THE IBM SYSTEM FOR THE NIST-2002 CELLULAR SPEAKER VERIFICATION EVALUATION

Ganesh N. Ramaswamy  Jiří Navrátil  Upendra V. Chaudhari  Ran D. Zilca
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, {ganeshr,jiri,uv,zilca}@us.ibm.com

ABSTRACT
This paper presents an overview of the architecture and algorithms implemented in IBM's text-independent speaker verification system developed for the 2002 NIST Speaker Recognition Evaluation, particularly for the 1-speaker detection task using cellular test data. We describe individual components including a Gaussianization front-end, cellular-codec post-processing, modeling, discriminative optimization and scoring steps. A combination of multiple, data-perturbed systems using a discriminative objective so as to achieve optimum performance for a low false alarm operating region obtained the top performance in the NIST 2002 1-speaker detection task.

1. INTRODUCTION
Text-independent speaker verification has seen significant progress in the past several years, due in part to the annual speaker recognition evaluations organized by the National Institute of Standards and Technology (NIST). These evaluations have motivated research efforts in several speaker recognition tasks, while also serving as a source for calibrating technical capabilities of various approaches. The task of interest in this paper is the one-speaker detection task, otherwise known as the speaker verification task.

We describe the architecture and algorithms implemented in IBM’s text-independent speaker verification system, which obtained the best performance in the NIST-2002 cellular one-speaker detection task. The system is based on the Gaussian Mixture model structure and the widely successful approach of adapted target speaker models [9]. In order to alleviate mismatch problems due to channel variability, we apply several steps of compensation and normalization during feature extraction and score calculation based on the Gaussianization technique [11] [6] and the T-Norm [1]. A coupled-Gaussian scoring technique [4] combined with a discriminative feature-space transform [5] help optimize our system for the NIST-defined operating region of low false alarms. The overall system consists of multiple systems utilizing different training datasets achieving an advantageous perturbation gaining performance in a combination. Furthermore a cellular codec simulation software was utilized to adapt landline-telephone data to the cellular task thus helping the overall data augmentation and a better testing-condition match.

The remainder of the paper is divided into three sections. Section 2 describes our system, including the front-end, model structure, detection and score generation, score normalization, score combination and threshold setting. Section 3 contains the experimental results on both the NIST-2001 and NIST-2002 test sets. Section 4 concludes the paper with a summary and outlook.

2. SYSTEM DESCRIPTION
Figure 1 shows the overall system architecture. The following sections detail on the individual components and their functionality.

2.1. Front-End
Our approach involves a combination of multiple systems (up to five) whereby each system is distinguished primarily by the training data used to create it. However, certain differences in the front-ends of the individual systems apply as follows. In all five systems, extracting features from the speech signal consists of generating the classic MFCC frame sequence comprising 19 static plus 19 derivative dimensions followed by either a marginal or a full Gaussianization transform as described in [11]. While the marginal Gaussianization, first introduced by Pelecanos [6] as feature warping, assumes independence of the input feature dimensions and carries out a series of non-linear dimension warpings independently of each other, the full Gaussianization first estimates and applies a linear feature-space transform with a maximum-likelihood objective to approximate the feature independence before applying the marginal warping. The Gaussianization step seeks to achieve normal distribution of the features within any window of a predetermined length - in our case three seconds. This leads to a reduction of the redundancy due to channel effect suppression alongside a partial information loss due to removal of long-term speaker properties, further extending the process of feature standardization. The full Gaussianization front-end is applied in System 1, involving gender-dependent linear transforms applied on each speech utterance, followed by a 3-sec-window warping, whereby the gender information is known a priori. Systems 2 through 5 use the marginal Gaussianization alone.

Feature frames with signal energy lower than a certain threshold are discarded during the extraction process before the Gaussianization step resulting in an average frame rejection rate of about 30%. Furthermore, in order to speed up the subsequent process of training background models, a frame subsampling of 2:1 was set in Systems 3, 4.

2.1.1. Signal Postprocessing via Cellular Codec
In order to achieve data perturbation across the multiple systems, additional data comprising the landline-telephone quality collection of the Switchboard-I corpus from the NIST-1996 evaluations was utilized. To add some simulated cellular characteristics absent in the landline speech signal, the landline data were additionally processed by a
number cellular software codes, based on the GSM specifications (610, 66, ETS200, 724) in System 3, as well as the CDMA specifications (EV R, QC13) in System 4.

2.2. Modeling
The core model structure in all systems is a Gaussian Mixture. Used with considerable success in the past [7, 8], individual speaker models are estimated as Maximum-A-Posteriori (MAP) adapted models using a speaker-independent Gaussian Mixture Model (GMM) trained in a relatively extensive fashion covering a variety of acoustic conditions, such as landline and cellular channels, male and female speakers, etc. In our system, the training of such a speaker-independent model, also known as Universal Background Model (UBM) [8], consists of the following steps:

1. Pre-Cluster the training data set via a fast deterministic top-down algorithm using binary data splits with an eigenvector-based criterion, as described in [4]
2. Iterate via K-Means using the Euclidean distortion measure until convergence
3. Carry out one iteration of the Expectation-Maximization algorithm on the mean and covariance parameters calculated from final K-Means clusters. Only the diagonals of the covariance matrices are computed
4. Estimate a feature-space Maximum-Likelihood Linear Transformation (MLLT) using the EM parameters from the previous step. The MLLT achieves optimum feature space for diagonal covariance modeling [2]
5. Repeat Step 3 and 4 iteratively until convergence, or until a maximum number of iterations is reached.

Using the speaker training data, the MAP adaptation is applied on the mean parameters of the UBM in the MLLT-transformed space to create each individual speaker model characterized by a set of adapted mean vectors with the diagonal covariances and the Gaussian weights being shared with the UBM.

2.2.1. fDETAC
An advantageous property of the mean-only adaptation, in conjunction with a single-best Gaussian scoring [4], is the fact that the likelihood ratio of the adapted and the UBM Gaussians is a linear function of the feature vector – a property exploited in the discriminative training of a \( \mathbf{A} \in \mathbb{R}^{d \times d} \) matrix with the DETAC objective [5], acting on the feature vectors (fDETAC). The procedure was applied in three systems with the objective to rotate counterclockwise and shift downwards the resulting system DET curve, i.e. to achieve a performance improvement in the operating region of interest – the low false-alarm area. Since the fDETAC optimization is a discriminative technique, two sets including true-target trials and impostor trials were allocated as a development set.

2.3. Detection
To process a particular trial, i.e. a test utterance against a claimed target identity, the individual systems calculate the (componentwise) Gaussian likelihoods of the vectors in the sequence using the system-dependent UBM s. As described in [5], for each feature vector, only the maximum-likelihood Gaussian component is taken into account in each system, followed by a likelihood ratio of that component and its corresponding adapted counterpart in the target speaker GMM. Such tying of single components was shown experimentally to cause a counterclockwise rotation of the DET curve, which is favourable to the operating region of interest. The framewise likelihood ratios are then averaged over the complete utterance and output as real-numbered scores.

2.3.1. Score Normalization
Beyond speaker properties, the resulting likelihood ratio values are typically influenced by irrelevant information such as channel properties and the generally non-linearly shaped acoustic space. The T-Norm [1] is applied to the ratio scores in order to compensate for certain shifts and scales due to the acoustic space. The T-Norm involves a set of additional speaker models created consistently with the regular target models, which serve as a basis for calculating the first and second order statistics of the score distribution at the given test point in the acoustic space. The trial score is normalized by subtracting the mean and dividing by the standard deviation of the T-Norm model scores. Due to the fact that the T-Norm takes the acoustic conditions of the given test into account, it is effective in suppressing score distribution shifts and scaling due to channel variations. In the final evaluation, we used a total number of 234 T-Norm speakers, whereby a gender-matched subset was used for each given trial. Furthermore a weighting and pruning scheme of the individual T-Norm scores is applied as described in [3].

2.3.2. Score Combination
To arrive at a single score for the given trial, scores from individual systems are linearly combined with the weights for individual system sources estimated via the pDETAC method [5] according to the objective of rotating counterclockwise and shifting the resulting DET curve.

2.3.3. Threshold Setting
The decision threshold, as a required part to be provided along with the scores for the NIST evaluation, is determined using a heuristic method as follows: Using a development dataset, the Detection Cost Function (DCF) is first examined for its minimum and the corresponding threshold value \( t_{\text{min}} \). Then two threshold values on the threshold axis are found (one to the left, \( t_0 \), one to the right, \( t_1 \), of the minimum-DCF) such that their DCFs is 10% higher than the minimum. Finally, the center of the interval \([t_0, t_1]\) is chosen to be the threshold applied on the evaluation data. This technique proved empirically to be a suitable way of avoiding thresholds overtrained due to data spuriousness. A typical DCF plot as a function of the threshold looks relatively smooth with a single minimum, however, with an asymmetry of the left and right parts around the minimum point. Thus, setting a threshold at an optimum but near a region with a steep growth may cause problems on unseen data shifting the threshold off the minimum point into a high DCF region easily. The described technique alleviates such a sensitivity to some degree.

3. EXPERIMENTAL RESULTS
3.1. Database
The performance of the described systems was developed and evaluated using data from the cellular part of the Switchboard (SWB) telephone corpus, in particular the set defined by the NIST for the 1-speaker cellular detection task (CT) in the 2001 and 2002 Speaker Recognition Evaluations (SRE) [10]. The 2001 set consists of 60 development, 174 test speakers, and a total of 20350 verification trials with the GSM codec being the prevalent type. For
3.2. System Setup
As mentioned earlier, a total number of five GMM systems with coupled UBM-Target modeling and a single maximum-likelihood linear transform (MLLT) as described in [4] were used in the evaluation, differing primarily in the dataset used to create the respective UBMs:

- System 1 (2 × 1024 Gaussians) consisting of two UBMs trained in gender-dependent fashion. Each UBM uses a gender-dependent linear transform in the full Gaussian front-end as described in Section 2.1.
- System 2 (1536 Gaussians) UBM was trained using the 60 development speakers of the 2001 CT plus 2 hrs of data from the internal data collection (cellular-GSM quality)
- System 3 (2048 Gaussians) had the 60 speakers plus the 1996 SRE landline-telephone recordings post-processed by three GSM-encoders (see Section 2.1.1.)
- System 4 (2048 Gaussians) used the 1996 SRE landline data only without post-processing
- System 5 (2048 Gaussians) had identical composition as System 3 except the codec used was of a CDMA type (see Section 2.1.1.)

The detection cost function (DCF) defined in the 2002 SRE [10] served as the primary evaluation measure. This DCF weights the two error types according to $Cost = 0.99 \times \varepsilon_{FA} + 0.1 \times \varepsilon_{MIS}$, and thus shifts the operating point towards low $P_{FA}$ and makes a counterclockwise DETAC rotation desirable.

3.3. NIST-2001 Cellular Task Results
Tables 1 and 2 summarize the results for individual systems in two development stages:

- (Table 1) Cross-evaluation test on the NIST-2001 cellular task set (two splits cross-evaluation to train the fDETAC transform). The 60 development speakers served, in addition to being used in the UBM training, as T-Norm speakers. Here, due to the fact that the T-Norm speakers were the sole data in the UBM of System 1, no T-Norm was carried out on this system. Also, at this stage System 5 was not evaluated.
- (Table 2) Test on a subset of 74 (out of 174) speakers of the NIST-2001 cellular task, whereby adding the remaining 100 speakers to the T-Norm set (leaving 160 speakers in total). At this stage, the threshold for the NIST-2002 evaluation was also estimated using this set.

Consistent gains can be seen by using the fDETAC applied on Systems 2-4, while the gender-dependency and consequently the extra data partitioning prevented robust estimation of the transform in System 1. The fDETAC gains relative to the T-Normed systems are around 5% relative DCF.

3.4. NIST-2002 Cellular Task Results
The DCF performances of the five systems used in the 2002 evaluation and their combination are shown in Table 3. The remaining 74 speakers used for evaluation in Table 2 were added to the T-Norm speaker set resulting in a total number of 234 speakers (122 female, 112 male).

While the fDETAC gains seem to hold when considering the systems before the T-Norm, the positive effect diminishes after applying the T-Norm. A possible factor may be

Note that, unlike in our paper, all NIST-released DCF results are divided by a constant 1.00.

Table 1. 2001 Cross-Evaluation Set performance of the GMM systems with feature-space DETAC, T-Norm and projectional DETAC

<table>
<thead>
<tr>
<th></th>
<th>opt. DCF 10^-3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>Baseline</td>
<td>37.6</td>
</tr>
<tr>
<td>+1DETAC</td>
<td>n/a</td>
</tr>
<tr>
<td>+ T-Norm (60)</td>
<td>n/a</td>
</tr>
<tr>
<td>Comb.</td>
<td>32.8</td>
</tr>
</tbody>
</table>

Table 2. 2001 Subset performance with 160 T-Norm speakers and projectional DETAC

<table>
<thead>
<tr>
<th></th>
<th>opt. DCF 10^-3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>Baseline</td>
<td>33</td>
</tr>
<tr>
<td>+ T-Norm (160)</td>
<td>30</td>
</tr>
<tr>
<td>Comb.</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 3. 2002 “All Trials” DCF performance of individual systems and pDETAC combination. In (%) are rel. gains of fDETAC after the T-Norm.

<table>
<thead>
<tr>
<th></th>
<th>opt. DCF 10^-3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
</tr>
<tr>
<td>Baseline</td>
<td>47.4</td>
</tr>
<tr>
<td>1DETAC</td>
<td>n/a</td>
</tr>
<tr>
<td>+ T-Norm (234)</td>
<td>37.5</td>
</tr>
<tr>
<td>Comb.</td>
<td>31.6 optimum / 39.2 threshold</td>
</tr>
</tbody>
</table>
false alarm probability (in %)

NIST 2002 Cellular Task (All Trials)

IBM 5−Comb. System
DCF=0.0346, EER=8.9%

Figure 2. Detection Error Trade-Off Curve for the All-Trials Condition of the NIST-2002 Evaluation

the prevalent codec difference between the 2001 and 2002 cellular tasks (GSM and CDMA respectively), and thus an overtraining effect of the fDETAC transform. In all experiments, however, there appears to be an interaction between the fDETAC and T-Norm techniques such that their gains are not independent and additive – an observation that deserves further study.

The five systems combined using the weights estimated via the pDETAC on the 2001 set, outperform the best individual system (System 5) by additional 4% relative DCF – a gain similar to those observed in the development stage. The corresponding Detection Error Trade-Off Curve (DET) of the combined system is shown in Figure 2.

4. CONCLUSIONS

The experimental results indicate the importance of appropriate data processing, particularly so as to minimize variability due to the channel as well as to achieve comparable conditions in the system training and test. The strength of score normalization, particularly the T-Norm using matched-condition speakers, proves essential to these objectives. Significant gains could be observed by combining systems with their UBM training perturbed by different datasets and data types. A discriminative training technique based on the DETAC objective applied on feature model as well as score level brings gains in performance in the operating region of interest, however an interaction between the score normalization T-Norm and the fDETAC transform appears to prevent additive gains of both steps - an observation that deserves further study.

REFERENCES


[10] (UR ).