THE MIT LINCOLN LABORATORY 2008 SPEAKER RECOGNITION SYSTEM*

D. E. Sturim†, W. M. Campbell†, Z. N. Karam†•, D. A. Reynolds†, F. S. Richardson†

†MIT Lincoln Laboratory, •MIT

ABSTRACT

In recent years methods for modeling and mitigating variational nuisances have been introduced and refined. A primary emphasis in this years NIST 2008 Speaker Recognition Evaluation (SRE) was to greatly expand the use of auxiliary microphones. This offered the additional channel variations which has been a historical challenge to speaker verification systems. In this paper we present the MIT Lincoln Laboratory Speaker Recognition system applied to the task in the NIST 2008 SRE. Our approach during the evaluation was two-fold: 1) Utilize recent advances in variational nuisance modeling (latent factor analysis and nuisance attribute projection) to allow our spectral speaker verification systems to better compensate for the channel variation introduced, and 2) fuse systems targeting the different linguistic tiers of information, high and low. The performance of the system is presented when applied on a NIST 2008 SRE task. Post evaluation analysis is conducted on the sub-task when interview microphones are present.

Index Terms— speech processing, speaker recognition

1. INTRODUCTION

A main theme in the 2008 NIST speaker identification evaluation (NIST-SRE) was the use of data gained from auxiliary microphones recorded in an office environment. The significant task in 2008 was an interview, simultaneously recorded on 14 microphones [1]. Auxiliary microphones were first introduced in the 2005 NIST SRE and were continued in 2006. The task in prior years used only the auxiliary microphone data in the speaker detection phase and not in enrollment. The 2008 NIST SRE uses the auxiliary microphone data in both enrollment and detection.

The potential variation of microphones and recording environment influenced our design philosophy for the 2008 NIST-SRE. Our goal was to utilize recent advances in variational modeling, factor analysis [2] and nuiance attribute projection [3], to mitigate the effects of microphone and room acoustics. To this end, we concentrated on the cepstral based, Gaussian mixture modeling (GMM) and support vector machine (SVM) systems.

In addition to the core spectral systems, we also maintained a goal of designing efficient high-level systems based on phone recognition. We have found in past years that speaker verification systems based on a phone recognizer yielded good performance while at the same time avoided the complexity of a speech-to-text STT system. Additionally, we have found that speaker verification performance of the STT-based systems suffered when applied to the auxiliary microphone data. The high-level systems developed for this evaluation were an SVM trigram, an SVM keyword system, and an SVM maximum-likelihood linear regression (MLLR) system with NAP.

Consideration of systems that would fuse well also influenced which systems were fielded for the evaluation. It has been historically verified from our organization, as well has many others in the SRE community, that fusion is most successful with "lower-level" spectral systems and higher-level STT-based systems. This point of view is based purely on empirical evidence over recent years. The belief that follows is that the two classes of speaker verification systems provide complementary information. The lower-level systems yield spectral information about the talker whereas the higher-level systems capture cues such as information about prosody, phonotactics, idiolect, and dialog.

We outline in this paper the systems, techniques, and experimental results for the new systems. Sections 2.2 and 2.3 describes the cepstral systems. The base phone recognition system is presented in section 3. Sections 3.1, 3.2 and 3.3 describe the high-level tokenizer and the multiple modeling systems. In Section 4, we describe experimental evaluation of the system on the NIST SRE 2008 data. Finally, we present some post evaluation analysis on the condition pertaining to the auxiliary microphone.

2. CEPSTRAL SYSTEMS

2.1. Front end processing for cepstral systems

The cepstral-based systems used a common set of speech activity detection marks from a GMM-based speech activity detection (SAD) system and an adaptive energy-based SAD.

The features used for recognition were MFCCs. The MFCCs consisted of 19 cepstral coefficients and deltas to produce a 38 dimensional feature vector. The feature vector stream is processed through SAD to eliminate non-speech vectors. RASTA, CMS, and variance normalization are then applied to the feature stream.

To combat additive noise in the microphone channel two noise reduction techniques were employed, 1) steady tone removal and 2) wideband noise reduction, were applied in series as preprocessor step to MFCC feature processing. The steady tone suppression method used very long analysis window, 8 seconds, to exploit the coherent integration of the Fourier transform. The wideband noise reduction algorithm used an adaptive Wiener-filter approach directed toward preserving the dynamic components of a speech signal while effectively reducing noise. Greater detail can be found in [4].

2.2. GMM LFA System

The base system was the MITLL GMM-UBM speaker detection system with 2048 mixtures, fully described in [5], and is similar to that used in previous evaluations.

The GMM Latent factor analysis (LFA) was based directly on the work presented in [6]. The approach models session variability through a low dimensional subspace projection in both training

*This work was sponsored by the Department of Defense 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.

and testing. The session variability is modeled as a low-dimensional additive bias to the model means:

$$m_i(s) = m(s) + Ux(s) \tag{1}$$

where $m_i(s)$ and $m(s)$ are supervectors of stacked GMM means [6, 7]. The $m_i(s)$ is the supervector from the i-th session of talker s whereas the $m(s)$ is the session independent term of talker s.

Training of the low-rank transformation matrix U was generated directly as described in [8] and not iteratively. Z-norm followed by T-norm was also performed on the scores.

The LFA system was applied gender dependently. Factor analysis was performed using session loading matrices generated with class-variation constrained to be speaker only. However, in the presence of a microphone channel the loading matrix used was one generated with class-variation constrained to be speaker and session. A corank of 64 was used thoughout conditions.

2.3. SVM GMM Supervector System (SVM GSV)

The SVM GMM supervector system is based on [3]. The frontend used 19 MFCCs and deltas followed by mean and variance normalization. For microphone data, stationary tone reduction followed by wideband noise reduction was used as in [4].

GMM supervectors were derived using MAP adaptation of means only with a relevance factor of 4 on a per utterance basis. The kernel inner product used was

$$\begin{aligned} K(g_a, g_b) &= \sum_{i=1}^{N} \lambda_i \mathbf{m}_{a,i}^t \boldsymbol{\Sigma}_i^{-1} \mathbf{m}_{b,i} \\ &= \sum_{i=1}^{N} \left(\sqrt{\lambda_i} \boldsymbol{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_{a,i} \right)^t \left(\sqrt{\lambda_i} \boldsymbol{\Sigma}_i^{-\frac{1}{2}} \mathbf{m}_{b,i} \right) \end{aligned} \tag{2}$$

as in prior work. In equation (2), $\mathbf{m}_{*,i}$ are the adapted means, λ_i are the mixture weight of the UBM, and Σ_i are the UBM covariances. SVMs were trained using SVMTorch. The NAP corank was 64.

3. HIGH-LEVEL SYSTEMS

This year we used a single cross-channel phonetic recognizer based on the Brno University (BUT) design [9] as the source of tokens for our high level systems. We used the tokenizer to generate 1-best and lattices from which we extracted N-grams, keywords and phonetic class alignments for our three high-level systems. Thus, we eliminated the need to run an STT system which greatly reduces the computational overhead for the high-level systems.

The tokenizer uses a standard three state left-to-right HMM and a null grammar. There are two key components for generating HMM state posteriors: TRAPS [10] which are long time-span time-frequency features, and feedforward artificial neural nets. The tokenizer was trained on approximately 10 hours of Switchboard-2 phase-1 data re-recorded through a subset of the microphones used in the this years evaluation. The data was phonetically segmented using an STT system, and the resulting system used 49 monophones including silence.

3.1. SVM Trigram

3.1.1. SVM N-gram Language Modeling

The SVM token systems use a bag-of-N-grams approach similar to [11]. For a lattice, W, joint probabilities of the unique N-grams, w_j, on a per conversation basis are calculated, $p(w_j|W)$ and are mapped to a sparse vector with entries $D_j p(w_j|W)$ where

$$D_j = \min \left(C_j, \sqrt{1/p(w_j|\text{all})} \right) \tag{3}$$

and $C_j = 10000.0$. The probability $p(w_j|\text{all})$ in (3) is calculated from the observed probability across all classes.

The general weighting of probabilities is then combined to form a kernel; for two lattices, W and V, the kernel is

$$K(W, V) = \sum_j D_j^2 p(w_j|W) p(\hat{w_j}, w_j|V). \tag{4}$$

SVM training and scoring are performed as in prior work [11].

3.1.2. Trigram SVM System Description

The approach described above in 3.1.1 was used to train trigram SVMs for the cross-channel tokenizer described above. The trigram system was only used in the MIT/LL primary system when the data (enrollment or verification) did not come from the interview microphone channel.

Two slightly different system configurations were used depending on whether the test message came from a telephone channel or from an auxiliary microphone channel. Both configurations used a NAP corank of 32, and T-norm speakers from the NIST SRE Eval04 data set. In the case where both training and testing were on 4w data, we used gender dependent models trained on a larger background data set. When the test data came from an auxiliary microphone channel, we used gender independent models trained on a smaller background data set. Also, NAP training consisted of Eval04 and Switchboard 2 part 1 data when testing on 4w data, and Eval05 auxiliary microphone data when training on the auxiliary microphone data.

3.2. SVM Keyword System

The SVM keyword system, which is based on the SVM trigram system described in 3.1, includes specially selected variable length N-grams of up to order 17. The varying length N-grams were selected by taking the most discriminative bigrams from an ASR based word SVM token system [12]. All of the phones in each pronunciation of the top and bottom ranking bigrams were concatenated to make the set of N-gram "keywords".

For the keyword system used in the 2008 SRE evaluation, we selected the 400 top and 400 bottom ranking word bigrams from models trained on the NIST 2004 SRE data set. The total number of keyword N-grams (including multiple pronunciations) totaled between 75 and 78 thousand. Posterior counts were collected efficiently using compacted parse trees across the lattices and accumulating counts at the parse tree leaf nodes [13].

The keyword system was only applied to the eight conversation (8conv) task as part of the MIT/LL secondary submission. The SVM kernel used in the keyword system was identical to the kernel described in 3.1.1. Two different configurations were used for 4w telephone and auxiliary microphone test messages mirroring the configurations used for the the trigram system described in 3.1.2. Both configurations used a NAP corank of 32, and T-norm speakers from the NIST SRE Eval04 8conv data set.

3.3. SVM MLLR NAP

The SVM MLLR system is a SVM GSV system with two class MLLR adaptation used instead of MAP adaptation, as in [14]. The system starts with a 512 mixture GMM (UBM), comprised of a weighted convex combination of two 256 mixture GMMs representing broad phonetic classes, sonorants and obstruents:

$$g_{UBM} = \mu_s \sum_{i=1}^{256} \lambda_{s,i} \mathcal{N}(\mathbf{m}_{s,i}, \mathbf{\Sigma}_{s,i}) + \mu_o \sum_{i=1}^{256} \lambda_{o,i} \mathcal{N}(\mathbf{m}_{o,i}, \mathbf{\Sigma}_{o,i}),$$

where $\mu_s = .71$ and $\mu_o = .29$ are class mixing weights, based on the class priors in the background. Open-loop phonetic classification is used to assign the frames of each utterance to either of the two classes. The means of the UBM are then adapted to each utterance using two-class MLLR adaptation, where the classes are the sonorants and obstruents. The adapted, sonorant and obstruent, means are then stacked to form GMM supervectors which are used as features for the SVM classifier with the following GSV kernel:

$$\begin{aligned} K(g_a, g_b) &= \mu_s \sum_{i=1}^{256} \lambda_{s,i} \mathbf{m}_{s,i}^{a\ T} \mathbf{\Sigma}_{s,i}^{-1} \mathbf{m}_{s,i}^{b} \\ &\quad + \mu_o \sum_{i=1}^{256} \lambda_{o,i} \mathbf{m}_{o,i}^{a\ T} \mathbf{\Sigma}_{o,i}^{-1} \mathbf{m}_{o,i}^{b}. \end{aligned}$$

This system also applies ZT-Norm and NAP with a corank of 32.

4. EXPERIMENTS

4.1. Experimental and System Setup

Experiments were performed on the NIST 2008 SRE data set. Enrollment/verification methodology and the evaluation criterion (minDCF) were based on the NIST SRE evaluation plan [15]. The systems as described in Sections 2 and 3 used background derived from Switchboard and Fisher corpora. For telephone conditions, T-Norm models and Z-norm utterances were drawn from the 2004 NIST SRE. For microphone conditions, T-Norm models and Z-Norm utterances were taken from the NIST SRE 2005 auxiliary microphone data. Nuisance subspace training for telephone conditions was performed using 2004 NIST SRE data. Subspace training for the microphone conditions was performed using 2005 SRE auxiliary microphone data.

Fusion was accomplished using a standard multi-layer perceptron as in prior evaluation systems [16]. Fusion was trained from scores on the NIST SRE 2006 evaluation set. Results were obtained for both the English only task (ENG) and for all trials (ALL) which includes speakers that enroll/verify in different languages. Additionally, when microphone data was present the noise-reduction frontend was applied as described in section 2.1.

4.2. Results

The system results of the 2008 NIST-SRE for our primary fusion system is presented in Table 1. The results are broken out according to the four training condition categories:

- Short - single training utterance from either telephone (tel) or conversational microphone or interview microphone (intmic)
- 3conv - three training utterances all from telephone channels
- 8conv - eight training utterances all from telephone channels
- Long - single long training utterance ($>$ 8min.) from interview microphone

Fusion was accomplished with the following combinations. For short telephone trials, all systems except the SVM keyword were used. For the 8conv case, all systems were used. For cross microphone-telephone trials, the SVM trigram, SVM GSV, and GMM LFA were used. Finally, for microphone only trials, only the cepstral systems were used. We note that specializing fusion and system design to every subcondition (training channel type, testing channel type, number of conversations) was critical for good performance.

From the results in Table 1, we can make the following observations. First, we note the performance hit incurred by the submitted telephone system when comparing English trials tasks versus all trials. The performance drop-off occurs for all the training categories: (short:tel-tel, 3conv:tel-tel and 8conv:tel-tel). When comparing to the English subsets for these conditions, (short:tel-tel-English, 3conv:tel-tel-English and 8conv:tel-tel-English), we draw the supposition that there may be language mismatch in the background model and T-norm-cohort training data. The performance degradation is noted here, but we will reserve analysis and mitigation of these effects for later study.

4.3. The Interview Task

A main goal entering the evaluation was to try to understand the effects of the auxilary microphones on speaker recognition performance. We can observe the degradation in the conditions involving the auxilary interview microphones Table 1. Considering the results for the two conditions 1) short:Tel-Tel-English and 2) short:intmic-intmic*, we observe the performance degradation in the presence of interview microphones. During post evaluation analysis, we noted three areas where we could improve performance of our systems. Since our submitted system consisted of a fusion of the GMM-LFA and GSV-NAP systems, we only considered improvements to these systems for this task.

First, we noted that performance of the GMM-LFA system could be vastly improved with the use of interview-microphone development data provided by NIST prior to the evaluation. The development consisted of speech utterances from six talkers (3 male and 3 female). Initially, we only used this data for threshold setting. However, performance can be improved if the data is utilized in training the transfer-loading matrix U of equation 1 and given as much weight as the other data used to train the loading matrix through stacking the loading matrices in a similar fashion to [17]. The second area for improvement was to use a speech activity detector utilizing the logical AND of VAD and ASR transcripts from a clean lapel microphone provided by NIST. The third improvement was to use both LPCC and MFCC SVM GSV systems and linearly fuse the scores [16].

Figure 1 and Table 2 present the results of implementing the improvements described above. The DET plot shows a significant overall improvement. Additionally, the equal error rate and DCF points of Table 2 are closer to the phone channel results of Table 1. One challenge with these results for future evaluations is that the processing is based upon NIST-provided oracle knowledge; that is, knowledge of the microphones and good VAD are not necessarily available in real speaker recognition applications.

5. CONCLUSIONS

We have presented the speaker recognition used for the NIST 2008 SRE. We have described systems for speaker recognition using fac-

* The interviews were conducted in English.

Table 1. Summary of performance on the NIST SRE 2008 task with the primary system

	Task *train-test*	EER (%)	minDCF (×100)
short	Tel-Tel	7.0	3.6
	Tel-Tel (English)	3.3	1.6
	Tel-Tel (US-English)	3.5	1.6
	Tel-Phnmic	5.5	2.2
	intmic-Tel	6.9	2.9
	intmic-intmic	5.2	2.6
3conv	Tel-Tel	4.9	2.8
	Tel-Tel (English)	2.2	1.0
	Tel-Tel (US-English)	2.0	1.0
	Tel-Phnmic	3.5	1.4
8conv	Tel-Tel	3.4	2.1
	Tel-Tel (English)	1.2	0.6
	Tel-Tel (US English)	1.4	0.5
	Tel-Phnmic	2.4	0.7
long	long-long	4.8	2.0
	long-Tel	5.7	2.3
	long-intmic	4.5	1.7

Table 2. Performance on the NIST SRE 2008 task with post evaluation improvements.

	Task *train-test*	EER (%)	minDCF (×100)
short	intmic-intmic	2.71	1.6

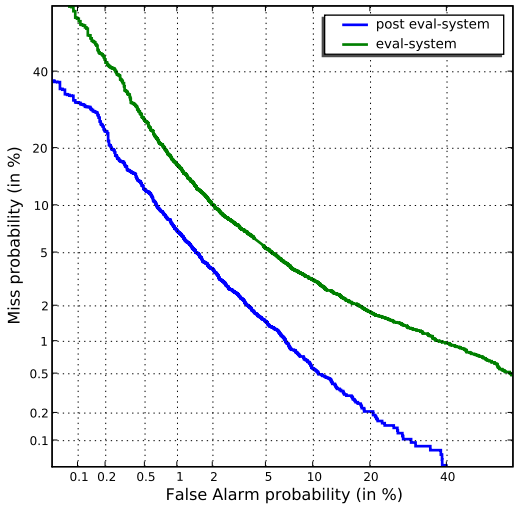

Fig. 1. Detection Error Trade-off comparing the submitted evaluation system and the post evaluation systems for the condition (short: intmic-intmic)

tor analysis, discriminative techniques, channel compensation, and high-level speaker recognition. Post-evaluation analysis showed that for the interview microphone in training and test, we can get close to telephone channel level performance. Future work will focus on the interview task, more in-depth analysis and new compensation techniques to create deployable systems.

6. REFERENCES

[1] Christopher Cieri, Linda Corson, David Graff, and Kevin Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Interspeech*, 2007.

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

[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] 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," *Proceedings of ICASSP*, vol. 4, pp. IV–49–IV–52, April 2007.

[5] Douglas A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *EuroSpeech*, 2006.

[7] 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.

[8] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[9] Petr Schwarz, Matejka Pavel, and Jan Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proceedings of ICASSP*, 2006, pp. 325–328.

[10] Hynek Hermansky and Sangita Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proceedings of ICASSP*, 1999, vol. 1, pp. 289–292.

[11] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "Phonetic speaker recognition with support vector machines," in *Advances in Neural Information Processing Systems 16*, Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.

[12] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, "High-level speaker verification with support vector machines," in *Proceedings of ICASSP*, 2004, pp. I–73–76.

[13] W. M. Campbell and F. S. Richardson, "Discriminative keyword selection using support vector machines," in *Neural Information Processing Systems*, 2008.

[14] Zahi N. Karam and William M. Campbell, "A multi-class MLLR kernel for SVM speaker recognition," in *ICASSP*, 2008.

[15] "The NIST year 2008 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2008/index.htm, 2008.

[16] W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced complexity recognition," in *Proceedings of ICASSP*, 2007.

[17] N. Dehak P. Kenny, R. Dehak, V. Gupta, and P. Dumouchel, "The role of speaker factors in the NIST extended data task," in *IEEE Odyssey*, 2008.