ROBUST SPEAKER RECOGNITION WITH CROSS-CHANNEL DATA: MIT-LL RESULTS ON THE 2006 NIST SRE AUXILIARY MICROPHONE TASK[*]

D.E. Sturim, W.M. Campbell, D.A. Reynolds, R.B. Dunn, T.F. Quatieri
{sturim,wcampbell,dar, rbd,tfq}@ll.mit.edu

MIT Lincoln Laboratory, Lexington, MA USA

ABSTRACT

One particularly difficult challenge for cross-channel speaker verification is the auxiliary microphone task introduced in the 2005 and 2006 NIST Speaker Recognition Evaluations, where training uses telephone speech and verification uses speech from multiple auxiliary microphones. This paper presents two approaches to compensate for the effects of auxiliary microphones on the speech signal. The first compensation method mitigates session effects through Latent Factor Analysis (LFA) and Nuisance Attribute Projection (NAP). The second approach operates directly on the recorded signal with noise reduction techniques. Results are presented that show a reduction in the performance gap between telephone and auxiliary microphone data.

Index Terms**—** Speaker recognition, Speech enhancement, Microphones, Acoustic noise

1. INTRODUCTION

One of the enduring challenges for automatic speaker verification is dealing with variability between training and testing speech. This variability arises from several intrinsic (speaker related) and extrinsic (audio quality related) factors, but many of the dominant and identifiable causes of train/test mismatch are extrinsic factors due to changes in the microphone used, and the acoustic environment in which the speech was recorded. Over the past 12 years, the NIST speaker recognition evaluations (SREs) have focused research effort on addressing aspects of the mismatch challenge by using telephone speech collected from a wide sampling of telephone instruments (landline, cellular, carbon-button, electret, etc.) from many different acoustic environments (indoors, outdoors, etc) and by designing train and test scenarios using speakers' speech recorded from different telephone numbers (and thus presumably different telephone instruments and locations). This emphasis has produced new compensation techniques in the feature, model and score domains, that have steadily driven down the error rates under mismatched telephone speech conditions. To further raise the bar on this challenge and to address new application domains, NIST introduced an auxiliary microphone (auxmic) task in the 2005 and 2006 SRE that provided a new cross-channel test scenario using multiple, non-telephony microphones.

In the auxmic task, the training speech for a speaker is collected from a single telephone number, but the test speech comes from one of eight different microphones. Such a scenario could be encountered in applications needing portability of speaker models across multiple recording domains, for example, in a forensic voice-comparison of a telephone threat to a suspect recorded with a room microphone. The mismatch in the auxmic task arises not only from the use of different microphones, but also from effects on the speech induced by the microphone placement and the recording room characteristics, such as reduced SNR and reverberation. Thus compensation techniques that address both microphone and noise mismatch are important for success on this task.

In this paper, we present an analysis of results from the MIT-LL systems applied to the 2006 NIST SRE auxmic task. We show how GMM and SVM microphone/session compensation techniques developed for telephone speech can be successfully applied to the auxmic task. We further demonstrate that applying speech enhancement algorithms for noise and tone removal to pre-process the auxmic data can greatly improve speaker verification performance.

2. AUXILIARY MICROPHONE DATA

The data used in the auxmic task comes from the cross-channel recordings collected as part of the Mixer telephone corpus by LDC [1]. The goal of the cross channel collection was to record one side of a series of Mixer telephone conversations on a variety of microphones. The microphones were chosen to represent certain target settings such as the microphones used in courtrooms, interview rooms, and cellular telephones. Participants placed calls to the Mixer robot operator while being recorded simultaneously on the cross channel recorder. Table 1 lists the eight microphones used in the auxmic task.

Table 1: Microphone types in the auxiliary microphone task.

Num.	Description	Microphone
1	Studio mic	Audio Technica AT3035 Cardioid Condenser
2	Courtroom mic	Shure MX418S Supercardioid Gooseneck Mic
3	Distant mic	Audio Technica Pro 45 Cardioid Condenser Hanging Mic
4	Microcassette mic	Olympus Pearlcorder S725
5	Ear miniboom mic	Jabra® EarWrap Headset Radio Shack #43-1914
6	Cell earbud	Motorola Earbud Handsfree (SYN8390)
7	Conference room mic	Crown SoundGrabber II pressure-zone mic (PZM)
8	PC-style stand	Radio Shack Desktop Mic with Noise Canceling #33-3031

During recording, the participant is wearing microphones 5 and 6, microphones 1, 2, 4, 7, and 8 are placed on a table in front of the subject, and microphone 3 is placed high and across the room from

[*] This work is 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.

the subject. The auxmic data was collected at three locations: ISIP (Mississippi), LDC (Pennsylvania) and ICSI (California). Some effort was made to use similar size rooms and microphone placements at the three sites, but this was not rigorously enforced (e.g., no control was made for reflecting surfaces in the rooms or noise sources, such as air vents). As shown later, the distant microphone (3) is the most affected by differences in room acoustics.

There were 8 speakers recorded at ISIP, 109 at LDC, and 61 at ICSI. The 2005 SRE auxmic data came from the 8 ISIP speakers and 89 LDC speakers, while the 2006 SRE auxmic data came from 20 LDC speakers and the 61 ICSI speakers.

Human analysis of the auxiliary microphone data reveals two main effects 1) distortions due to the microphone type, and 2) the presence of tones and broadband noise. Session compensation will try to address the first of these effects while noise reduction will address the later.

3. VERIFICATION SYSTEMS

For this task, state-of-the-art verification systems using Gaussian mixture models with universal background model (GMM-UBM) [2] and support vector machines using a generalized linear discriminant sequence kernel (SVM-GLDS) [3] developed at MIT-LL were applied. Both systems also used new channel/session compensation techniques of the models developed to help mitigate mismatch degradations for telephone speech.

3.1 GMM-UBM

In the GMM-UBM system, a speaker model is created by MAP adaptation from a speaker-independent UBM. The UBM was trained using Switchboard II and SRE-04 data. A standard MFCC front-end is used, producing 19 MFCCs extracted over the telephone bandwidth (300-3100 Hz) with 19 delta MFCCs appended. The features are processed with RASTA filtering and have mean and variance normalization applied to minimize session effects.

To compensate for session effects in the model domain, a form of Latent Factor Analysis (LFA) based directly on the work presented in [4] and related to initial work done by [5, 6] is applied. The session variability is modeled as a low-dimensional additive normal bias, *x(s)* 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 mean vectors. 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 session loading matrix *U* is generated directly without iteration as described in [7]. The selection of data used to train the loading matrix for the auxiliary microphone task is described in following sections.

Compensation in the score domain was done by applying Z-norm followed by T-norm. The Z-norm imposter test messages were taken from the Switchboard II phase 1-5 corpora and the T-norm speakers were drawn from the SRE04 corpus (364 female and 243 male models).

3.2 SVM-GLDS

The SVM-GLDS system uses a polynomial-based kernel with a degree 3 basis of monomials. Background data used in the SVM training were obtained from a subset of the English portion of the Fisher corpus. New in the SRE06 MIT-LL system, two front-ends were used. The first was the same MFCC front-end as used for the GMM-UBM system. The second was an LPCC front-end which generated 18 LPCCs plus deltas from 12 LP coefficients. Speaker models and scores are calculated independently for both feature sets, and the resulting scores are fused with a linear combination.

The SVM-GLDS system with NAP relies on the kernel computation between expansion vectors of two talkers, (n^a and n^b):

$$K(n^a,n^b) = b(n^b)^t \; P \; b(n^b) \tag{2}$$

where *P* is the NAP projection [8] and *b(·)* is the SVM polynomial expansion. The NAP projection matrix *P* is related to the LFA session loading matrix *U* see — [9]. This similarity in computation allows us to use the same data with the same constraints in training the *P* and *U* matrices. This will be expounded upon in Section 4.

NAP projection was applied in training to the backgrounds and speaker data. SVM models were obtained using the GLDS kernel and SVM Light. Model compaction was used to reduce the size of models.

For scoring, we computed an inner product between the average expansion and the SVM model. T-norm was also performed drawing from SRE04 cohorts.

4. AUXMIC SESSION COMPENSATION

The first approach is to mitigate the effects of auxiliary microphone on the testing channel by utilizing subspace modeling-modeling LFA and NAP. Both of these techniques compensate for unwanted variation through attenuation of that information. LFA models the variation with low-dimensional normally distributed latent factors, whereas NAP models variation in extremely high dimensions and excises out "nuisance" dimensions. In both techniques, information about what types of variability to suppress is obtained through the data used to estimate the sample covariance matrices (U and P).

The constraints on the training data for the LFA session loading matrix and NAP projection matrix were:

1) **Tel:** Session utterances coming only from telephone channel.
2) **Pool**: Session utterances drawn from the auxiliary microphone and telephone data are 'pooled' together.

The drawback in using the auxiliary microphone data is the limited number of speakers in the SRE 2005 auxmic data. Fortunately the 98 speakers also participated in the SRE 2005 telephony corpus collection, so additional 2-5 telephone sides could be added to each speaker's existing 8 auxmic sessions.

5. NOISE REDUCTION

Some of the auxmic data, most notably channel 3, was contaminated by both tones and wideband noise. To combat this contamination two noise reduction techniques, steady tone removal and wideband noise reduction, were applied in series as preprocessor steps to MFCC and LPCC features processing.

5.1 Steady Tone Suppression

Current methods of steady tone suppression using comb filters or short-time analysis/synthesis are inadequate for the closely spaced and inharmonic tones with low SNR observed in the data. The method we apply in this paper strives to address the limitations of other methods by using a very long analysis window to exploit the coherent integration of the Fourier transform [10]. An important aspect of this tone reduction method is that it introduces little amplitude and phase distortion in the surrounding signal, thus preserving components of the signal important for recognition by humans or machines.

The steps in the technique, which provides high frequency resolution and robustness, are as follows. First, the audio input is windowed using an 8 second long hamming window, and its Fourier transform is computed. Next, the magnitude spectrum is whitened by subtracting a smoothed version of the original. Tones are detected by applying a threshold to the whitened spectrum and at each tone a Gaussian shaped template with a 2-Hz bandwidth is subtracted from the magnitude. The resulting spectrum is inverted and a complete speech signal estimate is obtained through an overlap-and-add reconstruction with neighboring 8-second segments.

5.2 Wideband Noise Reduction

Standard noise suppression algorithms can distort dynamic speech-signal characteristics, such as transient plosives, formant motion, and vowel onsets which may be essential in contributing to distinguishing speech and speaker characteristics. In this paper, we use an adaptive Wiener-filter approach directed toward preserving the dynamic components of a speech signal while effectively reducing noise [11, 12]. A distinguishing component of the approach is an estimate of the speech spectrum, required by the Wiener filter, using a measure of spectral change that allows robust and rapid adaptation of the filter to speech events. The approach reduces speech distortion in Wiener filtering by making the time constant that controls smoothing of the speech spectrum a time-varying parameter. In particular, the time constant is selected so that little temporal smoothing is introduced in rapidly-changing regions and increased smoothing is performed in more stationary regions. Our measure of spectral change is provided by a dynamically-smoothed spectral derivative. The approach is consistent with temporally shaping noise to fall within certain regions of least perceptual sensitivity [11, 12].

The framework for implementing the suppression algorithm is an overlap-add approach, with the use of a very short 4-ms triangular analysis window and a 1-ms frame interval. Without suppression, the analysis/synthesis is an identity, while being consistent with sufficient temporal resolution to maintain discrimination of dynamic speech components.

6. EXPERIMENTS AND ANALYSIS

In this section, we present an analysis of results from the 2006 SRE auxmic task using the above verification systems and compensation techniques. For these experiments, we focus on the one conversation training condition, in which speaker models are trained using one side of a telephone conversation (approximately 2.5 minutes of speech) and tested against a conversation side recorded on one of the auxiliary microphones. Performance is assessed in terms of equal error rates (EERs) and minimum decision cost functions (minDCF). See the 2006 SRE evaluation plan at http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf for details.

Table 2 summarizes overall performance on the auxmic task for the GMM-UBM and SVM-GLDS systems using various forms of session compensation and with noise reduction applied to the auxmic speech. Comparison with results in the "telephone unprocessed" (telephone sides from the auxmic sessions) column shows that the auxmic data indeed is more challenging than the telephone data. It is also clear from the results that session compensation and noise reduction provide large gains in performance both separately and jointly.

Table 2: Summary of overall auxmic performance (EER in % / DCF in %) for the GMM-UBM and SVM-GLDS systems with and without session\ compensations and noise reduction.

Classifier	Session Compensation	Telephone Unprocessed	Auxmic Unprocessed	Auxmic NR
GMM	none	-	9.89 / 3.97	7.74 / 3.41
	LFA-tel	3.75 / 1.48	9.07 / 3.69	6.94 / 3.09
	LFA-pool	-	5.63 / 2.37	**4.60 / 2.17**
SVM	none	-	17.74 / 6.21	14.24 / 5.26
	NAP-tel	2.88 / 1.29	9.70 / 3.81	6.78 / 2.76
	NAP-pool	-	7.20 / 2.82	**5.71 / 2.29**

We next analyze results broken out by microphone and collection site. As discussed in the introduction, the degradations introduced by a microphone are a function not only of the transducer but also of the microphone placement and room acoustics. In Figure 1, we plot the estimated SNR for speech from the eight microphones for the two collection sites used in the 2006 SRE auxmic data. This plot provides both a rough ranking of noise degradations per microphone (a proxy for placement) and the correlation of microphone SNRs per site (a proxy for inter-site consistency). Based on placement, the low SNR of microphone 3 for both sites matches expectations. Microphone 5, however, appears to be abnormal at the LDC site (It was later found that this microphone had a battery failure which caused extreme under recording). In general the SNRs were correlated between sites, but lower at LDC.

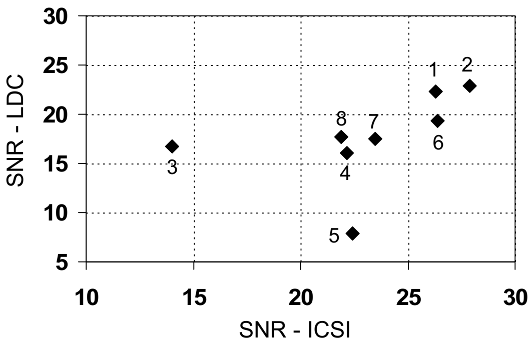

Figure 1: Scatter plot of SNR for auxmic microphones for LDC and ICSI collection sites used for 2006 SRE auxmic data.

The EER per microphone broken out by site is shown in Figure 2 for the GMM LFA-pool system without noise reduction. We can observe some limited correlation of performance with SNR. This figure demonstrates the microphone-dependent variability at a single site and the site-dependent variability of a single microphone arising from operator error (e.g., 5) and room acoustics and/or placement (e.g., 3 and 4).

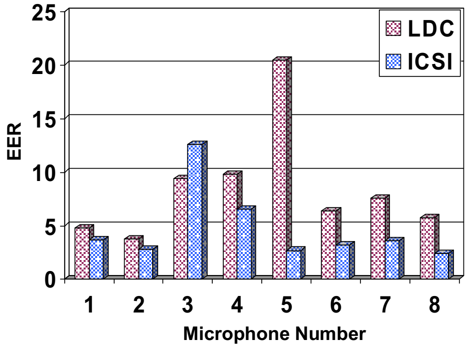

Figure 2: Collection site breakout of EER versus microphone for GMM-LFA-pooled system without noise reduction

The microphone-dependent performance effect of the noise reduction is shown in Figure 3 for the SVM with NAP-pool. These results combine auxmic data from the sites. A similar profile is seen for the GMM system when using LFA-pool compensation. As expected the noise reduction has the largest benefit on the distant microphone (3) but, also of importance, it is benign to those microphones without noise problems. The noise reduction appears to be most beneficial for SNRs < 20dB.

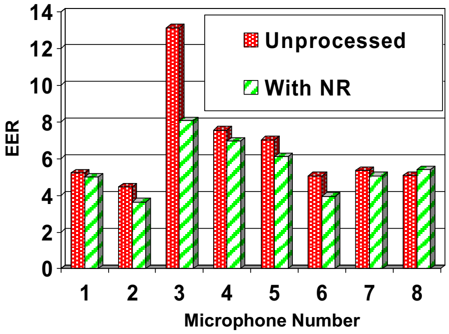

Figure 3: Per-microphone EER of the SVM with NAP-pool using unprocessed and noise reduced auxmic speech.

Figure 4 shows the microphone-dependent performance effects of session compensations for the GMM-LFA system after noise reduction is applied. Here we see the additional improvements in using pooled data to train the LFA compensation which reduces the ERR for all microphones, most notably for microphones 3 and 4.

Lastly, linear combination fusion of GMM and SVM systems with pooled session compensation and noise reduction pre-processing further reduces the error rate on the 2006 SRE auxmic task to EER= 4.04% and minDCF*100= 1.69.

7. CONCLUSIONS

In this paper we presented systems and compensations techniques for robust verification under challenging cross-channel conditions in the auxiliary microphone task of the NIST 2006 SRE. To deal with transducer mismatch, we demonstrated that session compensation techniques of LFA and NAP can be greatly improved by utilizing pooled data from telephone and auxiliary microphone data. To address mismatch from additive tones and broadband noise in the microphone speech, we demonstrated that pre-processing the audio with tone and noise reduction algorithms significantly improves accuracy for low SNR audio with no loss for high SNR audio. These compensations were shown to work very well individually and even better jointly for both a GMM and SVM based verifier. Results showed relative reductions of 55-67% in EER.

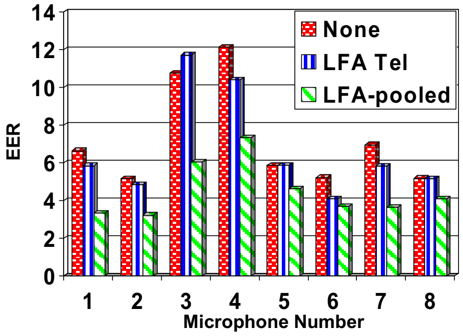

Figure 4: Per-microphone EER of the GMM with different session compensations using noise reduced auxmic speech.

REFERENCES

[1] J. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. Martin, and M. Przybocki, "The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition Research and Evaluation," in *In Odyssey Workshop*, Toledo, Spain, 2004, pp. 29-32.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[3] W. M. Campbell, "A SVM/HMM system for speaker recognition," in *ICASSP*, 2003, pp. II-209-212.

[4] R. Vogt, B. Baker, and S. Sridharan, "Modelling Session Variability in Text-Independent Speaker Verification," in *EuroSpeech*, 2006.

[5] P. Kenny, P. Boulianne, G. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," in *ICASSP* 2005.

[6] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling With Sparse Training Data," *IEEE Transactions On Speech And Audio Processing,* vol. 13, May 2005

[7] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation,* vol. 11, 1999.

[8] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances In Channel Compensation for SVM Speaker Recognition," in *ICASSP*, Philadelphia PA, 2005.

[9] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification Using a GMM Supervector and NAP Variability Compensation," in *ICASSP* 2006.

[10] R. B. Dunn and T. F. Quatieri, "Improved audio tape enhancement," MIT Lincoln Laboratory June 2004.

[11] T. F. Quatieri and R. Baxter, "Noise reduction based on spectral change," in *IEEE WASPAA*, New Paltz, NY, 1998.

[12] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *ICASSP*, 2002.