SPEAKER RECOGNITION AND BROAD PHONETIC GROUPS

Margit Antal
Mathematics and Informatics Department
Sapientia-Hungarian University of Transylvania
P-ta Trandafirilor 61.
Tg-Mures, Romania
email: manyi@ms.sapientia.ro

Gavril Toderean
Telecommunications Department
Technical University of Cluj-Napoca
Constantin Daicoviciu 15.
Cluj-Napoca, Romania
email:toderean@com.utcluj.ro

ABSTRACT

The aim of this study is to provide a quantitative assessment of the speaker discriminating properties of broad phonetic groups. GMM based approach to speaker modelling is used in conjunction with a phonetically hand-labelled speech database (TIMIT) to produce broad phonetic group ranking based on speaker identification scores. The broad phonetic groups nasals and vowels were found to be particularly speaker specific. Experiments show that the homogeneity of the speech material may improve the quality of speaker identification.

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-constraints can greatly improve the accuracy of a system. A great overview of speaker recognition systems can be found in [1].

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 [2] 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 discriminatory properties of several broad phonetic classes in this special pattern classification problem, the speaker identification. Previous work in this field using other speaker models were done in [3, 4]. While in [4] a well-known French speech database was used, in [3] 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 the third section we review the main characteristics of the broad phonetic classes. The fourth section presents experiments on speaker identification using different types of features and models trained with selective phonetic classes. Finally, in Section 5 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 [5]. 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 [6, 7]. Other papers compare the GMM and VQ models drawing conclusions based on measurements on different speech databases [8, 9]. Recently the two methods have been successfully combined resulting in the VQGMM method [10].

According to Jain et al. [11], clustering algorithm (CA) is the organisation of a collection of patterns, rep-
resented 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 \( X = \{x_1, x_2, \ldots, x_N\} \subseteq \mathbb{R}^D \) by a set \( Y = \{y_1, y_2, \ldots, y_M\} \), of \( M \) reference vectors in \( \mathbb{R}^D \). \( Y \) 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

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

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

\[
MQE = \frac{1}{M} \sum_{i=1}^{M} D_i, \quad \text{where} \quad D_i = \sum_{x_j \in S_j} d(x_j, y_i) \tag{2}
\]

where \( d \) is the Euclidean distance defined in \( \mathbb{R}^D \).

\( VQ \) can be done using different quantisation algorithms. The simplest one is the LBG algorithm [12], which was recently enhanced into ELBG in [13]. 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 [15, 16].

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 \( \mu_i \), and a covariance matrix, \( \Sigma_i \)

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

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) [17] 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 [2]. The paper propose 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 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 articulator 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, 19]. 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.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Utterances</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>training</td>
<td>2 SA, 3 SX, 3 SI</td>
<td>24.5s</td>
</tr>
<tr>
<td>test</td>
<td>2 SX</td>
<td>6.06s</td>
</tr>
</tbody>
</table>

Table 1. Corpus division

<table>
<thead>
<tr>
<th>Phonetic class</th>
<th>Training length</th>
<th>Test length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>9.66s</td>
<td>2.27s</td>
</tr>
<tr>
<td>Semivowels</td>
<td>2.38s</td>
<td>0.47s</td>
</tr>
<tr>
<td>Nasals</td>
<td>1.34s</td>
<td>0.38s</td>
</tr>
<tr>
<td>Fricatives</td>
<td>3.28s</td>
<td>0.95s</td>
</tr>
<tr>
<td>Stops</td>
<td>1.43s</td>
<td>0.35s</td>
</tr>
<tr>
<td>Affricates</td>
<td>0.20s</td>
<td>0.09s</td>
</tr>
<tr>
<td>Silence+Closures</td>
<td>6.28s</td>
<td>1.55s</td>
</tr>
</tbody>
</table>

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

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 Experiments

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

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

$$x_p(t) = x(t) - a \times x(t - 1)$$

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 [10].

<table>
<thead>
<tr>
<th>Features</th>
<th>VQ-LBG</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPCC-12</td>
<td>97.40%</td>
<td>98.26%</td>
</tr>
<tr>
<td>LPCC-16</td>
<td>99.05%</td>
<td>99.05%</td>
</tr>
<tr>
<td>LPCC-20</td>
<td>99.30%</td>
<td>99.70%</td>
</tr>
<tr>
<td>LPCC-24</td>
<td>100.0%</td>
<td>99.85%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Phonetic class</th>
<th>Training</th>
<th>Test</th>
<th>Mixt.</th>
<th>Id. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowels</td>
<td>9.65s</td>
<td>2.27s</td>
<td>8</td>
<td>95.39%</td>
</tr>
<tr>
<td>Nasals</td>
<td>1.34s</td>
<td>0.38s</td>
<td>1</td>
<td>70.31%</td>
</tr>
<tr>
<td>Fricatives</td>
<td>3.27s</td>
<td>0.95s</td>
<td>4</td>
<td>44.60%</td>
</tr>
<tr>
<td>Semivowels</td>
<td>2.38s</td>
<td>0.47s</td>
<td>4</td>
<td>41.74%</td>
</tr>
<tr>
<td>Stops</td>
<td>1.43s</td>
<td>0.35s</td>
<td>4</td>
<td>10.47%</td>
</tr>
<tr>
<td>All</td>
<td>24.55s</td>
<td>6.06s</td>
<td>32</td>
<td>96.20%</td>
</tr>
</tbody>
</table>

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

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 initialised 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 [5] and for VQ in [20], 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 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
### Table 5. Identification rates for 630 speakers using average 1.5s training and 0.5s test data and a GMM model with 2 density components

<table>
<thead>
<tr>
<th>Phonetic class</th>
<th>Id. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasals</td>
<td>64.92%</td>
</tr>
<tr>
<td>Vowels</td>
<td>20.31%</td>
</tr>
<tr>
<td>Semivowels</td>
<td>19.73%</td>
</tr>
<tr>
<td>Fricatives</td>
<td>11.42%</td>
</tr>
<tr>
<td>Stops</td>
<td>10.15%</td>
</tr>
<tr>
<td>All</td>
<td>12.38%</td>
</tr>
</tbody>
</table>

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 Conclusions and Future Work

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 plan to study the distribution of different phonetic features in the components of speaker models and to implement an enhanced GMM speaker model, which eventually represents more accurately the discriminative phonetic groups of the speakers.

### 6 Acknowledgements

This research has been partially supported by Sapientia-Institute for Research Programmes.

### References


