

SPEAKER VERIFICATION USING TARGET AND BACKGROUND DEPENDENT LINEAR TRANSFORMS AND MULTI-SYSTEM FUSION

Jiří Navrátil, Upendra V. Chaudhari, Ganesh N. Ramaswamy

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

e-mail: {jiri, uvc, ganeshr}@us.ibm.com

Abstract

This paper describes a GMM-based speaker verification system that uses speaker-dependent background models transformed by speaker-specific maximum likelihood linear transforms to achieve a sharper separation between the target and the non-target acoustic region. The effect of tying, or coupling, Gaussian components between the target and the background model is studied and shown to be a relevant factor with respect to the desired operating point. A fusion of scores from multiple systems built on different acoustic features via a neural network with performance gains over linear combination is also presented. The methods are experimentally studied on the 1999 NIST speaker recognition evaluation data.

1. Introduction

Maintaining data security and authenticity in speech-driven telephony applications can be done effectively through speaker verification. Despite the relatively long period of research, today's acoustic verification systems still face significant challenges caused by adverse acoustic conditions. Telephone band limitation, channel/transducer variability, as well as the natural speech variability all have a negative impact on the performance.

Among the most successful approaches to robust text-independent speaker verification is the Gaussian Mixture Model (GMM) formulation employed in many state-of-the-art systems [1, 2, 3]. In [4, 1] a particularly effective way of speaker modeling via Bayesian adaptation from speaker independent models combined with a likelihood-ratio detector was introduced, allowing for robust verification using limited training data. The likelihood ratio score is calculated typically between the target and a single or composite background model. Further significant progress has been made in score normalization techniques, particularly by introducing the H-norm [1] and T-norm [2], to help stabilize the likelihood-ratio scores obtained from the claimed target and the background model with respect to a systematic handset bias and scale (H-norm) and the acoustic variability given the test utterance (T-norm).

In this paper, we describe a GMM-based verification system involving the Bayesian adaptation scheme from a universal background model (UBM) created via a fast one-pass clustering method, followed by target-specific linear transforms applied to both the target and a target-dependent copy of the UBM. The effect of the transforms is studied in connection with specific ways of calculating the Gaussian likelihoods, to improve the accuracy at specific operating points. The H- and T-norm with some modifications are also presented. Finally, we describe a neural-network based fusion of multiple systems created on different acoustic features to exploit partial error decorrelations among the individual systems allowing for performance gains over the separate systems.

2. Speaker Models

2.1. Universal Background Model

The training of the common world model - or universal background model (UBM) - as a prior in the MAP adaptation, is carried out as follows. First, several smaller (or atomic) Gaussian Mixture Models (GMM) are created from the training data partitioned by gender and a handset type, similar to [1].

For clustering the data within the atomic segments, we use a one pass technique that operates significantly faster than a typical clustering algorithm, e.g. the LBG, without showing a performance degradation. Unlike the LBG, the method described below is completely deterministic and thus gives consistent performance over multiple enrollments. The process is recursive and operates on a set of vectors to produce an N-way partition (in this paper N=2). If \mathbf{X} is the input, then \mathbf{X}_1 and \mathbf{X}_2 are the output where $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2$. For each $\mathbf{x}_i \in \mathbf{X}$, a vector of projection coefficients \mathbf{c}_i is computed with respect to the set of eigenvectors of \mathbf{X} . Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ operate on each \mathbf{c}_i to produce a set of projected points $\{\gamma_i\}$. We split $\{\gamma_i\}$ into two sets, $\{\gamma_i\}_1$ and $\{\gamma_i\}_2$, via a thresholding operation. Since each γ corresponds to an \mathbf{x} , this defines the two sets \mathbf{X}_1 and \mathbf{X}_2 . The process repeats for each \mathbf{X}_i until a degeneracy condition is met or can be carried out for a fixed number of splits. This procedure imposes a tree structure on the data. Each node of this tree can be modeled. However, in this paper, we used only the leaf nodes for further modeling.

The resulting atomic GMMs were subsequently unified into a common UBM on which the Maximum Likelihood Linear Transform (MLLT) [3], was calculated to optimize the feature space with respect the UBM parameters.

2.2. Target Modeling

Let T_W denote the MLLT transform of the world (UBM) model W_0 . The MLLT can be interpreted as transforming the mean and covariance parameters μ and Σ of the UBM in the original feature space with M mixture components, that were estimated from N training vectors X_1^N :

$$M_{W_0} = \{\mu_j, \text{diag}(\Sigma_j), p_j | X_1^N\}_{1 \leq j \leq M}, \quad (1)$$

(p_j denotes the Gaussian prior) to a new GMM M_{W_1} with parameters estimated from training vectors transformed by T_W , i.e. $T_W X_1^N$

$$\begin{aligned} M_{W_1} &= \{\mu'_j, \text{diag}(\Sigma'_j), p_j | T_W X_1^N\}_{1 \leq j \leq M} \\ &= \{T_W \mu_j, \text{diag}(T_W \Sigma_j T_W^T), p_j | X_1^N\}_{1 \leq j \leq M} \end{aligned} \quad (2)$$

where the new feature space results in the least loss of likelihood. from the diagonal covariance assumption. In other words the transform is a form of feature space optimization given the particular GMM, in this case the UBM. The MLLT matrix is obtained by solving a nonlinear optimization problem involving both the covariance and the means of the original model [3].

As the first step in the enrollment, each target speaker’s GMM is created via Bayesian, or Maximum A-Posteriori (MAP) adaptation from the UBM, carried out in the MLLT space of T_W . Using a set of target adaptation vectors Y_1^K for the speaker i , the MAP yields a target GMM M_i whose parameters μ and $diag(\Sigma)$ are now based on the feature space of M_{W_1} , i.e.

$$M_i = \{\mu_j^{(i)}, diag(\Sigma_j^{(i)}), p_j | T_W Y_1^K\}_{1 \leq j \leq M} \quad (3)$$

In a second step, using the MAP-adapted model (3), a subsequent MLLT optimization is performed resulting in a new target-specific matrix T_{M_i} which further optimizes the feature space for the particular speaker model M_i , which becomes

$$M_i = \{\mu_j^{(i)}, diag(\Sigma_j^{(i)}), p_j | T_{M_i} T_W Y_1^K\}_{1 \leq j \leq M} \quad (4)$$

in which μ and Σ are estimated in the new space (for simplicity of indexing, the notations $\mu/\Sigma/p_j$ stand for parameters estimated in the given feature space and are different from those in (3)). We now have a set of target GMM’s with target-specific linear transforms and a common UBM with its corresponding transform. As a final step of the enrollment procedure, one “copy” of the UBM is tied to each of the targets by adopting the corresponding target-specific MLLT space, i.e. a set of M_i/M_{W_i} pairs is created, in which M_i is given by (4) and the (now) target-dependent UBM M_{W_i} by

$$M_{W_i} = \{\mu_j, diag(\Sigma_j), p_j | T_{M_i} T_W X_1^N\}_{1 \leq j \leq M} \quad (5)$$

Note, that the new UBM parameters μ and Σ are obtained by linearly transforming (2) and that no re-clustering of the model is carried out, thus maintaining the correspondence between Gaussians of the UBM and the target models. Within these model pairs, the feature space is somewhat more optimized for the target GMM than for the UBM and is expected to give a higher likelihood for target trials, as opposed to impostor trials. This “unidirectionally discriminative” way can achieve a sharper target/world separation. We try to underline this hypothesis by our experiments. Further discussion is provided in section 4.

3. Verification

3.1. Likelihood Ratio Scores and Gaussian Coupling

The speaker verification task is posed as a binary hypothesis test in which the loglikelihood ratio between the target model and the UBM is used as the discriminant function [4]. Given a claimed identity i^* , the vector scores are calculated as the log of the Gaussian mixture densities on the target model (4) and its corresponding UBM (5). Experimental experience suggests that the ratios of full-mixture densities are well approximated by ratios of their maximum Gaussian components which can be calculated more efficiently. Furthermore, the effect of Gaussian “coupling” - or fixing the Gaussian component index in the UBM and the MAP-adapted target - is of interest. As suggested in [4], due to the relatively distinct correspondence between the MAP-adapted Gaussians of the UBM and the targets, it is sufficient to calculate the several (e.g. five) best components in the UBM and then proceed with target calculations only on these selected components in all other models. We study the case of selecting a single best Gaussian, however, in two different ways, namely (i) best as determined in the UBM (or target-specific UBM) and imposed in the target, and (ii) vice versa. In the case (i) a given feature vector is examined in the UBM first thus determining its general location in the acoustic space. Then this vector is compared to the same Gaussian in the MAP-adapted target model. In order to gain a positive likelihood ratio, the corresponding target Gaussian must yield a higher likelihood for that vector. In the case (ii) the vector is first matched against the

target model, the best component is selected and then compared to the corresponding one in the UBM. When the test vector originates from an impostor, it seems intuitive that case (ii) tends to be more biased towards falsely accepting an impostor sample because it is given the chance of finding the best fit within the target model and then faces only one Gaussian from the UBM for comparison. The case (i), on the other hand, will tend to have less of such acceptance. When the vector is from a true target speaker similar hypothesis about lower and higher false rejection respectively can be posed. The effect of both ways of coupling Gaussians on the performance is presented in Sec. 4.

3.2. Score Normalization

Due to the fact that the verification task involves a threshold decision acting upon the ratio score, there is a direct connection between the detection accuracy and the stability of the score across acoustic environments. To prevent increasing error rates with varying handsets and channels, several score normalization schemes were developed. Among these, most recently, the T- and H-norm [2, 1] try to compensate for acoustic score variations induced by the test utterance or by a selected set of handset-specific impostor utterances respectively. The T- and H-normalization schemes were adopted in our system with the T-norm modified as follows. Utterance-adaptive cohort model weights were included in calculation of the T-norm so as to emphasize the cohort models yielding the best likelihood given the particular test utterance

$$L_{TNORM} = \frac{L_t - \sum_i w_i L_{C_i}}{\sigma_C} \quad (6)$$

where L denotes the likelihood ratio, C_i the i -th cohort model, w_i the corresponding weight and σ_C the standard deviation calculated on the cohort scores. The weights are derived from the cohort likelihoods on the test utterance as follows

$$w_i = \frac{L_i - L_{min}}{\sum_i L_i - K \cdot L_{min}} \quad (7)$$

in which L_{min} is the minimum score within the cohort of size K . Note that the weights are nonnegative and sum up to 1. Clearly, the weighting automatically adapts to the actual utterance such that the most relevant models are emphasized in the normalization and it can be viewed as a soft-decision way for determining the cohort set on the fly. This mechanism allows for a stable T-normalization in cross-gender trials which are often disregarded in cases of gender-matched cohorts.

The H-norm parameters, i.e. the mean and standard deviation of impostor scores, were calculated from a collection of impostor trials on individual target models similar to [1].

3.3. Final Classifier for Multiple Components

In this paper we also present a score-level combination of multiple systems by means of neural networks. A combination of several systems that are trained on different acoustic features exploits the potential of partially decorrelated outcomes of the individual systems due to differences in the acoustic speaker representation. In this context, neural networks represent a favourable alternative to linear combinations because they allow for modeling nonlinear dependencies between the system scores.

The structure of the neural network used in our experiments is shown in Fig. 1. The network has six inputs. Recall that for each system, an input test cell produces two scores, the target score and the UBM score. In our experiments, we have observed that letting these scores be separate inputs into the network gives better fused performance than if the log likelihood

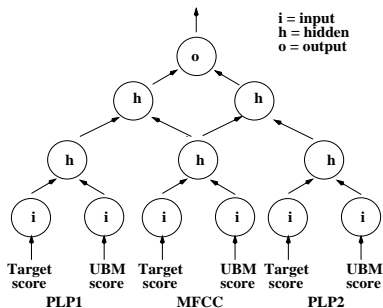


Figure 1: Flow of neural net score fusion.

ratio (subtracting the UBM score from the target score) was directly input.

4. Experiments

4.1. Database

The performance of the described text-independent verification system was experimentally evaluated on the Switchboard telephone corpus, namely using its two segments, selected for the NIST speaker recognition evaluations in 1996 and 1999 [5]. The Switchboard corpus contains a large number of speakers involved in topic-elicited conversations recorded in two separate channels, one for each conversation side. The recordings used in the NIST evaluations are accompanied by side information, such as gender and the likelihood of the handset transducer type for a conversation side (two handset types were distinguished: electret and carbon button). The 1996 data is a part of the Switchboard I corpus and, in our experiments, served as basis for creating the universal background model (UBM) as described in 2.1. A total of approximately 4.5 hours of speech was first partitioned into four parts by gender and the two handset types, from which the atomic GMMs, each with 256 (or 512) components were trained. These were then joined into a common 1024 (or 2048) component GMM forming the UBM. The 1999 data (Switchboard II, Phase 3), was used for training and testing the targets. This corpus, as specified by NIST, contains 2 minutes (2 1-minute sessions) of enrollment data for 539 speakers and provides for 37620 gender-matched verification trials. The test duration varies between a few seconds and one minute with the majority of tests falling into a range between 15 to 45 seconds. The ratio of target vs. impostor trials among the 37k tests is roughly 1:10. Furthermore, between 100 to 300 30-sec utterances of each handset from the 1996 set were also used for estimating the the H-norm parameters and 40 speaker models were built for the purpose of obtaining cohort scores in the T-norm calculation.

4.2. Features

Based on the same principles described in the previous sections, three separate systems were created differing only in the type of feature extraction using 1) 19-dimensional Mel-Frequency Cepstral Coefficients (MFCC) and their first derivatives, 2) 12-dimensional Perceptual Linear Prediction coefficients (PLP), and 3) 19-dimensional PLP coefficients and their first derivatives. The feature vectors were extracted every 10 ms and their derivatives calculated using the left and right context of two frames. Whereas the MFCC system served for primary investigations, the two additional PLP systems were used to study the potential of combining multiple systems on score level.

System	opt. DCF 10^{-3}	
	1024 GMM	2048 GMM
Plain	52.2	51.8
MLLT	48.4	50.9
MLLT cpl	46.9	48.5
MLLT-xUBM	47.7	48.0
MLLT-xUBM cpl	45.5	45.3

Table 1: Optimum detection costs for the 1024- and 2048-Gaussian MFCC systems

4.3. Evaluation Measure

Evaluating the system accuracy as a trade-off between the two types of detection error (false alarm, false miss) was based on Detection Error Tradeoff (DET) curves [6], which are a particular case of the Receiver Operating Characteristics. On the DET curve typically two specific operating points may be of interest, namely the Equal-Error Rate (EER) and the point having the lowest detection cost, as defined by specific cost factors C_{FA} , for false alarms, and C_{Miss} , for false false misses, as well as by the expected relative frequencies (priors) of target and nontarget tests, $P(T)$ and $P(N)$. For the 1999 evaluation, the NIST’s detection cost function was defined as [5]:

$$DCF = C_{FA} \Pr(FA|N) \Pr(N) + C_{Miss} \Pr(Miss|T) \Pr(T),$$

with $C_{FA} = 1$, $C_{Miss} = 10$, $\Pr(N) = 0.99$, and $\Pr(T) = 0.01$, thus shifting the point of interest towards low FA rates. It turns out to be important which operating point is of primary interest as different techniques may bring different gains dependent on the region of the curve.

4.4. Results

4.4.1. Single MFCC System

The performance of the MFCC GMM system in terms of the optimum detection cost function (DCF) for all 37k trials and two GMM sizes: 1024 and 2048 components, is shown in Table 1.

The plain configuration in the first row refers to MAP-adapted GMM without any linear transforms. Consistent improvements can be observed by adding the MLLT transforms to the common UBM and to the target models separately (“MLLT”) for both GMM sizes, whereby only marginal differences can be seen between 1024 and 2048 Gaussians. Tying (coupling) the Gaussians of the target and the UBM used in the likelihood calculation¹ brings gains in both cases the “MLLT cpl” and “MLLT-xUBM cpl”, the gain in the latter case being slightly higher. This observation goes along with the hypothesis that the target region in the native-MLLT feature space is more sharply separated from the “world” region in that same MLLT space. The effect of the way of Gaussian coupling can be better understood from the DET plot in Fig. 2, in which two curves with a comparable EER but quite different minimum DCF values are shown: The solid line for the case when determining the component in the UBM first and imposing on the target (U-T), and the dashed line for the opposite case of picking the Gaussian in the target GMM first (T-U). U-T apparently rotates the DET curve towards lower miss probability in the lower false alarm region, thus towards lower DCF values. In the T-U case (dotted line) the curve seems to have the opposite character to that of the

¹The coupling in the table was done in the UBM→Target direction, i.e. the Gaussian component was selected in the UBM and imposed on the target

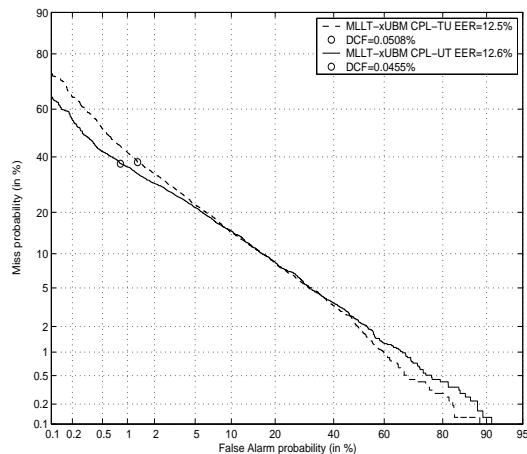


Figure 2: The effect of the two ways of Gaussian coupling: U-T (solid) and T-U (dashed).

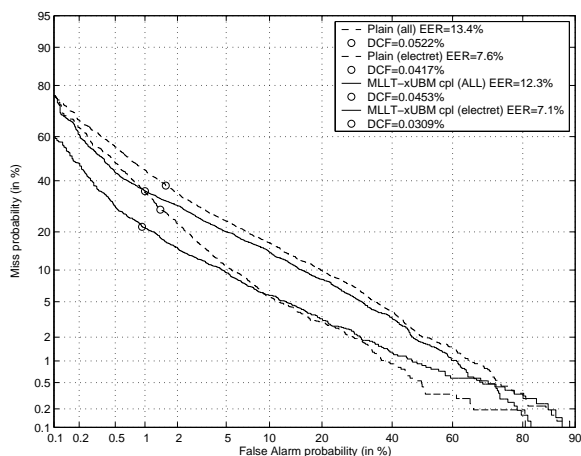


Figure 3: The DET plot of the baseline vs. the MLLT-xUBM system with T-U coupling for the 1999 evaluation and its electret trials subset.

UBM-first case, suggesting that the way of fixing the Gaussian components in the scoring plays a relevant role for different operating points. Figure 3 compares the performances of the plain system with the MLLT-xUBM system with $U \rightarrow T$ coupling in terms of the DET curves. The plot includes the complete 1999 evaluation set as well as a subset of purely electret-microphone trials. The rotation effect of the coupling is apparent especially for the electret condition and has a favourable effect in particular on the minimum DCF which is reduced by about 25%.

Table 2 shows the performance of the T-normalization and its modification as well as the H-norm applied on top of the MFCC MLLT-xUBM system with coupling. An improvement by using the weighting function from Eq. (6) is achieved in the T-norm, however, the handset normalization effect of the H-

System	opt. DCF 10^{-3}		
	T-norm	T-norm w/wgt	H-norm
MLLT-xUBM cpl	44.2	43.1	41.9

Table 2: Optimum detection costs after T- and H-norm

Config	Performance Measure	
	minDCF	EER
MFCC	46.4	12.9%
PLP1	50.6	15.7%
PLP2	50.7	14.8%
Linear Comb.	46.3	13.3%
Neural Comb.	44.1	12.7%

Table 3: Comparison of fused system performance.

norm alone outperforms the T-norm and reduces the optimum DCF to $41.9 \cdot 10^{-3}$. For further potential reductions, the H- and T-norms can be used in a combined fashion as mentioned in [1].

4.4.2. Neural Network System Fusion

The fused system was trained on multiple base feature sets described in Sec. 4.2.

After the input layer, we used one hidden layer with three nodes followed by a hidden layer with two nodes followed finally by one output node. The network was not fully connected as we enforced the structure in Fig. 1. We developed the fusion by creating a 129 speaker subset of the NIST99 evaluation data for training, with the complement used for testing (there was no speaker overlap in the training and testing data). A comparison of the performances is given in Table 3 showing the performance of the individual system, as well as that of score fusion using an equal weight linear combination. The neural net fusion performs the best in both equal error rate (EER) and minimum cost (minDCF).

5. Conclusion

Our experimental results show that through the use of linear transforms applied to the target-background model pair as well as by appropriately choosing the Gaussian components in the score calculation, the verification performance can be significantly improved. Further performance boost was demonstrated by fusing scores from multiple systems with different acoustic representations via a neural network. Results obtained on the 1999 speaker recognition evaluation set indicate reductions of the minimum detection cost of up to 13% and 25% for all tests and electret-only tests respectively, as compared to a baseline GMM system. The neural fusion of three systems gains further 5% cost reduction.

6. References

- [1] D.A Reynolds, T.F. Quatieri, and Dunn R.B. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, January/April/July 2000.
- [2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, January/April/July 2000.
- [3] U.V. Chaudhari, J. Navrátil, and S.H. Maes. Multi-grained data modeling for speaker recognition with sparse training and test data. In *Proc. of the ICSLP*, Beijing, October 2000.
- [4] D.A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. of the EUROSPEECH*, Rhodes, Greece, September 1997.
- [5] (URL). <http://www.nist.gov/speech/tests/spk/index.htm>.
- [6] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. of the EUROSPEECH*, pages 1895–8, Rhodes, Greece, September 1997.