DISCRIMINATIVE TRAINING AND MAXIMUM LIKELIHOOD DETECTOR FOR SPEAKER IDENTIFICATION

M. Mihoubi, G. Boulianne, P. Dumouchel

Centre de recherche informatique de Montréal (CRIM)
{mmihoubi, gboulian, pdumouchel}@crim.ca

ABSTRACT

This article describes a new approach for cues discrimination between speakers addressed to a speaker identification task. To this end, we make use of elements of decision theory. We propose to decompose the conventional feature space (MFCCs) into two subspaces which carry information about discriminative and confusable sections of the speech signal. The method is based on the idea that, instead of adapting the speakers models to a new test environment, we require the test utterance to fit the speakers models environment. Discriminative sections of training speech are used to estimate the probability density function (pdf) of a discriminative world model (DM), and confusable sections to estimate the probability density function of a confusion world model (CM). The two models are then used as a maximum likelihood detector (filter) at the input of the recogniser. The method was experimented on highly mismatched telephone speech and achieves a considerable improvement (averaging 16 % gain in performance) over the baseline GMM system.

1. INTRODUCTION

Recall that principal motivation for using a Gaussian mixture model (GMM) for text-independent speaker identification is the notion that each component density may model some underlying set of acoustic classes such as vowels, nasals, or fricatives [**?**]. But it is well known that not all acoustic classes have the same *capacity* to discriminate between speakers. The confusable classes are not the only problem affecting the performance of speaker recognition systems. Generally, the characteristics of the training and the testing environment are different, and if the test data comes from an environment not matched by the SI GMM model, the recogniser will fail to identify the claimed speaker. Thus feature extraction of a good set of acoustic parameters is the key to guarantee high accuracy recognition. For this purpose, feature selection and feature extraction methods are widely used. For example, linear discriminant analysis (LDA), with long-term parameters, was applied successfully to a speaker identification problem [**?**]. A variant of LDA, termed confusion discriminative analysis (CDA), was applied to a speech recognition task on a state-based feature space selection with hidden Markov models (HMM) [**?**, **?**, **?**]. The difference between LDA and CDA is that CDA attempts to identify specific confusion data for each state while LDA attempts to do that on a global basis.

Feature extraction with CDA [**?**] uses two approaches: a Viterbi based recogniser to generate multiple sentence hypotheses and a frame by frame Viterbi recogniser. For the first approach, the decision is made based on a comparison of likelihoods between the current hypothesis and correct transcription alignment, and for the second approach on rank orders within an N-Best framework. A multiple sentence hypotheses generation is used to ensure that confusable sections will be gathered. In both cases, the experiments have been performed using different thresholds to capture a sufficient amount of confusion data.

Two major problems can be identified in this approach. First the recognition on training data produces generally a small fraction of confusable frames which are not sufficient to robustly estimate the confusion distributions. Second, attempting to obtain more confusion data by using thresholds have led to a quiet poor classification. Also, these discriminative training schemes were applied to speech recognition task based on HMM, and thus cannot be used for comparison with our approach.

In our method, the confusable sections of speech are trimmed from a subset of the trainig data not used in the estimation of the model parameters. This avoids the use of free parameters (thresholds) and guarantees a sufficient amount of confusion data to train (or adapt) the confusion distributions. Furthermore, our approach attempts to search for the confusable sections between the different speakers instead of the confusable sections belonging to the same speaker as in the cited works. It should be noted that performing recognition on a frame by frame basis or on a short-term basis leads to a very expensive computation. Morever a short-term of speech segment (e.g., about 10 ms) cannot capture the essential information about the speaker [**?**]. To overcome these problems, we have concatenated several short-term analysis vectors into a longer one leading to a large sample classification rule. This approach enables us to use

a model with multiple mixture components in the gathering process.

2. DISCRIMINATIVE AND CONFUSION MODELS

As we mentioned in the introduction, our aim is to divide the space of training observations, for each speaker, into two mutually exclusive regions R_1 and R_2. The region R_1 will contain all the sequences coming from the considered speaker, and the region R_2, all the sequences coming from other speakers of the group. Two approaches for gathering confusion data from a recognition pass of the training data are proposed: one based on the discrimination of the speech segments between speakers, and the second, on the discrimination of the speech segments between each speaker and a background model.

2.1. Inter-speaker discrimination

The problem of testing between s hypotheses can be stated as a classification problem [**?**]. Let $S_1, ..., S_s$ be s speakers with models pdf $p_1(X/\lambda), ..., p_s(X/\lambda)$ respectively. Suppose that independent D-variate acoustic observations $X = \{x_1(s), ..., x_D(s)\}$ are available for each speaker j, $j = 1, ...s$. Our aim is to divide the space of observations into mutually exclusive regions $R_1, ..., R_s$. If an observation falls into R_i we can conclude that it comes from S_i.
Let the misclassification cost of deciding that an observation comes from S_i as coming from S_j be $c_{(j|i)}$. Let q_i be the a priori probability of drawing an observation from speaker S_i with density $p_i(X/\lambda)$, $i = 1, ...s$. The minimum risk decision rule for known parameters λ, is to assign X to R_k or S_k if

$$\sum_{\substack{i=1 \\ i\neq k}}^{s} q_i p_i(x) c_{(k|i)} < \sum_{\substack{i=1 \\ i\neq j}}^{s} q_i p_i(x) c_{(j|i)}$$

where $j = 1, ..., s$. Suppose further that all misclassification costs are equal, then the rule becomes: classify X to R_k if

$$\sum_{\substack{i=1 \\ i\neq k}}^{s} q_i p_i(x) < \sum_{\substack{i=1 \\ i\neq j}}^{s} q_i p_i(x)$$

Substracting $\sum_{i=1 i\neq k}^{s} q_i p_i(x)$ from both sides of the previous expression, we obtain

$$q_j p_j(x) < q_k p_k(x) \quad \text{or} \quad \frac{p_k(x)}{p_j(x)} > \frac{q_j}{q_k}$$

and the observation X is in R_k if k is the index for which $q_i p_i(x)$ is a maximum; that is, S_k is the most probable speaker. If we denote $\frac{q_j}{q_k}$ by T_{kj}, the expression becomes $\frac{p_k(x)}{p_j(x)} > T_{kj}$; where T_{kj} is some threshold. If we want the two probabilities of error to be equal, we should set $T_{kj} = 1$. We obtain *the log-likelihood ratio* rule.

$$\log \frac{p_k(x)}{p_j(x)} > 0$$

If an observation vector X comes from the speaker S_k (e.g., drawn from $p_k(x)$) and $p_k(x) > p_j(x)$, we can conclude that X is a discriminative segment otherwise X is a confusable segment. Applying the maximum likelihood rule, for each speaker S_i, we divide all the training sequences into two sets: a discriminative set X_d and a confusable set X_c. The confusable segments from all speakers are used to estimate a confusion world model (CM) distribution $p_c(X/\lambda)$, and the discriminative segments are used to estimate a discriminative world model (DM) distribution $p_d(X/\lambda)$. The discriminative model and confusion model are used to build a maximum likelihood detector (filter). The use of this filter is explained in the next section.

2.2. Speaker-background discrimination

To estimate a background model, the training data is pooled across a large pool of 272 speakers containing 140 male speakers and 132 female speakers from the Switchboard Corpus. We have used 10 sec of speech from each conversation side, and all the speakers are different from those used to train the system. The confusable segments are discarded from the training data and used to update the confusion model. The classification procedure is performed using the same principle as above.

$$\log \frac{p_k(x)}{p_{backg}(x)} > 0$$

This approach is less time consuming, because we do a pairwise comparison only.

3. MAXIMUM LIKELIHOOD DETECTOR

The confusable sections of speech are gathered from the testing data with the help of a maximum likelihood detector (filter) which is placed at the input of the recogniser. The filter will send to the recogniser only the observations matching the distribution function of the discriminative world model(DM). In others words, we impose to the test data to fit the characteristics described by the discriminative model. The trimming process is obtained as follow:
Let $X = \{x_1(s), ..., x_D(s)\}$ the testing observations coming from the speaker s. And let $p_d(X/\lambda)$ and $p_c(X/\lambda)$ be the probability distribution function of the discriminative model an the confusion model respectively. Let $\boldsymbol{\mathcal{D}}$ be the hypothesis that the segment $X = \{x_1, ..., x_n\}$ is drawn from the pdf $p_d(X/\lambda)$ and $\boldsymbol{\mathcal{C}}$ be the hypothesis that the same segment is drawn from the pdf $p_c(X/\lambda)$. We can state the

detection problem as the problem of testing between $D : X \sim p_d(X/\lambda), X \in R_D$ and $C : X \sim p_c(X/\lambda)$, $X \in R_C$ where R_D and R_C are non-empty sets which partition the parameter space into two disjoint regions ($R_D \cap R_C = \phi$). The observation X is in R_D if $\frac{p_d(x)}{p_c(x)} > T$; where T is some threshold. If $T = 1$, we obtain the *log-likelihood ratio* rule.

$$\log \frac{p_d(x)}{p_c(x)} > 0$$

Then, for each speaker, the test utterance is divided into two sequences $X_{dtest} \in R_D$ and $X_{ctest} \in R_C$: $X_{test} = (X_{dtest}, X_{ctest})$. The sequences X_{dtest} are sent to the recogniser and X_{ctest} are discarded.

4. DATABASE

The experiments were carried out using the SPIDRE Corpus (SPeaker IDentification REsearchCorpus) which is a subset of the much larger Switchboard Corpus. We have used the 180 target conversations coming from 45 speakers (27 males and 18 females). Three conversations are used to train the models, and one conversation for testing. More specifically:

- **Baseline system.** We trained the speakers models using 1 minute of data from each of 180 conversation sides representing 45 speakers. The average amount of data per speaker is 3 minutes.
- **Discriminative system.** For each speaker, the previous data used to train the baseline system, is divided into two sets: the training set which uses 90 seconds (30 seconds from each conversation side) to train the speakers models and the classification set which uses the rest of the data (90 seconds) to gather the confusable and discriminative sequences. The estimated models are used as auxiliary models to generate the discriminative and confusion data. Finally, the discriminative set is combined with the training set to form the new training data set. The confusion data set is used to train the world confusion model. More details about the classification process are given in the section 6.
- **Test protocol.** The test data consist of two segments t_1 and t_2 of 30 seconds length extracted from the test conversation. The segment t_2 is extracted from the end of the conversation.

5. EXPERIMENTS

The speech signal was transformed every 10 ms, using a window of 25 ms, into 12-dimensional MFCC along with the corresponding first and second derivatives for a vector dimension of 36. The baseline and the discriminative systems are based on a GMM with 128 Gaussians. Each Gaussian mixture in the GMM has a diagonal covariance matrix. Let Δt and δt be the discriminative train frame length and the discriminative test frame length respectively. Experiments carried out on the discriminative system were performed with different values of Δt and δt.

	Baseline System	
	$t_1 = 30\ sec$	$t_2 = 30\ sec$
BW128-3-min	73.33	75.55

Table 1. *Baseline system. Speaker identification performance with the segments t_1 and t_2.*

	$\Delta t\ (sec)$			
	2	1	0.500	0.250
δt= 2	82.22	82.22	87.00	82.22
δt= 1	80.00	82.22	84.50	82.22
δt= 0.500	**82.22**	**84.50**	**87.00**	**84.50**
δt= 0.250	80.00	82.22	82.22	77.80

Table 2. *DM-CM detector. Speaker identification performance for different values of Δt and δt.*

5.1. Experiments with the baseline system

The results of the experiment carried out on the baseline system are reported in Table 1.

5.2. Experiments with the discriminative system

The first experiment is carried out using the inter-speaker discriminative approach and a test segment t_1 of length 30 seconds. The maximum likelihood detector (filter) is built with the pair (DM-CM). The results (Table 2) show that a considerable improvement is achieved (averaging 14 % gain in performance) over the baseline system (from 73.33 % to 87 %) for Δt = 500 ms and δt = 500 ms. The results for different values of Δt and δt are also compiled and shown in Table 2. The second experiment is carried out using the second method, when the background model is used as a confusion model. We have considered two cases: the case BM, where the background model is the same as in the training classification processs and the case $UPBM$, when the background model is updated with the confusable frames. The results (Table 3) are given for Δt = 2 sec, and different values of δt. The method performs worse thant the inter-speaker discrimination, however, an appreciable gain is achieved (averaging 11.00 % gain in performance) over

	$\Delta t = 2sec$	
	BM	UPBM
δt= 2	77.80	80.00
δt= 1	77.80	82.22
δt= 0.500	**77.80**	**84.50**

Table 3. *BM-DM detector. Speaker identification performance for Δt =2 sec, and different values of δt.*

	$\Delta t(sec)$	
	500	250
δt= 2	80.00	80.00
δt= 1	84.50	82.22
δt= 0.500	**88.90**	80.00

Table 4. *DM-CM detector with segment t_2. Speaker identification performance for Δt =500 ms and 250 ms and different values of δt.*

the baseline system. Generally the last segments of the conversation contain more information about the speaker. We carry out the same experiment using the DM-CM filter and with a segment t_2 of length 30 seconds extracted from the end of the test conversation. The results(Table 4) achieve 88.90 % of accuracy in recognition (averaging 16 % of gain in performance) over the baseline system for Δt = 500 ms and δt= 500 ms.

6. DISCUSSION

We have presented a method to discriminate features between speakers using a decision theory approach. Before we conclude, we would want to point to the following:
Classification process. As mentionned previously, a half of training data is used to train the speaker models and the second half to generate the confusion and discriminative data. The reader has to take in mind that the data set used for classification must contain the data set used for training the models. However, the very good recognition performance on the training data (100 % of the training utterances were recognised for Δt = 2, 1 sec and 95 % for Δt = 500 ms) has led us to exclude the training data from the classification set for the mentioned values of Δt.
Amount of classification data. The model parameters have been estimated using the MLE principle. The computation, and probably the performance, can be improved by the use of small amount of data and adaptation methods such as MAP or MLLR.

7. CONCLUSION

This paper has discussed the use of inter-speaker discriminative training for GMM-based speaker identification system. It has been shown that a significant gain in performance can be obtained for highly mismatched telephone speech (averaging 16 % of gain in performance) over the baseline system . We conclude that frames of length 500 ms are more discriminative between speakers when the component mixture is used to compute the likelihood ratio. We conclude also that the good results obtained with large sample reclassification would motivate investigation of small sample reclassification process (in order of syllabic length) using mixture components. We intend to further investigate this issue and other modifications in a global intra and inter-speaker discriminative framework to improve generalisation performance.

8. REFERENCES

[1] D.A. Reynolds and C. Rose,"Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *IEEE Trans. On Speech And Audio Processing*, vol. 3, no 1, January 1995.

[2] S. van Vuuren and H. Hermansky. "Data-driven design of rasta-like filters". In EUROSPEECH, volume 1, pages 409– 412, Rhodes, 1997.

[3] C. Tadj, P. Dumouchel, M. Mihoubi, P. Ouellet,"Environment Adaptation and Long Term Parameters in Speaker Identification". *Proc. EUROSPEECH-1999*, vol. 2. Budapest, Hungary, September 1999.

[4] P.C. Woodland and D.R. Cole. Optimising Hidden Markov Models using Discriminative Output Distributions. *Proc. ICASSP*. Vol. 1, pp. 545-548. 1991.

[5] D. Povey & P.C. Woodland. Frame Discrimination Training of HMMs for Large Vocabulary Speech Recognition. *Proc. ICASSP'99*,pp. 333-336, Phoenix.

[6] J.C. Leggetter. *Improved Acoustic Modelling for HMMs using Linear Transformation*. PhD thesis, Cambridge University, 1995.

[7] H. Hermansky. Exploring temporal domain for robustness in speech recognition. In Proceedings of the 15th International Congress on Acoustics, pages 61– 64, Trondheim, Norway, 1995.

[8] T.W. Anderson. An Introduction to Multivariate statistical Analysis. Wiley publications in statistics. New York, 1958.