THE MIT LL 2010 SPEAKER RECOGNITION EVALUATION SYSTEM: SCALABLE LANGUAGE-INDEPENDENT SPEAKER RECOGNITION*

Douglas Sturim†,William Campbell†, Najim Dehak‡, Zahi Karam†*, Alan McCree†, Doug Reynolds†, Fred Richardson†, Pedro Torres-Carrasquillo†, Stephen Shum‡

†MIT Lincoln Laboratory, ‡MIT CSAIL, *DSPG Research Laboratory of Electronics at MIT

ABSTRACT

Research in the speaker recognition community has continued to address methods of mitigating variational nuisances. Telephone and auxiliary-microphone recorded speech emphasize the need for a robust way of dealing with unwanted variation. The design of recent 2010 NIST-SRE Speaker Recognition Evaluation (SRE) reflects this research emphasis. In this paper, we present the MIT submission applied to the tasks of the 2010 NIST-SRE with two main goals—language-independent scalable modeling and robust nuisance mitigation. For modeling, exclusive use of inner product-based and cepstral systems produced a language-independent computationally-scalable system. For robustness, systems that captured spectral and prosodic information, modeled nuisance subspaces using multiple novel methods, and fused scores of multiple systems were implemented. The performance of the system is presented on a subset of the NIST SRE 2010 core tasks.

1. INTRODUCTION

The 2010 NIST speaker recognition evaluation (NIST-SRE) was largely a continuation of the previous 2008 NIST-SRE. The same basic recording channels were used as in the 2008 SRE—auxiliary microphone and telephone channels. Auxiliary microphones were first introduced in the 2005 evaluation and have been a part of the main tasks in following evaluations. Speaking styles were first changed in the 2008 SRE. All previous evaluations since their inception of the NIST-SRE have used the speaking style found in conversational-telephone speech. The 2008 evaluation introduced the interview-condition in which a subject was simultaneously recorded on multiple microphones. Recorded speakers in the 2010 evaluation were required to speak English. The majority of test subjects were native US English speakers with a large minority of non-native talkers.

Microphone and recording channel variation motivated continued research in techniques to reduce these nuisance factors. Nuisance mitigation has dominated the speaker identification field in recent years. Research has progressed on two fronts, 1) factor analysis [1] and 2) weighted nuisance attribute projection [2, 3]. Our submitted systems reflect this research emphasis.

All of our systems have the following attributes—inner product based, subspace compensation, and language-independent modeling. This makes the systems easily adapted to new languages, scalable, and robust to channel nuisances.

*This work was sponsored by the Federal Bureau of Investigation under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

To address the tasks of the NIST evaluation we fielded mostly low-level cepstral-based algorithms and one system relying on high-level 'prosodic' features [4]. The MIT Lincoln Laboratory/CSAIL/RLE system submission can be categorized into three areas; 1) GMM factor analysis, 2) SVM and kernel based, and 3) high-level prosodic modeling. The two most notable additions this year was a total variability (TV) system [5] and inner product discriminant functions (IPDF) system [6].

Depending on the task, different combinations of systems were combined with a logistic regression algorithm. The goal of fusion is to gain performance by combining techniques that contain complimentary information. The speaker ID community has historically fused low-level systems (i.e. spectral based) and high-level systems (i.e. prosodic based). The philosophy is that the systems discern information that the other systems cannot find. In contrast to prior evaluations (e.g., NIST SRE 2008), we found that our different low-level cepstral systems fused well giving substantial performance gains.

We present in the paper the system techniques and experimental results of the submitted systems. The speaker recognition systems are presented in Section 2. Section 3 presents the experimental results as well as post evaluation analysis. We conclude in Section 4, with reflections on the 2010 NIST-SRE as well a indications of future directions of research.

2. RECOGNITION SYSTEMS

2.1. Features

Two types of pre-processing of the data were performed—echo cancellation for telphone speech and noise reduction [7] for microphone speech. Then, two types of features were extracted, MFCCs and LPCCs. MFCCs were extracted with a standard front-end [8]. LPCCs were extracted using HTK. For both feature sets, changes were made based upon experiments at the 2008 JHU Summer workshop—bandwidth was changed to [0,4 kHz], RASTA was turned off, acceleration and energy were added, and 0/1 normalization of features was switched to feature warping. The feature vector size per frame was 60 for MFCCs and 57 for LPCCs. Speech activity detection depended on the channel-type and was performed with a combination of ASR, GMM-based, and energy-based detectors.

2.2. IPDF-KL system

Inner product discriminant functions (IPDFs) are described in [6]. We use a comparison function from the IPDF framework based on approximations to the KL divergence between two GMMs [3, 6]. For a sequence of feature vectors from a speaker i, we adapt a gender-independent 512 mixture GMM UBM using a relevance factor of 0.01 for the means and an ML estimate of the mixture weights. The adaptation yields new parameters which we stack into a parameter vector, $\mathbf{a}_i$.

The IPDF-KL inner product, C_{GM}, is given by

$$C_{GM}(\mathbf{a}_i, \mathbf{a}_j) = (\mathbf{m}_i - \mathbf{m})^t (\boldsymbol{\lambda}_i^{1/2} \otimes I_n) \Sigma^{-1} (\boldsymbol{\lambda}_j^{1/2} \otimes I_n)(\mathbf{m}_j - \mathbf{m}) \quad (1)$$

where $\mathbf{m}_i$ and $\mathbf{m}_j$ are the adapted means, $\mathbf{m}$ is the vector of stacked UBM means, Σ is the block diagonal matrix of UBM covariances, $\otimes$ is the Kronecker product, I_n is the identity matrix of size n, and $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ are diagonal matrices of adapted mixture weights.

For compensation, weighted NAP (WNAP) [2] was used. Weighting was based on the number of frames of speech in the nuisance training space. WNAP used a fixed matrix multiply.

To obtain scores, we applied gender independent WNAP to both enroll and verification mean parameter vectors. The WNAP corank was fixed at 128. We then scored using the C_{GM} kernel. Both Z- and T-Norm were applied.

2.3. JFA System

The base system for our Joint Factor Analysis (JFA) work was the MITLL GMM UBM speaker detection system. Our JFA setup is based on the work of [9], where the mean supervector is decomposed as:

$$M = m + Vy + Dz + Ux, \quad (2)$$

where m is the speaker-independent mean supervector of GMM means, U defines the within-class (session/channel) variability subspace, V defines the across-class (speaker) variability subspace, and D is a diagonal matrix describing the remaining speaker variability.

We used gender-dependent UBMs with 1024 mixtures. 300 eigenvoices were trained using a variation of PCA of the across-class variability covariance matrix. To reduce over-estimation bias of the eigenvalues, a cross-validation approach was used where the eigenvectors were estimated from one partition of the training data and the eigenvalues were estimated as the energy in these directions over the other partition. We found that using this approach, the diagonal matrix could be estimated from the same data. In a similar way, 100 eigenchannels were estimated from the within-class covariance matrix. Two of these estimates were generated, one for telephone channels and the other for microphone conditions, and stacked together into a combined 200-dimensional matrix.

Enrollment of speakers in this system consists of estimating $Vy + Dz$ in the presence of Ux, and is done by stacking all the parameters together and extracting the speaker model. Testing is done by removing Ux from the test utterance. To speed up the Gaussian scoring, only the linear (inner product) term is calculated as in [10]. ZT-norm was applied to these output scores.

2.4. Prosodic System

A more extensive overview of the prosodic system and its features is detailed in [4]. These features are extracted at the pseudo-syllabic level and correspond to a Legendre polynomial approximation of the pitch and energy contours. We used six Legendre polynomial coefficients each for pitch and energy, as well as the duration of the pseudo-syllable to obtain a feature vector of 13-dimensions. We used a gender-dependent Universal Background Model (UBM) composed of 512 Gaussians per gender and gender-dependent total variability matrices of 200 eigenvectors trained only on telephone speech [11]. LDA was used to reduce the dimension to 75, while WCCN normalized the cosine scoring. We used cosine scoring and applied zt-norm to normalize the final decision scores.

2.5. Eigenvoice Comparison System (ECS)

For ECS, speaker model enrollment consists of generating speaker factors for the enrollment and test utterance $\mathbf{s}_{train}$, without any session variability modeling or compensation. The speaker factors are assumed to have a Gaussian distribution. We use a standard log-likelihood ratio test and keep only the inner product. We perform an inner product with normalized vectors,

$$LL(\mathbf{s}_{test}|\mathbf{s}_{train}) = \frac{\mathbf{s}_{test}^T \boldsymbol{\Sigma}_{wc}^{-1} \mathbf{s}_{train}}{\sqrt{(\mathbf{s}_{test}^T \boldsymbol{\Sigma}_{wc}^{-1} \mathbf{s}_{test})(\mathbf{s}_{train}^T \boldsymbol{\Sigma}_{wc}^{-1} \mathbf{s}_{train})}} \quad (3)$$

The speaker loading matrix for obtaining the factors is from the JFA system. The matrix $\boldsymbol{\Sigma}_{wc}$ is full covariance and is estimated from the session variability data. ZT-norm is applied to the scores.

2.6. SVM GSV

The SVM GMM supervector system is based on [3]. The system was applied almost unchanged from NIST SRE 2008 [8] except for different input features. GMM supervectors were derived using MAP adaptation of means only with a relevance factor of 4 on a per utterance basis. NAP [3] was used for nuisance compensation. A KL-based SVM kernel was used for enrollment. Scoring was an inner product. ZT-norm was applied to the scores.

2.7. Total Variability System

The total variability system is composed of two subsystems, one exclusively for telephone speech and another for microphone or interview data. The parameters for the first subsystem were trained on telephone data. We use a gender-dependent UBM with 2048 Gaussians and gender-dependent total variability matrices consisting of 600 eigenvectors trained on telephone speech [5]. The use of Linear Discriminant Analysis (LDA) reduces our dimensionality to 250, and Within Class Covariance Normalization (WCCN) carries out the channel compensation in the total variability space [5]. Similar to [11], we use cosine scoring and zt-normalization to make the final decision. As with everything else so far, the impostors for zt-norm were entirely selected from telephone speech data.

The second subsystem is used when we have microphone and interview data in training or in testing. This system is based on the total variability space and its 600 total factors estimated on telephone speech and an additional 200 total factors trained in microphone and interview data. We then use Probabilistic LDA [5] to project all microphone and telephone total factors of dimension 800 into speaker space of dimension 600. The PLDA consists of a mean vector of dimension 800 estimated from telephone data, an eigenvoice matrix of dimension 800x600 trained on telephone speech, an eigenchannels matrix of dimension 800x200 trained exclusively on microphone and interview speech, and a full covariance matrix trained from telephone speech. After the projection with PLDA, we used LDA to reduce the 600 dimensions to 250 and WCCN to normalize the cosine kernel. These channel compensation matrices are estimated using telephone, microphone and interview data all together. And as before, the decision score is computed using cosine scoring, but the final scores are normalized using s-norm [5].

2.8. Adaptive Norming

Adaptive norming of scores showed promise in our development set for minimizing the new minDCF criterion and was applied to three systems—IPDF, JFA, and TV. Adaptive normalization techniques were applied with inspiration from several sources including cohort normalization [12], T-Norm, Z-Norm, and adaptive variants [13].

As with classic cohort selection [12] and Z- and T-norm, there are several issues in adaptive methods—cohort selection, cohort normalization function, and whether the model or test score is normalized. Cohorts selection was accomplished by a simple method. For a given message or model, cohorts were selected as the highest scoring

models or messages (respectively) from a large dataset. Cohort normalization was performed with the mean and standard deviation of the cohort scores. More details on our methods are in a companion paper [14].

2.9. Fusion

Fusion was performed using a separate logistic regression for nine subconditions of the core NIST SRE task. The criteria function of the logistic regression, normalized conditional cross-entropy, was adjusted to use the new target prior for the 2010 NIST SRE. Although the criterion function is not the same as NIST performance metric C_{Det}, C_{Det} was generally improved when we optimized using the same effective target prior that matches Bayes' optimal decision rule.

3. EXPERIMENTS

3.1. Experimental Setup

The corpora used for NIST SRE 2010 development lists consisted entirely of data from the 2008 NIST speaker evaluation. These lists were used for system tuning, system selection, and backend fusion/calibtration. Several adjustments were made so that the data would be suitable for developing a system designed to perform well in the 2010 NIST SRE:

- **additional background training data**: the NIST SRE 2008 interview microphone data was partitioned into two approximately equal sets one of which was used for subspace, ZT-norm or background model data and the other was used for development train and test data.
- **increased non-targets**: an exhaustive set of non-target trials was created for each development test data set in order to match the much lower target prior in the 2010 NIST SRE.

3.2. Results on Selected Core Tasks

System fusion results for 2010 NIST-SRE for our primary fusion system are presented in Table 1. The results are broken out into four of the nine training condition categories of the 2010 NIST-SRE. We additionally show results for trials of 2.5 minutes of training/testing data—the short condition—for comparison to prior evaluations.

- Sub-Conditions 1 and 2 – Int-Int Same-Mic and Int-Int Different-Mic: Single training and testing utterances from interview microphones.
- Sub-Condition 3 – Int-4w: Single training from an interview microphone and testing utterance from a telephone.
- Sub-Condition 5 – 4w-4w: Single training and testing utterances from a telephone.

Table 1 shows the performance of our fused systems for multiple tasks. Note that "old minDCF" corresponds to the DCF from prior (2008 and before) evaluations. The fused system varied per subtask and was: conditions 1, 2 (GSV+TV+JFA), condition 3 (IPDF+TV+JFA), condition 5 (IPDF+TV+JFA+Prosodic+ECS). Performance in terms of minDCF and EER is very good for the overall system. Calibration for the fused system had issues which we explore in the following experiments.

During the post evaluation analysis, we discovered that two of the spectral systems (JFA of section 2.3 and IPDF-KL system of section 2.2) had calibration and performance issues. Both the JFA and IPDF-KL systems exhibited sensitivities to front-end processing that did not appear in the experimentation with the development corpus. We investigated three changes to the front-end processing (Section 2.1). The processing changes we consider here are: 1) preforming feature warping instead of feature norming, 2) using the full bandwidth (0-4KHz) instead of our standard band-limiting to (300-3100 Hz), and 3) turning rasta off. Table 2 shows the experimental progression in turning each of these processing changes off in-turn. In Table 2 we see that the performance of both systems in condition 2 was slightly worse. In condition 5 of Table 2, the JFA system got slightly better, but the IPDF-KL system improved significantly. After investigating performance issues, we looked at system calibration, see Table 3. The table shows that the new features significantly narrows the gap between minDCF and actDCF.

We then fused three systems (IPDF-KL/GSV+TV+JFA) using the configuration 001—results are shown in Table 4. We can see from the table that the new feature configuration improves system performance and calibration. Note that condition 1 is still difficult in actDCF since the meta-data is not available to detect matched and mismatched microphones for fusion.

3.3. Challenges in the NIST SRE 2010 Experimental Setup

The 2010 NIST-SRE posed challenges with its experimental design and presentation. The first most notable change was the new minDCF point. Previous years had the decision cost function (DCF) point with the cost of miss, cost a false alarm and probability of tar-

Table 1. Eval system performance on Eval10 Core task for Fused System

Cond	Duration	EER (%)	Old minDCF	minDCF	actDCF
1	all	2.63	0.126	0.458	25.302
1	short	3.12	0.140	0.480	24.909
2	all	3.21	0.158	0.517	1.835
2	short	3.58	0.171	0.533	1.971
3	all	2.86	0.117	0.454	0.538
3	short	3.18	0.128	0.478	0.566
5	all & short	1.86	0.090	0.380	0.507

Table 2. Performance on short train and test for Eval10 Core task, Multiple Feature Types. Mask is FW (1) or FN (0); full bandwidth (1) or telephone bandwidth (0); rasta on (1) or off (0).

Mask	GSV minDCF Cond 2	JFA minDCF Cond 2	IPDF-KL minDCF Cond 5	JFA minDCF Cond 5
110	0.671	0.756	0.713	0.523
111	0.712	0.758	0.649	0.501
100	0.756	0.821	0.560	0.486
010	0.701	0.761	0.706	0.511
001	0.764	0.826	0.483	0.472

Table 3. Selected individual system performance on short train and test for Eval10 subconditions

Cond	Feat	IPDF/GSV minDCF	IPDF/GSV actDCF	JFA min DCF	JFA actDCF
1	Eval	0.654	29.855	0.691	43.032
1	001	0.612	2.202	0.587	1.292
2	Eval	0.687	1.760	0.765	6.992
2	001	0.764	0.800	0.788	0.835
3	Eval	0.602	0.660	0.666	0.749
3	001	0.580	0.619	0.628	0.630
5	Eval	0.713	1.078	0.504	0.617
5	001	0.483	0.548	0.468	0.472

Table 4. System performance on train and test on short for Eval10 Core task with 001 features for fused system

Sub-Cond.	EER (%)	Old minDCF	minDCF	actDCF
1	1.89	0.092	0.324	1.439
2	3.04	0.151	0.541	0.546
3	3.15	0.125	0.454	0.563
5	2.03	0.095	0.380	0.393

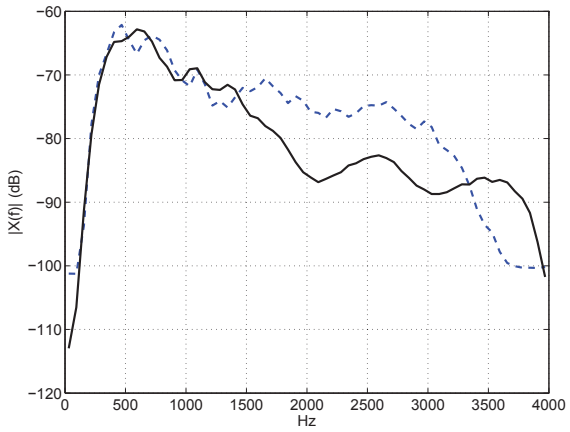

Fig. 1. Comparison of spectra of 2008 (dashed blue) and 2010 (solid black) interview microphones recordings

get defined at $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, and $P_{tgt} = 0.01$. The new DCF point was set at ($C_{Miss} = 1$, $C_{FalseAlarm} = 1$, $P_{tgt} = 0.001$). The most notable impact of the new operating point was the need for more impostor trials. The number of trials should be on the order of 1 million (depending on system accuracy). The initial 2010 SRE trials were far too small to interrogate this new DCF point. The extended evaluation was then created to correct this deficiency. The consequence is that all results using the new DCF point should be reported on the extended evaluation.

The other impact of the new operating point, is that the output score distributions of individual systems may not be Gaussian. We discovered this phenomenon in both the development and evaluation datasets. The development data were more non-Gaussian at the DCF point when compared to the evaluation speech data. This is counter to the Gaussian score assumptions made in most state-of the art fusion techniques. Investigations at the old DCF point show that the Gaussian score assumption still holds. This is certainly an area for future study since it is not clear how to predict how much future score distributions will diverge from Gaussian. It is also not clear what is causing the divergence. If changes in the collection paradigm influences these divergences, then systems are in greater danger of becoming over-tuned to particular corpora.

In the post evaluation analyses of Section 3, we can see that the greatest impact on performance and calibration is the limiting of the bandwidth to $300 - 3140KHz$. Figure 1 shows a clear difference in the spectrums between 2008 and 2010 evaluation speech data. This spectral mismatch between collection years impacted both of our IPDF and JFA systems.

Another characteristic of the 2010 NIST-SRE setup was that some of the conversational telephone speech was recorded with the high vocal effort (HVE) apparatus. The same talkers where also recorded with normal conversational speech. This constrains a portion of the recorded speech to be collected over a limited number of handsets. This changes the problem in these instances to be closer to the access control concept of operations rather then a speaker identification problem over a broad corpus.

4. CONCLUSIONS

We have presented the MIT site speaker recognition system used for the 2010 NIST-SRE. We have described the systems for speaker recognition using total variability, factor analysis, discriminative function techniques, channel compensation, and high-level speaker recognition. Post-evaluation analysis showed calibration/performance issues for the IDPF-KL and JFA systems. Post evaluation analysis was presented for these systems. The corrections were reapplied to the evaluation fusion system and was shown to correct the calibration and performance for the system overall.

5. REFERENCES

[1] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. Odyssey04*, 2004, pp. 219–226.

[2] W. M. Campbell, "Weighted nuisance attribute projection," in *Proc. IEEE Odyssey*, 2010.

[3] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proceedings of ICASSP*, 2006, pp. I–97–I–100.

[4] N. Dehak, P. Kenny, and P. Dumouchel, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2095–2103, Sept. 2007.

[5] Najim Dehak, Zahi N. Karam, Douglas A. Reynolds, Reda Dehak, William M. Campbell, and James R. Glass, "A channel-blind system for speaker verification," *submitted to ICASSP*, 2011.

[6] W. M. Campbell, Z. N. Karam, and D. E. Sturim, "Inner product discriminant functions," in *Advances in Neural Information Processing Systems 22*, Cambridge, MA, 2009, MIT Press.

[7] D. E. Sturim, W. M. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proceedings of ICASSP*, 2007, pp. IV–49–IV–52.

[8] D. E. Sturim, W. M. Campbell, Zahi N. Karam, Douglas Reynolds, and F. S. Richardson, "The MIT Lincoln Laboratory 2008 speaker recognition system," in *Proc. Interspeech*, 2009, pp. 2359–2362.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions On Speech And Audio Processing*, vol. 13, no. 3, pp. 345, May 2005.

[10] Ondrej Glembek, Lukas Burget, Najim Dehak, Niko Brummer, and Patrick Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," in *Proceedings of ICASSP*, 2009.

[11] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, 2009.

[12] Aaron E. Rosenberg, Joel DeLong, Chin-Hui Lee, Biing-Hwang Juang, and Frank K. Soong, "The use of cohort normalized scores for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 599–602.

[13] D. E. Sturim and D. A. Reynolds, "Speaker adaptive cohort selection for Tnorm in text-independent speaker verification," in *Proceedings of ICASSP*, 2005.

[14] N. Dehak Z. Karam, W. M. Campbell, "Towards reduced false-alarms using cohorts," in *submitted to ICASSP*, 2011.