

# TRANSFORMATION ENHANCED MULTI-GRAINED MODELING FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Upendra V. Chaudhari, Jiří Navrátil, Stéphane H. Maes, and Ramesh Gopinath

IBM T.J. Watson Research Center  
Rt. 134, Yorktown Heights, NY 10598  
Email: uvc@us.ibm.com

## ABSTRACT

We describe our formulation of transformation enhanced data modeling used to develop a multi-grained data analysis approach to text independent speaker recognition. The broad goal is to address difficulties caused by sparse training and test data. First, our development of maximum likelihood transformation based recognition with diagonally constrained Gaussian mixture models is detailed. We give results to show its robustness to decreasing training data. Then using these models as building blocks, a multi-grained model structure is developed. For this, the training data must be labeled, e.g. with an HMM based phone labeler. A graduated phone class structure is then used to train the speaker model at various levels of detail. This structure is a tree with the root node containing all the phones. Subsequent levels partition the phones into increasingly finer grained linguistic classes. We demonstrate the effectiveness of the modeling with identification and verification experiments.

## 1. INTRODUCTION

Reliable authentication based on speech must be flexible enough to allow conversational speech input [3] [4]. i.e. it must be text-independent. Because it is impractical to collect a large amount of training data, we cannot properly create voice-print models for all of the possible usage scenarios. We address this by first improving the general robustness of our basic models, via a feature space transformation and then subsequently use them to build multi-grained models that allow a detailed phonetic analysis where possible and a natural back-off otherwise. We describe novel methods both for constructing individual speaker discriminant functions as well as for deriving target dependent background discriminant functions for use in verification. Results are given for both identification and verification comparing the performance of our system to standard techniques. We show the benefits of our basic formulation as the training data is gradually reduced. Then we give verification results based on multi-grained modeling.

## 2. SPEAKER RECOGNITION FRAMEWORK

Our initial features are a set of vectors,  $\{\mathbf{x}\}$  in  $R^n$ , containing 19 Mel-frequency cepstral coefficients (MFCC) computed using 24 filters, with delta parameters concatenated. Further, we use cepstral mean subtraction for robustness.

### 2.1. Gaussian Mixture Density Model

We describe the standard Gaussian Mixture Model (GMM) formulation [7]. Given training data from a speaker  $j$ , define the GMM  $M^j$  to be the statistical model, parameterized by  $\{\mathbf{m}_i^j, \Sigma_i^j, p_i^j\}$ , where one has respectively the Maximum likelihood (ML) estimates of the mean vector and covariance matrix along with the weight for each component  $i$  of the mixture induced by clustering. For stability the clustering is done using a speaker independent initial seed. Then, the expectation maximization (EM) algorithm is used to optimize the speaker model parameters. Constraints on the amount of data available to train a target makes it necessary to use diagonal covariance models.

#### 2.1.1. Discriminant Function

Mathematically, both identification and verification can be formulated as hypothesis tests. The basic discriminant is  $\log P(\mathbf{X}|M^j)$ , the log likelihood function which can be written as

$$\log P(\mathbf{X}|M^j) = \sum_{\mathbf{x} \in \mathbf{X}} \log \left[ \sum_{i=0}^{N-1} p_i^j p(\mathbf{x}|\mathbf{m}_i^j, \Sigma_i^j) \right] \quad (1)$$

where, the number of mixture components is  $N$ . Restricting to diagonal covariances, the component likelihoods are

$$p(\mathbf{x}|\mathbf{m}_i^j, \text{diag}(\Sigma_i^j)) = \frac{1}{(2\pi)^{n/2} |\text{diag}(\Sigma_i^j)|^{1/2}} \times e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i^j)^t \text{diag}(\Sigma_i^j)^{-1}(\mathbf{x}-\mathbf{m}_i^j)}. \quad (2)$$

For completeness, we mention that  $M_d^j$  is used to denote the diagonal version of the model.

### 3. TRANSFORMATION BASED RECOGNITION FRAMEWORK

In this section we describe our enhancement of the approach to pattern recognition based on the Gaussian mixture statistical model. The resulting formulation will be the basis for the multi-grained model structure (Section 4).

#### 3.1. Feature Transformation

The scarcity of training data used to construct  $M^j$  implies poor covariance estimates for some of the components. Thus, we restrict to diagonal covariances. However, when modeling data with diagonal Gaussians it can be shown that the unmodified cepstral feature space is sub-optimal [2], from the maximum likelihood point of view, in comparison to one which can be obtained via an invertible linear transformation. Thus, we make the restriction by choosing a pattern specific maximum likelihood linear transformation that gives the feature space which reduces the loss in likelihood caused by the restriction to the diagonal. The transform takes on significance due to the fact that we are using Gaussian mixture models *and* the transformation will apply to more than one mixture component. Single component modeling would be the same as full covariance modeling [2]. By this method, the data corresponding to mixture components with good covariance estimates contributes most to the determination of the transformation, which is applied to all mixture components.

#### 3.2. Transform Enhanced Building Block

We start with  $M^j$  as defined in Section 2.1. Next, a transformation  $\mathbf{T}^j$  is obtained for each speaker  $j$  by performing the optimization described in Section 3.2.1. Note that this requires a definition of a mixture component membership function  $I_j(i, \mathbf{x})$  (see the next section for an explicit definition) which is readily available after the clustering. For the transform enhanced system, the voice-prints are given by  $M^j$  and  $\mathbf{T}^j$ .

##### 3.2.1. Transformed Discriminant Function

Let  $M_{\mathbf{T}^j}^j$  represent the transformed model  $\{\mathbf{T}^j \mathbf{m}_i^j, \mathbf{T}^j \Sigma_i^j \mathbf{T}^{j,t}, p_i^j\}$ . Note that both the means and covariances are affected, but not the weights. The log likelihood becomes

$$\log P(\mathbf{X}_{\mathbf{T}^j} | M_{d, \mathbf{T}^j}^j) = \sum_{\mathbf{x} \in \mathbf{X}} \log \left[ \sum_{i=0}^{N-1} I_j(i, \mathbf{x}) \frac{1}{(2\pi)^{n/2} |\text{diag}(\mathbf{T}^j \Sigma_i^j \mathbf{T}^{j,t})|^{1/2}} \times \right.$$

$$\left. e^{-\frac{1}{2}(\mathbf{T}^j \mathbf{x} - \mathbf{T}^j \mathbf{m}_i^j)^t \text{diag}(\mathbf{T}^j \Sigma_i^j \mathbf{T}^{j,t})^{-1} (\mathbf{T}^j \mathbf{x} - \mathbf{T}^j \mathbf{m}_i^j)} \right] \quad (3)$$

where  $M_{d, \mathbf{T}^j}^j$  represents the transformed model with diagonal covariances,  $I_j(i, \mathbf{x})$  equals 1 if  $\mathbf{x}$  is assigned to class  $i$ , and 0 otherwise, and  $\mathbf{X}_{\mathbf{T}^j}$  represents the set  $\{\mathbf{T}^j \mathbf{x} : \mathbf{x} \in \mathbf{X}\}$ . Note that  $I_j(i, \mathbf{x})$  has replaced  $p_i^j$ . This step is required for the optimization described next.

The MLLT [2] transform  $\mathbf{T}^j$  is determined by maximizing the maximum value, over class restricted inputs, of equation 3 when the model parameters are chosen to be the maximum likelihood parameters.

The discriminant functions are now parameterized by a pattern specific maximum likelihood transformation in addition to the component parameters. This allows us to evaluate the relationship of data to a given model in that model's optimal feature space, given that diagonal models are used.

## 4. MULTI-GRAINED MODELS

First, all the data is labeled with an HMM based phone labeler. Next, consider a tree structure with the root node containing all the phones. The next level has seven nodes. One for each linguistic class (vowels, nasals, voiced and unvoiced fricatives, plosives, liquids), plus silence. The last level is comprised of the individual phones. Each training vector is assigned to any tree node containing its label. So the same data is seen from different viewpoints. In the sequel we refer to each node simply by an index number running from 1 to the number of nodes. The model for the root node of the tree will be called the Global model. For each node in the tree, we construct a transformation enhanced model as in Section 3.2. A multi-grained model for a speaker  $j$  is a set of node models denoted  $M_{\{\mathbf{T}\}^j}^j$ , where the braces in  $\{\mathbf{T}\}^j$  indicate the multiplicity of transformations associated with the full model. To represent the multi-grained model without transformations, we will reuse the notation  $M_d^j$ . We now discuss the discriminants used for identification and verification based on the multi-grained models.

### 4.1. Identification Discriminant Function

Given a set of vectors  $\mathbf{X}$  in  $R^n$ , the likelihood based discriminant function (PickMax [5]) for any individual target (or background) model is

$$D(\mathbf{X} | M_{d, \{\mathbf{T}\}^j}^j) = \sum_{\mathbf{x} \in \mathbf{X}} \max_{k,i} \left[ \log p(\mathbf{T}_k^j \mathbf{x} | \mathbf{T}_k^j \mathbf{m}_{k,i}^j, \text{diag}(\mathbf{T}_k^j \Sigma_{k,i}^j \mathbf{T}_k^{j,t})) \right]. \quad (4)$$

Here,  $I_j(i, \mathbf{x})$  is defined implicitly in terms of the max over  $i$  in equation 4. That is for each node  $k$ ,  $I_j(i, \mathbf{x})$  is set

to 1 for the component that has maximum probability and 0 otherwise.

The method of speaker identification is as follows. For the baseline case: Given test data  $\mathbf{X}$ , compute the discriminant based on equation 1 for each  $j$ , and let the decision be given by  $\arg \max_j D(\mathbf{X}|M_d^j)$ . The subscript  $d$  denotes diagonal models. For the transform enhanced case: Given test data  $\mathbf{X}$ , compute equation 4 for each  $j$ , and let the decision be given by  $\arg \max_j D(\mathbf{X}|M_{d,\{\mathbf{T}\}_j}^j)$ .

#### 4.2. “World” Model Discriminant Function for Verification

During training, the  $N_{BG}$  background speakers with the highest discriminant score on the target training data are selected for that target to represent the “world” model. For speaker  $j$ , let  $M_{BG}^j$  denote this set.

A verification claim consists of an identity claim  $j$  along with validation data. Given these, the world model discriminant score is a weighted combination of the individual scores for each model in  $M_{BG}^j$ .

$$D_{BG}(\mathbf{X}|M_{BG}^j) = \sum_{M^{bg} \in M_{BG}^j} w_{M^{bg}}^j D(\mathbf{X}|M_{d,\{\mathbf{T}\}_{M^{bg}}}^{bg}) \quad (5)$$

where  $w_{M^{bg}}^j$  is the weight for the background model  $M^{bg}$  for target  $j$  and  $M_{d,\{\mathbf{T}\}_{M^{bg}}}^{bg}$  indicates the diagonally restricted, transformation enhanced, version of the multi-grained background model  $M^{bg}$ . Note, that the weights were also subject to

$$\sum_{M^{bg} \in M_{BG}^j} w_{M^{bg}}^j = 1.$$

We use a non-linear weighting function deriving the weights from the actual discriminant function of the corresponding background model according to

$$w_{M^{bg}}^j = \frac{D(\mathbf{X}|M_{d,\{\mathbf{T}\}_{M^{bg}}}^{bg}) - m}{\sum_{M^{bg} \in M_{BG}^j} D(\mathbf{X}|M_{d,\{\mathbf{T}\}_{M^{bg}}}^{bg}) - N_{BG}m}, \quad (6)$$

with  $m = \min_{\sum_{M^{bg} \in M_{BG}^j} D(\mathbf{X}|M_{d,\{\mathbf{T}\}_{M^{bg}}}^{bg})}$ . This function achieves a model emphasis adaptively to  $\mathbf{X}$  and increases cohort competitiveness. This is important because the test data will vary from trial to trial making it necessary to adjust the relative importance of the individual background speakers.

## 5. EXPERIMENTAL RESULTS

First we give results on identification. The purpose is to highlight the relative robustness of the transform based enhancement to a reduction of the training data. As such we give results on the Global data model. Then we address verification with the Multi-Grained approach.

<i>system</i>	<i>female</i>	<i>male</i>
baseline	13%	10.5%
MLLT	11.5%	3.5%

Figure 1: Error rate for 7.5 seconds of training.  $N = 16$

<i>system</i>	<i>female</i>	<i>male</i>
baseline	4.5%	3%
MLLT	2%	0.5%

Figure 2: Error rate for 15 seconds of training.  $N = 64$

### 5.1. Training and Test Data

The Lincoln Lab Handset Database (LLHDB), available from the Linguistic Data Consortium, was used for the experiments described in this section. The target and imposter populations each consisted of 20 speakers (equal numbers of males and females taken alpha numerically). Training data was taken from the rainbow passages for each speaker. The background population consisted of the other speakers together with those from other telephony data sets. For each speaker, all of the TIMIT-text sentences were used for the tests, giving a total of 200 target trials. In addition, for verification, each imposter was claimed to be each target.

### 5.2. Identification

Identification is very sensitive to model quality. We demonstrate the robustness of the “building block” model to a reduction of training data. Figure 1 shows not only a big overall performance improvement when using the transformation, but also shows a greater improvement for the males than females, due perhaps to properties of the data set. Comparing figures 2 and 3, the results show that the transformation enhanced performance is comparable to the baseline performance with double the training data. We propose that the transformation indeed carries speaker specific information.

Next, we give results for speaker verification. The multi-grained models, constructed using the building blocks, are

<i>system</i>	<i>female</i>	<i>male</i>
baseline	2%	0.5%
MLLT	1%	0%

Figure 3: Error rate for  $\approx 30$  seconds of training.  $N = 64$

Config	Type of weighting			
	Unif.	Lin.	Likelihood	Pruned
Global	8.6%	7.6%	7.1%	5.9%
Multi	5.5%	4.4%	4.0%	4.1%
MLLT Global	4.0%	3.8%	3.8%	3.4%
MLLT Multi	3.1%	3.0%	3.0%	2.6%

Table 1: Equal-error rates on the LLHDB matched case

compared to the global level (as used here for id) models.

### 5.3. Verification

For the verification experiments the system was trained (with the full rainbow passages) and tested in text-independent mode using the data from the LLHDB partition as described in Sec. 5.1. The background population was created using internal large telephony database involving approximately 500 speakers.

Table 1 shows the system performance in the verification task. We make some observations. As with identification, the Pattern Specific MLLT drastically improves the accuracy, decreasing the error rates by approximately 50% for the global models and by about 20-30% for multi-grained models. The multi-grained model structure together with the PickMax scoring technique outperforms the global level model trained without phonetic structuring. The adaptive weighting (Likelihood) compares favorably to a uniform

weighting, which corresponds to averaging the individual background models to obtain the world model score, or a linear weighting. Furthermore, Pruning the set of background speakers for the likelihood weighting further improves performance.

## 6. CONCLUSION

We have presented an enhancement to Gaussian mixture model based pattern recognition, via the use of pattern specific feature space transformations. The technique is particularly suited to limited data situations that require the use of diagonal covariance matrices. The robustness of this approach to data reduced training was demonstrated. Building on this, a multi-grained approach based on a phonetic labeling of the data was developed. Often, data collection for training is performed on-line and must be relatively short. Consequently, the phonetic content of possible test speech may not be well modeled. The multi-grained modeling provides us with a refined acoustic structuring and allows for further improvements of the transform-based system. It does so in a way that incorporates a natural back-off mechanism. If fine grain models are appropri-

ate, they will be used, otherwise the more global models will prevail. As the described system is designed for text-independent speaker recognition, it is a suitable component for a number of more complex authentication systems.

## 7. REFERENCES

- [1] F. Beaufays and M. Weintraub, "Model Transformation for Robust Speaker Recognition from Telephone Data", *Proc. ICASSP97*,
- [2] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions For Classification", *Proc. ICASSP98*,
- [3] S. H. Maes, "Conversational Biometrics", *In proc. of the European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, 1999*,
- [4] J. Navrátil, J. Kleindienst, and S. H. Maes, "An instantiable speech biometrics module with natural language interface: Implementation in the telephony environment", *In proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, June 2000. IEEE*,
- [5] J. Navrátil, U. V. Chaudhari, and S. H. Maes, "A SPEECH BIOMETRICS SYSTEM WITH MULTI-GRAINED SPEAKER MODELING", *Proc. Conference for Natural Speech Processing (KONVENS2000), Ilmenau, Germany, October 2000*,
- [6] D. A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification", *EuroSpeech*, Rhodes, Greece, September, 1997.
- [7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, January 1995.