Proceedings of the 7th International Conference COMMUNICATIONS, , June 5-7, 2008, Bucharest, Romania, pp. 73-76.

PHONETIC SPEAKER RECOGNITION

Margit Antal
Computer Science Department, Sapientia University Tg-Mures
Soseaua Sighisoarei 1C., 540485, Tg-Mures, Romania
phone: +40-265-208170, email:manyi@ms.sapientia.ro

Abstract

The aim of this study is to answer two questions regarding the use of phonetic information for speaker modelling. We formulate answers for (1) what are the discriminative powers of broad phonetic classes for the task of speaker identification? (2) Are the phonetic speaker models more suitable for speaker recognition than standard models?

Key words: Speaker Recognition, Phonetic Speaker Models

1 INTRODUCTION

Speaker recognition is the process of recognising the speaker on the basis of information obtained from speech waves. There are two problems included in speaker recognition, one is speaker identification and the other is 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. This paper deals with the problem of speaker identification. All speaker modelling techniques investigated in this paper belong to text independent speaker modelling despite of the fact that in some models we use phonetic information for creating the speaker model. Text-independent means that in the recognition stage the speakers are allowed to utter any text. Great overview of speaker recognition systems are [2, 3]

Conventional speaker recognition systems rely only on spectral features extracted from very short time segments of speech. This approach fails to capture longer-range stylistic features of a person's speaking behaviour, such as lexical, prosodic, and discourse-related habits. The use of only acoustic features is limiting because they suffer direct degradation in the presence of noise and environmental mismatch. Due to this, more recent research directions have broadened to incorporate high level features in an effort to make speaker recognition systems more robust. From a speech signal we can extract several types of features. At the lowest level, acoustic features like Mel Frequency Cepstral Coefficients can be extracted. At the next level prosodic features, such as pitch and energy contours and speaking rate can be extracted. However these types of features are more difficult to extract, they are based on theoretical constructs which are independent of acoustic noise or channel mismatch. Finally, it is possible to consider speech in terms of phonemes, words and sentences. The high level features will not provide replacement for acoustic features, but if we combine these high level features to acoustic features we would be able to improve the accuracy of speaker recognition systems. An investigation of the use of high level features is described in [9]. Doddington modelled idiolectal differences among speakers by means of word n-grams [4]. Discriminative powers of broad phonetic classes for speaker recognition were studied in paper [1]. The aim of this paper is to answer two questions regarding the use of phonetic information for speaker modelling: 1. What are the discriminative powers of broad phonetic classes for the task of speaker identification? 2. Are the phonetically structured speaker models more suitable for speaker recognition than standard models?

This paper is organized as follows. Section 2 describes the phonetic classes that were considered and the methodology used for ranking these classes. Section 3 is dedicated to the presentation of the phonetically structured speaker model including the recognition stage description too. Comparative analysis results are presented in Section 4. Finally, in Section 5 the main conclusions of this paper are drawn.

2 PHONETICALLY PURE 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. Gaussian mixture model (GMM) is the main modelling technique used for speaker recognition. Good adaptation of this model is also known [8]. Given a collection of training vectors, the expectation-maximisation 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. Unfortunately, this algorithm is very sensitive to the initial values of the parameters. There are several techniques to initialize the mean vectors. Usually a clustering algorithm is used to find good initial values for the mean vectors.

For broad phonetic classes we used those recommended in TIMIT corpus, which are the following: vowels, semivowels, nasals, stops, fricatives, and affricates. Silence and closure parts of stops were excluded from this study. Because we had a limited training speech material and there were very few affricates, we grouped affricates to the fricatives group resulting in 5 broad phonetic classes. In order to rank the discriminative properties of these broad phonetic classes, we split the training data for each speaker in 5 sets, each set containing feature vectors only from one broad phonetic group. Using these 5 training sets, for every speaker we trained 5 Gaussian mixture speaker models, one for each broad phonetic class. The proper number of Gaussians was experimentally determined for each broad class. We started by using one Gaussian and increased the number of Gaussians until the identification rate reached its maximum point for the given phonetic class. After the training stage we obtained five models for every speaker. In each test stage we used only one of these models. For example when we considered speaker models created using vowels, from the test data we used all feature vectors belonging to vowels.

3 PHONETICALLY STRUCTURED GMM

In the standard GMM based speaker recognition system each speaker is modelled by a single GMM. In our approach, we first divided the speech in five broad phonetic classes: vowels, semivowels, nasals, stops and fricatives with affricates. Then, for each broad phonetic class we collected the speech data and trained a GMM for this part of speech. Training ended by obtaining a separate GMM model for each broad phonetic class of a speaker.

A weight factor to each broad phonetic class GMM can be attached. These weights can be set to be equal or can be used the speaker discriminative power of broad classes to initialize them. Similar studies were reported in [5, 7, 10], however in these studies the number of groups in the structured GMM were selected to model more narrow phonetic groups. The majority of research papers focus on creating speaker models from the speaker specific phoneme models. These speaker specific phoneme models are adapted from speaker independent phoneme models. Our approach is different: we limit ourselves to model strictly the broad phonetic classes of a given speaker. We will denote this new model as suggested by [5] Phonetically Structured Gaussian Mixture Model (PSGMM).

The probability of a feature vector x in a PSGMM will be computed using the modified formula

$$p(x|\lambda_{PSGMM}) = \sum_{i=1}^{n} w_i p_i(x) \qquad (1)$$

where $w_i,\ i = \overline{1,n}$ are the weight factors attached to the broad phonetic groups and $p_i(x)$ represents the probability of x in the GMM attached to the ith broad phonetic group. Let us suppose that the ith broad phonetic group is modelled by a GMM having n_i Gaussian components. Then $p_i(x)$ will be computed by

Broad class	Training	Test
V	9.66	2.27
W	2.38	0.47
N	1.34	0.38
F	3.48	1.04
S	1.43	0.35

Table 1: Average length of broad phonetic training and test material; V-vowels, W-semivowels, N-nasals, F-Fricatives and affricates, S-stops

$$p_i(x) = \sum_{k=1}^{n_i} c_{ik} b_{ik}(x), \qquad \sum_{k=1}^{n_i} c_{ik} = 1 \qquad (2)$$

and b_{ik} is a Gaussian density function defined by equation

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \qquad (3)$$

The difference between the set of weights $\{w_i | i = \overline{1,n}\}$ and $\{c_{ik} | k = \overline{1,n_i}\}$ is that the latter are computed by the EM algorithm and the former can be set by us. In the next section we present two methods for setting these weights.

We should note that there is no need for phonetic labelling during recognition. Each set of feature vector is scored against each broad class GMMs and the weighted sum of these probabilities given by equation 1 is used as the probability in the given speaker model.

4 EXPERIMENTS

4.1 Corpus parameters and feature extraction

All the experiments were conducted on the TIMIT speech corpus, using all 630 speakers for speaker identification. The corpus contains 10 utterances from every speaker, each utterance having a unique identifier. There are three types of sentences denoted as SA, SX and SI sentences. Altogether there are 2 SA, 5 SX and 3 SI sentences. For training we used 8 sentences from each speaker (2 SA, 3 SX, 3 SI; 24.5s on average) and for test the remaining 2 sentences (2 SX; 6.06s on average).

Table 1 shows the average training and test data for each broad phonetic group, which were used in the experiments. In case of phonetically pure GMM, we used only the segments belonging to a broad class both for training and test.

Before segmenting the signal into frames, a FIR filter with the transfer function $H(z) = 1 - az^{-1},\ a = 0.97$ was applied. The analysis of speech signal was done locally by the application of a window whose duration

Broad class	#mixt.	Id.rate	Weight
V	8	95.39%	0.37
N	1	70.31%	0.27
F	4	44.60%	0.16
W	4	41.74%	0.16
S	4	10.47%	0.04

Table 2: Speaker identification rates for 630 speakers using phonetically pure models (PPGMM) and the corresponding set of weights

in time is shorter than the whole signal. This window is first applied to the beginning of the signal, and 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. MFCC cepstral features were used in all the experiments. (The detailed description of the extraction of this feature set can be found in [2, 11]).

4.2 Phonetically pure GMM

The aim of these experiments is to rank broad phonetic classes according to their speaker discriminative power. The optimal number of mixtures for each broad phonetic class was selected carefully. We started to model the group using one Gaussian and increased the number of Gaussians until the identification rate reached its maximum point. Table 2 summarizes the optimal number of Gaussians and their identification rates.

The last column in table 2 represents the normalized discriminative weight factors, which will be used for phonetically structured GMMs. The vowels were found to be the best broad phonetic class for speaker recognition. We have to note that the vowels represent approximately 40% of the speech corpus. The second best was the nasals class, despite of the limited amount of training and test data. The phonemes belonging to this class capture well the nasal cavity parameters, which demonstrated their speaker discriminative properties.

4.3 Phonetically structured GMM

In these experiments we used a total number of 6 Gaussians for each speaker model. In the case of phonetically structured models, vowels were modelled by 2 Gaussians and each of the other broad phonetic class by a single Gaussian. We measured identification rates for using both test parts, 2 sentences, which are altogether 6s on average, and identification rates for only one sentence, which is 3s on average. Because we had 2 sentences in the test, for the 3s test, we run the identification twice and computed the average identification rate, which is shown in the third column of table 3. Each row of the table contains the results for a certain type of speaker modelling. The first row is for the standard GMM. In the second row the results are obtained by using phonetically structured speaker models with equal weight coefficients in formula 1. The last row differs from the second one by the weight coefficients; in this case we used those obtained from broad phonetic class ranking, from the last column of table 2. Interestingly, using equal weights for each broad phonetic class produces identification performance slightly better than using fine tuned weights. The phonetically structured models did not perform better or worse than the standard GMM using the same number of Gaussians.

Model	#mixt.	Id.rate 3s	Id.rate 6s
GMM	6	93.02%	98.58%
PSGMM1	6	92.14%	98.74%
PSGMM2	6	92.06%	98.16%

Table 3: Speaker identification rates using PSGMM

Model	#mixt.	Id.rate 6s
GMM	5	97.47%
PIGMM	5	98.89%

Table 4: Speaker identification rates using usual and phonetic initialisation

4.4 GMM models with phonetic initialisation

In all the previously presented experiments Gaussian mixture parameters (weights, mean vectors, covariance matrixes) were computed using the expectation maximisation algorithm (EM). This iterative technique is very sensitive to the initial values of the parameters; therefore we used the k-means algorithm to find a set of good values for the mean vectors. Covariance matrixes were initialised with the identity matrix, and weights were initialised with equal values satisfying the constraint (sums to one).

In this subsection we present a new set of experiments, where we changed the initialisation step of the EM algorithm. Instead of using the centroids computed by the k-means algorithm, we used the broad phonetic centroids for the mean vectors' initialisation. These broad phonetic centroids were obtained by classifying each feature vector into one broad class and computing the centroids of these broad classes. In this step we used the phonetic information contained in the speech corpus. We denote this type of model by Phonetically Initialised Gaussian Mixture Models (PIGMM). Of course, this type of initialisation is suitable for models whose number of mixtures coincides with the number of broad phonetic groups. Surprisingly this type of initialisation proved to be very effective and increased the overall speaker identification rate. As we mentioned earlier we grouped the affricates to the fricatives and in this way we had 5 broad classes instead of 6. Table 4 presents the identification rates for all the 630 speakers of the TIMIT corpus using 5 mixtures using both the usual and the phonetic initialisation.

5 CONCLUSION

In this paper we presented several modifications to the standard GMM model trying to improve the speaker identification rates for a closed set of speakers. The first modification was to use only a suitable broad phonetic class. We proved experimentally that some broad phonetic classes are more speaker specific than others. For example using only vowels we obtained a very high identification rate and a very good result for fricatives although these are not very frequent in usual speech. The second modification was to combine the phonetically pure speaker models in order to achive a better identification rate. We tried two sets of weights, one uniform and one fine-tuned set of weights but these modifications did not improve the speaker identification rate of the system. The third modification was one affecting the initialisation step of the EM algorithm, namely using the broad phonetic centroids for the initialisation of the mean vectors. This proved to be very effective and increased the identification rate comparing to the same complexity GMM model.

Experiments show that it is possible to identify speakers using only a suitably selected broad phonetic class, such as vowels. The answer for the first question was given in subsection 4.2. We succeeded in ranking the broad phonetic classes due to their speaker discriminative power. Vowels and nasals produced very high identification rates. This result can be used by the designers of speech materials for speaker identification systems. The second question should be answered very carefully. Based on our experiments we can state that the broad phonetically structured speaker models are not better than standard ones. We should remark that the phonetic information can be used in various ways. Speaker models produced by the phonetic initialisation were more accurate than models produced by the standard initialisation. So we can conclude that using this limited amount of speech material, we obtained better models not by creating separate models for the different broad phonetic groups but by the more special initialisation of the model parameters.

References

[1] ANTAL, M., TODEREAN, G., Speaker Recognition and Broad Phonetic Groups, IASTED International Conference on Signal Processing, Pattern Recognition and Applications, ICSPPRA'06 Proceedings, pp. 155-159, 2006.

[2] BIMBOT, F., BONASTRE, J-F, FRESOUILLE, 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, pp. 430-451, 2004.

[3] CAMPBELL, J. P., Speaker Recognition: A Tutorial, Proc. IEEE, Vol. 85, no. 9, pp. 1437-1462, 1997.

[4] DODDINGTON, G., Speaker Recognition based on Idiolectal Differences between Speakers, Eurospeech, pp. 2521-2524, 2001.

[5] FALTHAUSER, R., RUSKE, G., Improving Speaker Recognition Performance Using Phonetically Structured Gaussian Mixture Models, Eurospeech'01 Proceedings, pp. 751-754, 2001.

[6] FERRER, L., BRATT, H., GADDE, V. R.R., KAJAREKAR, S. S., SHRIBERG, E., SONMEZ, K., STOLCKE, A., VENKATARAMAN, A., Modeling duration patterns for speaker recognition, Eurospeech'03 Proceedings, pp. 2017-2020, 2003.

[7] GUTMAN, D., BISTRITZ, Y., Speaker Verification Using Phoneme-Adapted Gaussian Mixture Models, European Signal Processing Conference, Euspico'02 Proceedings, vol. III., pp. 85-88, 2002.

[8] 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, ICSLP'96 Proceedings , pp. 1321-1324, 1996.

[9] MASON, M., VOGT, R., BAKER, B., SRIDHARAN, S., The QUT NIST 2004 Speaker Verification System: A fused acoustic and high-level approach, Proc. of the 10th Australian International Conference on Speech Science & Technology, pp. 398-403, 2004.

[10] PARK, A., HAZEN, T. J., ASR dependent techniques for speaker identification, International Conference on Speech and Language Processing, ICSLP'02 Proceedings, pp. 1337-1340, 2002.

[11] REYNOLDS, D.A., Speaker identification and verification using Gaussian mixture speaker models, J. Speech Communications 17, pp. 91-108, 1995.