VERY LARGE POPULATION TEXT-INDEPENDENT SPEAKER IDENTIFICATION USING TRANSFORMATION ENHANCED MULTI-GRAINED MODELS

Upendra V. Chaudhari, Jiří Navrátil, Ganesh N. Ramaswamy, and Stéphane H. Maes

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

ABSTRACT

The paper presents results on speaker identification with a population size of over 10000 speakers. Speaker modeling is accomplished via our Transformation Enhanced Multi-Grained Models. Pursuing two goals, the first is to study the performance of a number of different systems within the modeling framework of multi-grained models. The second is to analyze performance as a function of population size. We show that the most complex models within the framework perform the best and demonstrate that, in approximation, the identification error rate scales linearly with the log of the population size for the described system. Further, we develop a candidate rejection technique based on our analysis of the system performance which indicates a low confidence in the identity chosen.

1. INTRODUCTION

Research in the field of Speaker Recognition covers a variety of topics related to the basic premise, which is to make a judgment on the identity of an individual who has given a speech sample. Text-independent speaker recognition places no restrictions on the content of the sample, enabling a wide variety of on-line and adaptive applications [3].

Among the many factors that can influence identification accuracy, the uniqueness of speakers and recording environment (e.g. channel and microphone) are the most significant: (1) The environment altering the spectral characteristics of speech with its negative impact on the performance, and (2) speaker uniqueness, whose extent is largely unknown, determining overlap of speaker characteristics. These factors can be seen as competing forces in that both affect the speech signal in similar ways, which means that in mismatched conditions it is difficult to attribute an error to one factor or the other.

To get a better understanding of the uniqueness factor, we conducted experiments in matched conditions with increasing population sizes, with our largest experiment involving over 10000 speakers. The intent was to study the effects of scale in the speaker identification problem. We

also developed an effective confidence measure based on an analysis of the N-best list of results that can be formulated to reject incorrect answers in approximately half of the cases without significantly compromising the correct recognition rate.

2. SPEAKER MODELS

In this work, the transform enhanced, multi-grained speaker modeling framework [1] is used. Multi-grained modeling consists of the concurrent use of fine and coarse grained models where the granularity of a model is characterized by the specificity of phonetic content in its training data. In identification, the nature of the testing data determines which granularities will be used when computing the score for a given multi-grained model. The transformation enhancement is due to the use of the MLLT [2] in carrying out speaker dependent feature space optimizations on top of the initial feature set consisting of mean normalized 19 Mel-frequency cepstral coefficient (MFCC) vectors computed using 24 filters, with delta parameters concatenated. The feature space optimization is carried out independently for each constituent model of the multi-grained speaker model.

For a speaker j, the multi-grained model is a collection of transformation enhanced Gaussian Mixture Model (GMM) units on varying levels of granularity, which are described below. The standard N-component Gaussian Mixture Model [5], denoted M^j, is parameterized by $\{\mathbf{m}_i^j, \mathbf{\Sigma}_i^j, p_i^j\}_{i=1,\dots,N}$, where one has respectively the ML estimates of the mean vector and covariance matrix, along with the mixture weight for each component i induced by clustering. For stability, the clustering is done using a speaker independent initial seed. The amount of data available to train a model unit often makes it necessary to use diagonal covariance models. However, this restriction to the diagonal can be done in a feature space, via the MLLT, that minimizes the effect of the restriction. This effect can be described as a loss in likelihood [2]. A transformation (MLLT) $\mathbf{T}^j$ can be chosen to minimize this loss and is constructed, via a gradient descent, for each model unit of the
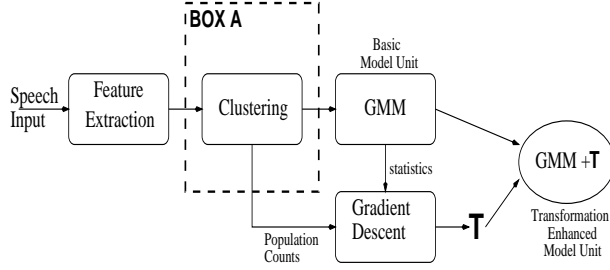
Fig. 1. Flow of model construction with transformation.

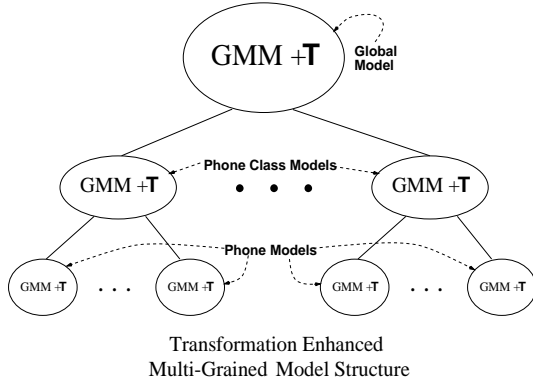Transformation Enhanced
Multi-Grained Model Structure

Fig. 2. Transformation Enhanced Multi-Grained Model.

multi-grained speaker model.

Figure 1 shows the model construction procedure for a single GMM unit with transformation. Note that the statistics of the initial GMM are used along with the cluster counts from the original data in deriving the MLLT transformation [2].

Before constructing a multi-grained model, the training data must be labeled. In our case this is done with an HMM-based phone labeler. The data is distributed into bins corresponding to each phone. Model units (GMMs) for these bins constitute the finest grained modeling in the system. Next, we consider seven linguistic classes: vowels, nasals, voiced and unvoiced fricatives, plosives, liquids, plus silence. Each phone is assigned to one of these. The bin for each phone class is filled with the data from all of the constituent phone bins. These class models make up the middle-grain level. Finally, all of the data is collected together in the root model bin. We call the GMM here the global, or coarsest-grained, model (unit). We refer to each bin, or node, simply by an index number running from 1 to the number of bins. A graphical description is shown in Figure 2. The use of the term node is appropriate given the tree structure described.

For the transform enhanced system, the constituent models are given by M^j and $\mathbf{T}^j$. We use the notation $M^j_{d,\{\mathbf{T}\}^j}$. The subscript d indicates that all of the Gaussians are diagonal and the set $\{\mathbf{T}\}^j$ denotes the fact that each constituent

model of the multi-grained model has its own transformation.

Corresponding to our transform enhanced multi-grained model, we use a modified likelihood based discriminant function. Given a set of vectors $\mathbf{X} = \{\mathbf{x}\}$ in R^n, the discriminant function (PickMax) [4] for any individual target model is

$$D(\mathbf{X}|M^j_{d,\{\mathbf{T}\}^j}) = \sum_{\mathbf{x} \in \mathbf{X}} \max_{k,i}$$
$$\left[\log p(\mathbf{T}^j_k \mathbf{x}|\mathbf{T}^j_k \mathbf{m}^j_{k,i}, diag(\mathbf{T}^j_k \mathbf{\Sigma}^j_{k,i} \mathbf{T}^{j,t}_k))\right], \quad (1)$$

where the index k runs through the nodes and the index i through the mixture components at the nodes. Speaker identification is carried out by computing equation (1) for each speaker j, and letting the decision be given by $arg \max_j D(\mathbf{X}|M^j_{d,\{\mathbf{T}\}^j})$.

3. DATA

Our experiments are based on an internal database consisting of 10013 speakers with telephone-quality speech. As for the acoustic channel properties, the data in the training and the test are matched. For each speaker, approximately 30 seconds of speech is used for training and 3 to 5 seconds utterances for test. The speech was recorded over real telephone landlines and the different speakers were talking about various topics. It has to be noted that the database is somewhat segmented in terms of spontaneous vs. read speech, which, however, plays rather a minor role in text-independent speaker modeling, as opposed to speech recognition tasks. We emphasize that the main goal of the experiments was to study identification performance as a function of the population size, and thus matched condition tests are appropriate.

4. EXPERIMENTS

The results are presented on three different systems that fall within the framework of transformation enhanced multi-grained modeling. First, we use simply the root node, i.e. the global level, of the multi-grained tree. This corresponds to modeling all of the training data with one transformation enhanced GMM. In this case, it contains 8 mixture components. Next, we use the multi-grained models based on data labeled by the HMM decoder. Note that each such model unit has 8 components. Finally, we approximate the advantages of the multi-grained approach, which is a form of hierarchical clustering, in the single GMM case by using a method called *centroid inclusive* clustering. We refer to *BOX A* in figure 1. Centroid inclusive clustering involves a modification to *BOX A* whereby after the initial clustering, the centroid of the original data is added back as a new cluster centroid. The data is then reclustered (Figure 3).
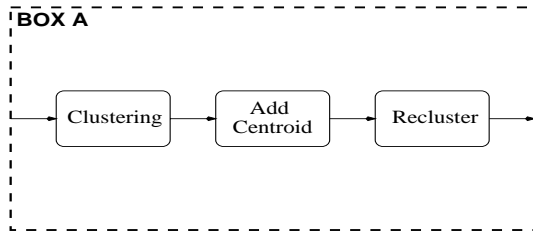
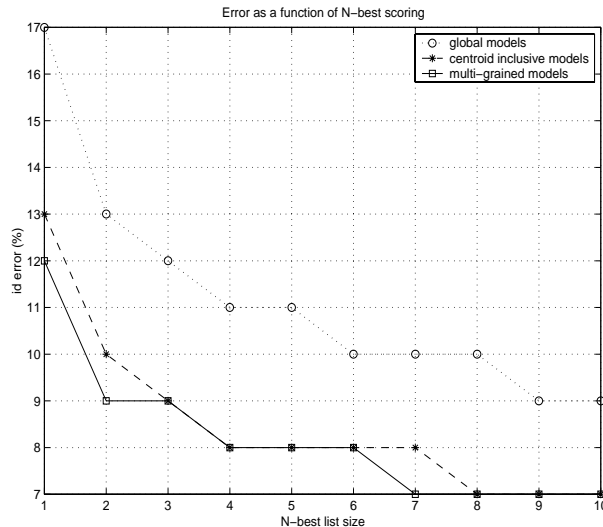**Fig. 3**. Centroid inclusive clustering.

Fig. 4. System comparisons with N-best scoring.

This is an approximation to the hierarchical model in the sense that information near the total data centroid, which would be present in the root node of an hierarchical model, is added to the global level GMM. Multi-grained modeling is not used for this experiment. The rest of the modeling is identical. The resulting model thus has 9 components as opposed to 8 in the first experiment.

For each case we present results on 10013 speakers in the form of an N-best list in Figure 4. The vertical axis is the identification error percentage. The horizontal axis indicates the size of the N-best list used in scoring. A trial was considered correct if the true identity was in the list.

From the results in Figure 4 we observe first that the multi-grained models have significantly better performance than the global model alone, which we use as the baseline. We point out that the multi-grain models do have many more Gaussians than the global model. However, this approach allows for the larger number of Gaussians without partitioning the data into too many *sparse* clusters. The centroid-inclusive clustering, though not as good as the multi-grained modeling, gave a big improvement over the global model baseline as well. The significance of this result is that it allows fast identification as compared to the multi-grained approach which can take up to M times
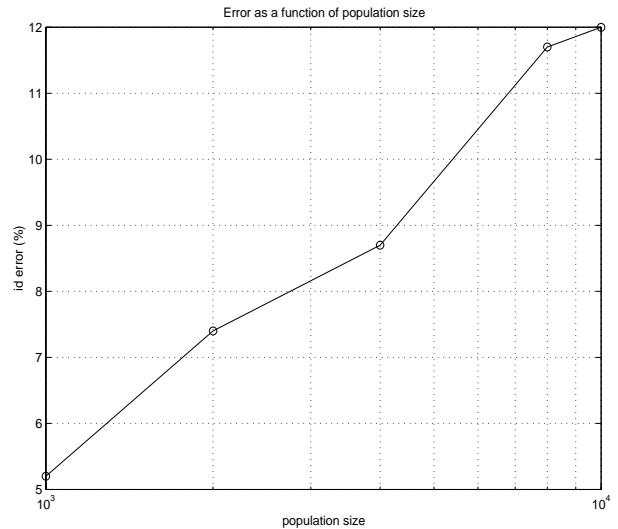

Fig. 5. Id error vs. population size (semilog).

longer, where M is the number of model units used. This is an important advantage for large population identification. Further, there is a big gain in accuracy if we allow the correct answer to be in the top 4 candidates in the N-best list. This lead us to develop a scheme to accept or reject identification results that is much simpler than the standard technique of performing a verification on the top candidate in the list (see Section 5).

4.1. Scaling with Population

For the best system, we charted the performance as a function of the population size. The plot is shown in Figure 5. Note that this is a semi-log plot with the horizontal axis indicating the log of the population size. The plot shows that the identification error rate is approximately a linear function of the log of the population size.

A priori, it did not seem unreasonable to expect a linear relationship between the population size and the identification error rate. If we estimate the performance as a linear function using the first data point, we get the result shown in Figure 6.

We give the plot to emphasize how much better the actual performance is compared to the linear estimate. Clearly, for any given model, adding more speakers to the enrolled population does not necessarily add strong competitors for the identification task. However, overall each added speaker does serve as competition for some subset of speakers. The character of the curve gives an indication of the size of this subset.

5. CANDIDATE REJECTION

Examining the results shown in Figure 4, we note that the top 10 scoring models contain a significantly larger per-
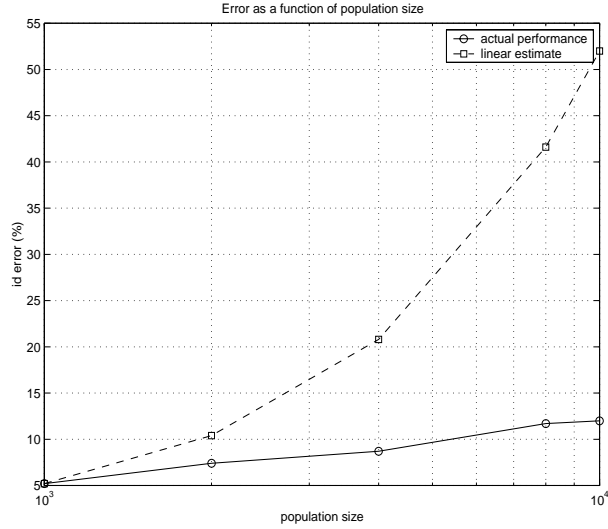
Fig. 6. Id error vs. population size (semilog).

centage of correct identities than the top scoring models. Moreover, we noted that the differentials among the top 10 scores had a different character when the top answer was correct as compared to when it was not. Based on this evidence, we have developed a confidence measure which is derived from the score differential statistics:

For a given identification test i, let

$$s_i^1, \ s_i^2, \ ..., \ s_i^N$$

be the top N scores in decreasing order. (We use $N = 10$.) Let

$$d_i^1, \ d_i^2, \ ..., \ d_i^{N-1}$$

be the score differentials. i.e.

$$d_i^j \ = \ s_i^j \ - \ s_i^{j+1}.$$

Given a set of identification trials I, we divide them into two groups. Those that have the correct model with the top score, and those that do not. Respectively, these are the accept, I_{accept}, and reject, I_{reject}, classes for the identification result. We create simple statistical models of the corresponding differential sets

$$\{d_i^1, \ d_i^2, \ ..., \ d_i^{N-1}\}_{i \in I_{accept}}$$

and

$$\{d_i^1, \ d_i^2, \ ..., \ d_i^{N-1}\}_{i \in I_{reject}}.$$

We compare the likelihood of the differential set of a given test trial with respect to the two models (with a log-likelihood ratio test) and accept or reject the result based on a threshold. Thus, the likelihood ratio serves as a confidence measure for the identification output of the system. This approach is similar to, but much simpler and faster than, that of performing a verification on the top scoring model. Using this approach, we have achieved a false alarm rate of

2.5% with a miss rate of 44.7%, meaning that approximately 55.3% of the time we correctly reject a false identification result with the consequence that 2.5% of the time we reject a correct identification result. Ideally, the N-best lists associated with the rejected trials may be re-scored with perhaps a more powerful approach to achieve a greater accuracy, which, naturally, has to be decided with respect to the computation expense to be spent on this task.

6. CONCLUSION

We have presented a number of different results on very large population speaker identification with size exceeding 10000 speakers. Comparisons among various systems revealed that our most detailed multi-grained models gave the best performance. We demonstrated performance approaching these models with our centroid inclusive approach, which has the benefit of speed, a necessity for large population identification. Further, we observed a log-linear relationship between population size and performance for the best system. Finally, a candidate rejection technique was presented with the potential to improve performance, given a re-scoring mechanism for the rejected trials.

7. REFERENCES

[1] U.V. Chaudhari, J. Navrátil, and S.H. Maes. Transformation Enhanced Multi-grained Modeling for Text-Independent Speaker Recognition. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, October 2000.

[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, 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,*

[5] 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.