEAKER INDIVIDUALITY

MARGIT ANTAL, GAVRIL TODEREAN

Abstract. Vector quantisation and Gaussian mixture modelling methods are very popular methods for automatic speaker identification. First we give a concise overview of these methods, then present some measurements comparing them on behalf of the TIMIT corpus. The aim of this paper is to study the influence of the speech material on performances of such methods. For this purpose pure phonetic speaker models were created containing speech data from a single broad phonetic class. The speaker discriminative property of these pure phonetic speaker models had been investigated. Among the broad phonetic classes nasals and vowels were found to be particularly speaker specific.

Key Words: Speaker Identification, Gaussian Mixture Models, Pure phonetic speaker models

1. Introduction

A variety of signals and measurements have been proposed and investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face and voice. There are two main reasons for using voice instead of other measurements. First, there is a well-developed infrastructure for speech signal transmission which can be accessed almost everywhere using a cell phone. Second, speech is the most natural way of communication, therefore is not intrusive for users to provide speech sample for authentication.

Speaker recognition is the process of recognising the speaker on the basis of information obtained from speech waves. Speaker recognition can be divided into speaker identification and speaker verification. While speaker identification is a classification problem performed on a closed set of speakers, speaker verification is a binary decision, determining whether an unknown voice is from a particular enrolled speaker. If the speaker is recognised based on unconstrained speech, the system is called text-independent. However, text constrains can greatly improve the accuracy of a system. A great overview of speaker recognition systems can be found in [10].

Received by the editors: July 28, 2005.

2000 *Mathematics Subject Classification.* 68T10, 62H30.

1998 *CR Categories and Descriptors.* I.5.2 [**Pattern Recognition**]: Design Methodology– *Classifier Design and Evaluation*; I.5.4 [**Computer Systems Organization**]: Special-purpose and Application-based Systems – *Signal Processing Systems* .

The special recognition task addressed in commercial systems is that of verification rather than identification. In spite of that, for this project, we chose to confine our experiment to the task of closed-set identification rather than speaker verification. The motivation for doing so was to measure the classification capability of the system without having to consider the effect of different background model normalisation schemes required for the verification task.

The majority of research papers focuses on feature extraction and selection methods or classifier combinations for obtaining higher identification rates rather than on analysing the content of speaker models. Flanagan's group in [9] selectively used the speech spectrum for speaker identification and found that the higher portion of the speech spectrum contains more reliable idiosyncratic information on the speaker.

The aim of this paper is to investigate the discriminative properties of several broad phonetic classes in this special pattern classification problem, the speaker identification. Previous works in this field using other speaker models were done in [8, 4]. While in [4] a well-known French speech database was used, in [8] the authors worked on a private English database. No similar results were reported on TIMIT database.

Section 2 briefly presents the speaker modelling techniques. In Section 3 we review the main characteristics of the broad phonetic classes. Section 4 presents experiments on speaker identification using different types of features and models trained with broad phonetic classes. Section 5 presents statistical analysis of the content of a VQ based speaker model. Finally, in Section 6 we discuss the results and draw the main conclusions of our paper.

2. Speaker models

Over the past several years, Gaussian mixture models have become the dominant approach for modelling in text-independent speaker recognition applications [19]. However, in special cases, simpler speaker models could perform similarly well or even better. One simpler model is the vector quantisation (VQ) model, which was investigated in several papers [5, 20]. Other papers compare the GMM and VQ models drawing conclusions based on measurements on different speech databases [21, 16]. Recently the two methods have been successfully combined resulting in the VQGMM method [9].

According to Jain et al. [11], CA (Clustering Algorithm) is the organisation of a collection of patterns, represented as multidimensional feature vectors, into clusters based on similarity. Patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster. Vector Quantisation is not so much interested in finding the clusters, but in representing the data by a reduced number of elements that approximate the original data set as well as possible. We can say that in many cases CA and VQ are practically equivalent, grouping the data into a certain number of groups, so that an error function is minimised.

2.1. **Vector Quantisation - VQ.** The objective of VQ is the representation of a set of feature vectors $X = \{x_1, x_2, \ldots, x_N\} \subseteq R^D$ by a set $Y = \{y_1, y_2, \ldots, y_M\}$,

of M reference vectors in R^D. Y is called codebook and its elements codewords. VQ can be represented as a function $q : X \rightarrow Y$. The function q permits us to obtain a partition S of X constituted by M subsets S_i, where each cell S_i has the form

$$S_i = \{x \in X \quad : \quad q(x) = y_i\}, \quad i = 1, \ldots, M. \tag{1}$$

We measure the goodness of partitioning by the means of quantisation error(MQE), which can be defined as follows

$$MQE = \frac{1}{M} \sum_{i=1}^{M} D_i, \; where \; D_i = \sum_{x_j \in S_j} d(x_j, y_i) \tag{2}$$

where d is the Euclidean distance defined in R^D.

VQ can be done using different quantisation algorithms. The simplest one is the LBG algorithm [15], which was recently enhanced into ELBG in [17]. All variants of LVQ introduced by [14] can be applied equally well. All the clustering algorithms developed by Artificial Intelligence researchers might work as well. A comparison of several clustering algorithms used in speaker identification was done in [13, 2].

2.2. **Gaussian Mixture Models - GMM.** Finite mixture is a flexible and powerful probabilistic tool. Mixtures can also be seen as a class of models that are able to represent arbitrarily complex probability density functions.

For a $D-$dimensional feature vector, x, the mixture density used for the likelihood function is defined as

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{3}$$

The density is a weighted linear combination of M unimodal Gaussian densities, $p_i(x)$, each parameterised by a mean vector μ_i, and a covariance matrix, Σ_i

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-\frac{(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}{2}} \tag{4}$$

The mixture weights, w_i, satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. A GMM model can be denoted as

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, \ldots, M. \tag{5}$$

Given a collection of training vectors, the expectation-maximisation (EM) [7] algorithm can be used to estimate the model parameters. This algorithm iteratively refines the GMM parameters in order to monotonically increase the likelihood of the estimated model for the observed feature vectors.

A new parameter estimation method was proposed in paper [9]. The paper proposes to cluster the whole acoustic space into several subspaces. Within a subspace, the feature vectors are relatively more homogeneous. Each subspace is then characterised by a number of Gaussian mixture models whose parameters are

TABLE 1. Corpus division

Dataset	Utterances	Length
training	2 SA, 3 SX, 3 SI	24.5s
test	2 SX	6.06s

determined using only those relevant acoustic features belonging to the subspace. This means that feature vectors far from the subspace are not used to estimate model parameters for that subspace.

3. Phonetic classes

The basic theoretical unit for describing how speech conveys linguistic meanings is called a phoneme. Each phoneme can be considered to be a code that consists of a unique set of articulatory gestures. These articulatory gestures include the type and location of sound excitation, as well as the position of movement of the vocal tract articulators. There are some phonetic alphabets in use. European phoneticians developed the International Phonetic Alphabet (IPA), which is appropriate for handwritten transcription but its main drawback is that it cannot be typed on a conventional typewriter or a computer keyboard. Therefore, a more recent phonetic alphabet was developed by the United States Advanced Research Projects Agency (ARPA), and is accordingly called ARPAbet.

There are a variety of methods for classifying phonemes. Phonemes can be grouped based on properties related to the time waveform or frequency characteristics. A phoneme is continuant if the speech sound is produced by a steady-state vocal-tract configuration. A phoneme is non continuant if a change in the vocal-tract configuration is required during production of the speech sound. Vowels, fricatives, affricates, and nasals are all continuant sounds. Diphthongs, liquids, glides, and stops all require a vocal-tract reconfiguration during production. An exhaustive study of these classes can be found in the following books [18, 6]. In our study liquids and stops are grouped together forming the semivowels group.

Experiments were performed on the TIMIT corpus, which is phonetically segmented and annotated using the ARPAbet symbols. For broad phonetic classes we used those recommended in TIMIT corpus documentation, which are the following: Vowels, Semivowels, Nasals, Stops, Fricatives, Affricates, Silence+Closures. The speech corpus consists of 10 spoken utterances from 630 speakers covering the 8 major dialect regions of the United States. Table 1 shows the speech corpus division in training and test utterances and table 2 shows the average length of speech material for each broad phonetic class. There is no point of making speaker models from silence and we could not use the affricates, their average length were not enough to train the models.

4. Speaker Identification Experiments

All the experiments were conducted on the TIMIT speech corpus, using all 630 speakers for speaker identification.

TABLE 2. Broad phonetic classes and their training and test length

Phonetic class	Phonemes	Training length	Test length
Vowels	iy,ih,eh,ey,ae,aa,aw ay,ah,ao, oy,ow,uh,uw ux,er,ax,ix,axr,ax-h	9.66s	2.27s
Semivowels	l,r,w,y,hh,hv,el	2.38s	0.47s
Nasals	m,n,ng,em,en,eng,nx	1.34s	0.38s
Fricatives	s,sh,z,zh,f,th,v,dh	3.28s	0.95s
Stops	b,d,g,p,t,k,dx,q	1.43s	0.35s
Affricates	jh,ch	0.20s	0.09s
Silence+Closures	pau,epi,h# bcl,dcl,gcl,pcl,tck,kcl,dcl,tcl	6.28s	1.55s

Before segmenting the signal into frames, a filter was applied to enhance the high frequencies of the spectrum. We used the following filter:

$$x_p(t) = x(t) - a * x(t-1)$$

where $a = 0.97$.

The analysis of speech signal was done locally by the application of a window whose duration in time is shorter than the whole signal. This window is first applied to the beginning of the signal, then moved further and so on until the end of the signal is reached. For the length of the window we used 32ms with 22ms of overlapping between consecutive frames. Each frame was multiplied by a Hamming window in order to taper the original signal on the sides and thus reduce the side effect. After these steps we extracted cepstral parameters from each frame. In the following experiments we used two types of cepstral features, MFCC and LPCC. The detailed description of these features can be found in [3]

4.1. **VQ and GMM comparison.** For these speaker identification experiments we used LPCC features, which performed slightly better than the MFCC ones. Both VQ and GMM models were trained with 32 components. For VQ we used the LBG algorithm and the GMM models were initialized by the mean vectors provided by the LBG algorithm. The weights were set to be equal. We used diagonal covariance matrices initialised with the identity matrix. The standard ML estimation of the parameters was used with 10 iterations.

The results are summarised in Table 3. We used the training-test division presented in Table 1.

Similar results were reported for the case of GMM in [19] and for VQ in [12] , all measured on the same speech corpus and using cepstral features.

4.2. **Phonetic pure GMM.** The aim of these experiments is to determine the speaker discriminative phonetic broad classes. Table 4 summarises the identification rates obtained for the phonetic broad classes. The amount of data used for training and test is presented in Table 2. In these experiments we used 12 MFCC parameters. The number of mixture's density components were selected

TABLE 3. Speaker identification results using all the 630 speakers from TIMIT

Features	VQ-LBG	GMM
LPCC-12	97.40%	98.26%
LPCC-16	99.05%	99.05%
LPCC-20	99.30%	99.70%
LPCC-24	100.%	99.85%

TABLE 4. Speaker identification rates for 630 speakers using pure phonetic GMMs

Phoneme class	Training	Test	Mixtures	Id. rate
Vowels	9.65s	2.27s	8	95.39%
Nasals	1.34s	0.38s	1	70.31%
Fricatives	3.27s	0.95s	4	44.60%
Semivowels	2.38s	0.47s	4	41.74%
Stops	1.43s	0.35s	4	10.47%
All	24.55s	6.06s	32	96.20%

carefully, running several times the classification for different number of mixtures and selecting the one, which gives the best result.

The result obtained for the vowels is amazing. This means that using homogeneous data, which represents only 40% of the whole training data, we could almost reach the performance of the models using all training data. Another impressive result was produced by the nasals group, which represents approximately 5-6% of the whole speech data.

The results summarised in Table 4 are representative for the TIMIT database but are not comparable due to the variety of training and test time. For a correct ranking of the discriminative effects of the broad phonetic classes on speaker identification, we limited all training data to 1.5s and the test data to 0.5s for every speaker. For every classification we used a GMM with two density components. Table 5 ranks the identification rates obtained in similar training and test conditions for broad phoneme classes. We included for comparison a similar test using all phonetic classes, 1.5s training and 0.5s test data. We can see that limiting the training and test material seriously affected vowels and fricatives.

5. PHONETIC CONTENT OF SPEAKER MODELS

In this section we are going to analyse the phonetic content of VQ based speaker model. This model consists of a set of clusters. Eeach cluster is represented by its centroid, which is chosen as a prototype of its cluster. Our goal is to verify how well the broad phonetic classes are separated in this type of speaker model. We selected several speakers and made a statistical analysis of the content of their models. Our analysis has the following steps:

TABLE 5. Identification rates for 630 speakers using in average 1.5s training and 0.5s test data and a GMM model with 2 density components

Phonetic class	Id. rate
Nasals	64.92%
Vowels	20.31%
Semivowels	19.73%
Fricatives	11.42%
Stops	10.15%
All	12.38%

(1) First, we obtained the feature vectors from all audio data belonging to the selected speaker. We excluded the feature vectors for silence, because these vectors do not contain speaker specific data. Let us denote this set by $X = \{x_1,\ x_2,\ \ldots, x_T\}$, where $x_i \in R^D$and D is the dimensionality of the feature space.
(2) In the second step we labeled each feature vector with its broad phonetic class label. Let us denote by $Y = \{(x_i,\ f_i) \mid i = 1, 2, \ldots T\}$ the resulted set, where $f_i \in F$. $F = \{A,\ F,\ S,\ W,\ V,\ N\}$ is the set of broad phonetic class labels. Label A denotes the affricates, F is for fricatives, S is the symbol for stop phonemes, W and V denote the semivowels and vowels and N stands for nasals.
(3) Using these labeled feature vectors we applied the VQ algorithm and obtained the M clusters, whose content we analysed statistically.

Figure 1 shows the distribution of broad phonetic classes in the $M = \{2,\ 3,\ 4,\ 5,\ 6\}$ clusters created by the clustering algorithm. We repeated the clustering using up to 6 clusters in order to be able to capture the broad phonetic classes separation tendency. We stopped at 6, because there are altogether 6 broad phonetic classes.

The left top figure shows that if we use only two clusters, the first one will be populated by affricates, fricatives and stops and the second one by vowels, semivowels and nasals. This clustering tendency is explainable by the acoustical similarities between these broad phonetic groups. The affricates group is the least numerous one, and these phonemes are always situated in the same cluster together with the vast majority of fricatives and a representative part of the stop broad phonetic group.

Analysing the last figure, in which we used exactly six clusters, we can see that the vast majority of affricates, fricatives and nasals are concentrated in one cluster and only a small amount are spread among other clusters. Because we used all frames from phonemes, some of these frames are situated at the boundaries of phonemes, not being very representative for any broad phonetic class.

This experiment shows that a speaker model does not separate perfectly the broad phonetic classes. It is also true that every broad phonetic group has its own specific clusters. The bigger the acoustic variety inside a broad phonetic group the more clusters are spread among the feature vectors belonging to these groups.

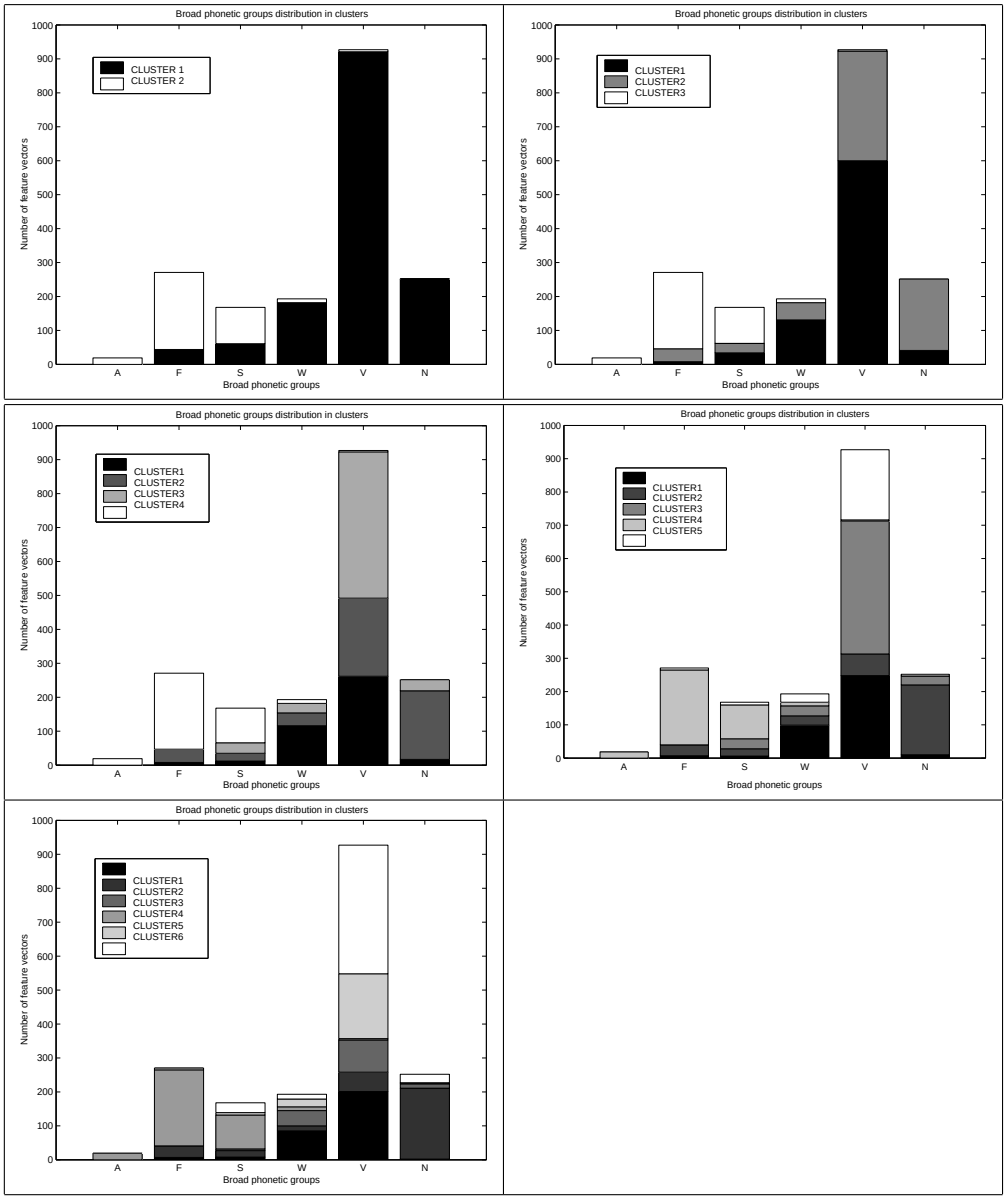

FIGURE 1. Broad phonetic classes distribution in clusters

Broad phonetic class	Variance
Affricates	23.22
Fricatives	47.35
Stops	37.85
Semivowels	46.76
Vowels	41.33
Nasals	20.53

TABLE 6. Variances of broad phonetic classes

The acoustic variety of a broad phonetic group can be characterized by the variance of the group. The higher this variance the more powerful the separating tendency of the group. We computed for each speaker model the variances of the broad phonetic classes. For this experiment we used the training part of the TIMIT speech corpus (462 speakers). Table 6 shows the average values of these variances. The higher the speaker discriminating ability of a broad phonetic group, the lower the variance of this class in a speaker model.

6. Conclusions

The main purpose of this paper was to compare the relative speaker discriminating properties of broad phonetic classes. For this purpose pure phonetic speaker models were created. We found that the pure phonetic speaker models using exclusively vowels, almost reached the performance of models using the whole speech data from a speaker. We should mention that the vowels represent 40% of the whole corpus. We also found that when pure phonetic speaker models were trained using the same amount of training data, the nasals produced the best identification rate. We can conclude that for a very good speaker model one should use speech materials which contain as much nasals as possible. Another conclusion is that the phonetic content of the training speech material is more important than its quantity. We have also studied the distribution of broad phonetic classes in the components of speaker models. We showed that the clusters of a speaker model are not phonetically pure. However, every broad phonetic class has its specific components. We showed experimentally that nasals have the best speaker discriminating ability and also the lowest variance per speaker.

Part of this work was published in [1].

References

[1] Antal, M., Toderean, G., Speaker Recognition and Broad Phonetic Groups, *Proceedings of the 3rd IASTED International Conference on Signal Processing, Pattern Recognition, and Applications*, February 15-17, Innsbruck, Austria, 2006, 155-159.

[2] Antal, M., A Comparison of Parametric Clustering Techniques used in Speaker Identification, *Proc. of the 1st International Conference on Intelligent Knowledge Systems*, Assos, Turkey, 2004, 19-25.

[3] Bimbot, F., Bonastre, J-F, Fredouille, C., Gravier, G., Margin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Gracia, J., Petrovska-Delacretaz, D., Reynolds, D. A., A Tutorial on Text-Independent Speaker Verification, *EURASIP Journal on Applied Signal Processing*, 4, 2004, 430-451.

[4] Chagnolleau,I.M., Bonastre, J-F., Bimbot, F., Effect of utterance duration and phonetic content on speaker identification using second order statistical methods, *Proc. Eurospeech*, Madrid, Spain, 1995, 337-340.

[5] Campbell, J. P., Speaker Recognition: A Tutorial, *Proc. of the IEEE*, vol. 85(9), 1997, 1437-1462.

[6] Deller, J.R., Hansen, J. H.L., Proakis, J. G., *Discrete-Time Processing of Speech Signal*(IEEE Press, 2000).

[7] Dempster, A., Laird, N., Rubin, D., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39(1), 1977, 1-38.

[8] Eatock, J.P., Mason, J. S., A quantitative assessment of the relative speaker discriminating properties of phonemes, *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Adelaide, Australia, 1994, 333-336.

[9] Lin, Q., Jan, E-E.,Che, C., Yuk, D-S, Flanagan, J., Selective use of the speech spectrum and a VQGMM method for speaker identification, *Proc. 4th International Conference on Spoken Language Processing*, Atlanta, USA, 1996,1321-1324.

[10] Furui, S., An overview of speaker recognition technology, in C-H. Lee, F.K. Soong, K.K. Paliwal (Eds.) *Automatic Speech and Speaker Recognition, Advanced Topics*, (Kluwer Academic Publisher, 1996) 31-56.

[11] Jain, A.K., Murty, M.N., Flynn, P.J., Data clustering: A review, *ACM Computing Surveys*, 31(3), 1999, 264-323.

[12] Kinnunen, T., Karpov, E., Franti, P., Real-Time Speaker Identification, *Proc. of 8th Int. Conference on Spoken Language Processing*, 2004, 1805-1808.

[13] Kinnunen, T., Kilpelainen, T., Franti, O., Comparison of clustering algorithms in speaker identification, *Proc. 4th IASTED International Conference on Signal Processing and Communications*, Marbella, Spain, 2000, 222-227.

[14] Kohonen, T., *Self Organizing Maps* (Berlin, Springer, 2001).

[15] Linde, Y., Buzo, A., Gray, R. M., An algorithm for vector quantizer design, *IEEE Transactions on Communications*, 28(1), 1980, 84-94.

[16] Matsui, T., Furui, S., Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs, *Proc. ICASSP*, San-Francisco, California, 1992, 157-160.

[17] Patene, G., Russo, M., The enhanced LBG algorithm, *Neural Networks*, 14(9), 2001, 1219-1237.

[18] Rabiner,L.R., Juang, B. H., *Fundamentals of speech Recognition*(Englewood Cliffs NJ: Prentice-Hall, 1993).

[19] Reynolds, D.A., Speaker identification and verification using Gaussian mixture speaker models, *Speech Communications* 17(1-2), 1995, 91-108.

[20] Soong,R. K., Rosenberg, A. E. , Juang, B. H., Rabiner, L. R., A Vector Quantization Approach To Speaker Recognition, *AT&T Technical Journal*, 66 (6), pp. 14-26, 1987.

[21] Stapert, R., Mason, J.S., Speaker Recognition and the Acoustic Speech Space, *Proc. Odyssey Speaker Recognition Workshop*, Crete, Greece, 2001, 195-199.

Sapientia - Hungarian University of Transylvania,Faculty of Technological and Human Sciences, 540053 Tg.-Mures, Romania, Technical University of Cluj-Napoca, Faculty of Electronics and Telecommunications, Romania

E-mail address: manyi@ms.sapientia.ro, toderean@com.utcluj.ro