Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Exploring the Impact of Advanced Front-End Processing on NIST Speaker Recognition Microphone Tasks*

W. M. Campbell, D. Sturim, B. J. Borgstrom, R. Dunn,
A. McCree, T. F. Quatieri, D. A. Reynolds

MIT Lincoln Laboratory

{wcampbell,sturim,jonas.borgstrom,rbd,mccree,quatieri,dar}@ll.mit.edu

Abstract

The NIST speaker recognition evaluation (SRE) featured microphone data in the 2005-2010 evaluations. The preprocessing and use of this data has typically been performed with telephone bandwidth and quantization. Although this approach is viable, it ignores the richer properties of the microphone data—multiple channels, high-rate sampling, linear encoding, ambient noise properties, etc. In this paper, we explore alternate choices of preprocessing and examine their effects on speaker recognition performance. Specifically, we consider the effects of quantization, sampling rate, enhancment, and two-channel speech activity detection. Experiments on the NIST 2010 SRE interview microphone corpus demonstrate that performance can be dramatically improved with a different preprocessing chain.

1. Introduction

The 2005-2010 NIST SREs have had microphone data in a variety of formats. In the 2005 and 2006 evaluations, recordings of one side of a telephone conversation were made from multiple microphone channels at 48 kHz, downsampled to 8 kHz, and μ-law quantization was applied. In these recordings, co-talker interference was minimal and typically speech activity detection (SAD) could be performed with one (microphone) side of the conversation. In the 2008 and 2010 evaluations, the conversational microphone task was kept, and, in addition, an in-person interview task was added. The interview process introduced the problem of speaker diarization of the recording. In 2008, only one channel was provided, so diarization was performed with either NIST provided two-side SAD or ASR transcripts. In 2010, a noise masked version of the interviewer's lapel microphone recording was available and paired with the interviewee's recording. The continuing paradigm changes of the style and preprocessing impacted system performance and were not systematically examined.

Systems have evolved considerably since the original 2006 NIST SRE, so the effect of preprocessing on recent state-of-the-art systems is important to understand. For the purposes of this paper, we selected two of the MIT Lincoln Laboratory systems which are representative of performance—the inner-product discriminant function (IPDF) system and an iVector system. The IPDF system (or more precisely IPDF-KL) is based on a KL-divergence between adapted GMM UBM models; details can be found in [1, 2, 3]. The iVector system is a Wiener-filter based approach [4] based on the work of Dehak [5].

*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.

For the purposes of this paper, our goal was to understand the basic trends in system performance under a variety of microphone preprocessing conditions:

- Resampling from 16 kHz to 8 kHz and mu-law quantization
- Speech enhancement of data
- Speech activity detection (SAD)
- Methods for diarization

Although we used systems that had a complete suite of processing techniques—channel subspace compensation, z-, t-, s-norm, etc.—we did not attempt to produce the absolute best performance results. One difficulty is that preprocessing changes require extensive hyperparameter retraining and have a long experimental cycle. Another difficulty is that only limited data at higher rates (16 kHz) was available for hyperparameter training (including UBM, subspaces, and cohorts). Therefore, the results in this paper should be interpreted in terms of relative performance and trends.

We first examine the effects of bandwidth on accuracy of the speaker recognition system. We contrast NIST/LDC preprocessing (8 kHz, μ-law) with alternate techniques that are less destructive. We demonstrate that dramatically improved performance is possible.

For speech enhancement, we look at the interplay between SAD, subspace compensation (NAP, TV, Wiener), and speech enhancement. We break-out via experiments the impact that speech enhancement has on SAD and features. Through experiments we show that enhancement methods are providing gains in multiple areas.

For diarization, we consider alternate methods of two-channel SAD rather than the standard technique of using automatic speech recognition (ASR) transcripts. We describe a frequency-dependent method that uses two-side interviewer/interviewee recordings to perform SAD.

In summary, we provide a baseline set of experiments that demonstrate the nuances of processing microphone data. We show that the richness of microphone data leads to alternate processing methods not typically considered for telephone data. In many situations, this alternate processing has a simple form and can be applied to achieve substantial calibration and accuracy improvements on NIST microphone data.

2. Recognition Systems

For our experiments, we used two systems from our NIST 2010 SRE submission [6]. We describe the top-level approaches in the following subsections.

2.1. IPDF-KL system

Inner product discriminant functions (IPDFs) are described in [2, 3]. We use a comparison function from the IPDF framework based on approximations to the KL divergence between two GMMs [1, 3]. 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) [7] 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 64. We then scored using the C_{GM} kernel. Both Z- and T-Norm were applied.

2.2. iVector System

The iVector system is a variant of the total variability system first proposed by [8] and with refinements from [5]. However in this implementation we use the Wiener filtering approach presented in [4].

The iVectors are formed with the following equation:

$$\hat{\mathbf{z}} = U_{tot}^T (\boldsymbol{\Sigma}_{tot} + \boldsymbol{\Sigma}_n)^{-1} (\bar{\mathbf{x}} - \mathbf{m}_o) \quad (2)$$

where $\boldsymbol{\Sigma}_{tot}$ is the total variation that may be seen across all observations (within-class and across-class). PCA modeling is used in forming total variation covariance $\boldsymbol{\Sigma}_{tot} = U_{tot}^T U_{tot}$. A diagonal covariance $\boldsymbol{\Sigma}_n$ is used to model the observation noise. The vector $\bar{\mathbf{x}}$ are the observed supervectors with noise and $\mathbf{m}_o$ is the supervector of the universal background model.

Equation 2 can be interpreted as projection from a Wiener filtered supervector down to the low-dimensional iVector, $\hat{\mathbf{z}}$. The dimension of the supervector is 81920. The dimension of the iVectors are 400.

The system used a gender-independent UBM with 2048 Gaussians. Total variability matrices consisting of 400 eigenvectors trained on gender independent training data. Scoring is performed using WCCN followed by cosine scoring in the iVector space [8]. S-norm was used for score normalization.

2.3. Features

For front-end features, we used a standard MIT LL processing chain for microphone data [6]. Input speech is pre-processed with wide-band noise reduction and tone removal. Speech activity detection was performed in two stages. For interview speech, the first stage performed is an "and" of the ASR transcipts "not"-interviewer segments with a GMM-based SAD system to produce a waveform with interviewee-only speech. For conversational speech, only GMM SAD is used. In the second stage, an energy-based SAD (non-aggressive) was used to eliminate the remaining non-speech frames. MFCCs are extracted every 10 ms using a 25 ms Hamming window. Delta coefficients are found. The MFCCs and deltas are stacked to form feature vectors. Both RASTA and $0/1$ feature normalization are applied to the feature vectors. A total of 40 features is available at 8 kHz bandwidth. For 16 kHz bandwidth, the number of filter banks and cepstra is increased and correspondingly the number of features is 58. The iVector system and IPDF use all available features.

3. Experiments with Sampling Rate and Quantization

3.1. Experimental Setup

Recently, NIST has made available microphone data at 16 kHz for the NIST 2010 SRE. In this section, we consider the effect of the higher sampling rate, conversion to 8 kHz linear, and μ-law on system performance.

Given the large breadth of our experiments, we focused only on a small part of the NIST 2010 SRE. Experiments were limited to short-interview train, short-interview test trials. For all experiments, we limited trials to the NIST SRE 2010 list. Evaluation of systems was done using condition 2—the cross-microphone case. For male speakers, the test set had 691 speaker models with 1,068 true trials and 38,241 false trials. For female speakers, the test set had 829 speaker models with 1,335 true trials and 46,086 false trials. Performance measures were EER and old minimum DCF (oldDCF) from 2008 and prior NIST SREs.

3.2. Data selection

Several sets of data at higher rates were available for experiments:

- The NIST supplied 16 kHz linear-PCM SRE 2010 microphone data.
- LDC data from the original Mixer conversational microphone data from SRE 2005 and SRE 2006 sampled at 48 kHz using linear PCM encoding (full 24 kHz bandwidth).

We note that the 48 kHz data is somewhat challenging to use. The data has original file names and must be mapped to NIST evaluation data using keys. The data is "raw"; that is, the conversations are not duration limited to the standard 5 minutes. Also, the time offsets NIST used for extracting speech segments for the evaluations were not readily available. For our experiments, we used all of the 48 kHz data corresponding to the SRE 2005 and 2006 data set and eliminated the problematic channel 5 recordings. This selection resulted in approximately 4800 utterances with 83 male speakers and 98 female speakers for hyperparameter training (including z- and t-norm cohorts).

3.3. Data processing

To match the rates of the NIST supplied 2010 data (16 kHz) and the original NIST SRE 2010 data (8 kHz), resampling of all of the speech data to both 8 kHz and 16 kHz was performed using multistage multirate methods [9]. All filters were linear-phase FIR and odd length. For downsampling from 48 kHz to 8 kHz, a two stage down-by-3, down-by-2 setup was used. To reduce computation, an interpolated FIR filter design was used. The two-stage filters were length 41 and 621, respectively. Stop-band attenuation was approximately 80 dB to eliminate aliasing; the pass band was approximately 0 to 3800 Hz. For the conversion from 16 kHz to 8 kHz, a length 421 filter was used. Stop-band attenuation and passband specs were the same as the 48 kHz system. For conversion to 16 kHz, a similar design was produced with double the bandwith (0-7600 Hz).

To avoid issues with quantization, all rate conversion was performed in floating point. Signals were gain normalized to have maximum absolute value of 32766 and then quantized to linear 16-bit PCM.

3.4. Quantization and Rate Conversion Results

Our first set of experiments focused on 8 kHz data in both μ-law and linear PCM format. We compared the performance of IPDF and iVector systems and a simple linear fusion (equal weighted) with multiple configurations. Only SRE 2010 evaluation (not extended) trials were scored. Results are shown in Table 1.

Table 1: Performance of systems on 8 kHz data with linear and μ-law quantization applied to NIST SRE 2010 short-interview train and test, condition 2.

Hyper-params	Eval	IPDF EER/oldDCF	iVector EER/oldDCF	Fuse EER/oldDCF
μ-law	μ-law	6.62/0.303	5.78/0.295	5.37/0.250
μ-law	linear	5.12/0.249	5.08/0.278	4.33/0.215
linear	linear	4.45/0.217	4.12/0.224	3.54/0.180

In the table, hyperparameters indicates the quantization type of the data used for training the UBM, subspaces, transforms, and cohorts. Quantization for the NIST SRE Eval10 data (both enroll and verify) is given in the Eval column. Results show that a 30% reduction in error rate can be achieved by switching to linear quantization.

3.5. Sampling Rate Results

We applied the same methodology used for building the 8 kHz linear system to the 16 kHz linear system. Results are shown in Table 2. In the table, the rate column indicates the sampling rate for the hyperparameter training, cohorts, and NIST SRE 2010 enroll/verify. All results use linear PCM encoding. In the table, we have broken out results by gender.

The table shows several interesting results. First, there is a big absolute male/female performance gap for systems at 8 kHz. This absolute gap is reduced when the sampling rate is increased to 16 kHz. Second, there is a substantial reduction in error rates when using an all 16 kHz system as compared to an 8 kHz system.

Table 2: Performance of systems on 8 kHz and 16 kHz data with linear PCM encoding applied to NIST SRE 2010 short-interview train and test, condition 2.

Rate	Sex	IPDF EER/oldDCF	iVector EER/oldDCF	Fuse EER/oldDCF
8 kHz	M	1.87/0.106	2.81/0.155	1.69/0.088
8 kHz	F	6.14/0.302	5.09/0.272	5.02/0.250
8 kHz	All	4.45/0.217	4.12/0.224	3.54/0.180
16 kHz	M	1.03/0.060	1.31/0.068	0.66/0.044
16 kHz	F	3.00/0.150	2.02/0.116	1.87/0.116
16 kHz	All	2.12/0.110	1.62/0.096	1.33/0.085

3.5.1. Analysis

Somewhat surprisingly, both quantization and sampling rates substantially impact system performance. More future work is needed to understand the exact mechanism for degradation of performance. For instance, for μ-law quantization, the systems may be impacted if no gain normalization is applied before quantization. Alternatively, the quantization or resampling methods may be the degrading factor.

The improvement of error rates at higher sampling rates is quite compelling. The result clearly shows that current systems can (and probably should) take advantage of more bandwidth if available in an application; i.e., substantial performance improvements are possible. In addition, the dramatic drop in error rate for female speakers is worth investigation, since a male/female gap at telephone bandwidth has been problematic in many systems.

Another interesting area of exploration is to understand how to build a better 16 kHz system. The availability of more high-rate data would be a key enabler to this process. Also, dealing with additional non-speech artifacts (e.g., breath noise is more apparent) could improve performance. Finally, feature extraction tuned for 16 kHz for MFCCs might be of interest. The resulting optimized 16 kHz would provide a good benchmark for 8 kHz systems.

4. Effects of Speech Enhancement

Enhancement has been a feature of the MIT LL system since the first microphone data appeared in the NIST evaluation in 2005 [10]. The use of enhancement has been motivated by the presence of tones and wide-band noise in the NIST data. In this section, we review the techniques for enhancement and recent modifications to the algorithms to increase efficiency. The combination of tone-suppression and wide-band noise reduction systems is denoted as the noise preprocessing (NPP) system. We explore the role of enhancement in SAD and feature processing. Experiments demonstrate where gains in performance are achieved and the efficacy of enhancement.

4.1. 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. 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:

1. The audio input is windowed using an 8-second long Hamming window, and its Fourier transform is computed.
2. The magnitude spectrum is whitened by subtracting a smoothed version of the original.
3. Tones are detected by applying a threshold to the whitened spectrum and at each tone a Gaussian-shaped template with a 4-Hz bandwidth is subtracted from the magnitude. The tone reduction threshold is fixed relative to the mean whitened spectrum.
4. The resulting spectrum is inverted and a complete speech signal estimate is obtained through an overlap-and-add reconstruction with neighboring 8-second segments.

Computational time for the algorithm is about 0.01 times real time. Figure 1 shows an example of the various algorithmic steps in detection. The red spectral curve provides the spectral average that is divided out of the composite measured spectrum. After this spectral normalization, the interfering tones stand out from the resulting uniform background and a threshold is set for tone detection relative to the mean of the whitened spectrum.

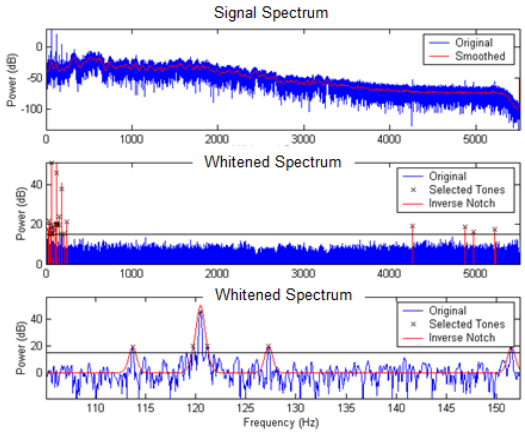

Figure 1: Illustration of steps in tone suppression: (top panel) Smooth spectral estimate superimposed on original spectrum; (middle panel) Whitened magnitude spectrum obtained by removing smooth spectrum; (bottom panel) Detected tones relative to whitened spectrum via a detection threshold.

4.2. 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 property of the approach is an estimate of the speech spectrum, as well as a possibly time-varying background spectrum, required by the Wiener filter, using a measure of spectral change that allows robust and rapid adaptation of the filter to speech and background events. The approach reduces speech distortion in Wiener filtering by making the time constants that control smoothing of the speech and background spectra a time-varying parameter. In particular, time constants are 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 [11, 12]. The approach is consistent with temporally shaping noise to fall within certain regions of least perceptual sensitivity [11].

An important component of the system is a speech activity detector that guides the time constants in speech-spectral smoothing for the adaptive suppression, as well as time constants in smoothing a possibly time-varying background spectral estimate during non-speech regions. The detector is a highly-sensitive, yet robust, multi-band detector. The method works by modeling the per frame energy distribution (in the log domain) of different frequency bands as a weighted sum of two Gaussian distributions. The lower of the two Gaussian means is an estimate of the background noise level, the higher mean is an estimate of the signal level, and the ratio of the two is the SNR. We then normalize the log energy in each band by the mean and variance of the lower Gaussian. This normalization allows the signal in low-energy bands to contribute to a detection statistic, responsive to low-energy regions of unvoiced speech. The detector threshold is set by fitting two Gaussians to the detector statistic and then finding a point between the two Gaussian means where the Gaussians cross. If they do not cross the midpoint between the two Gaussians is used. The detector statistic

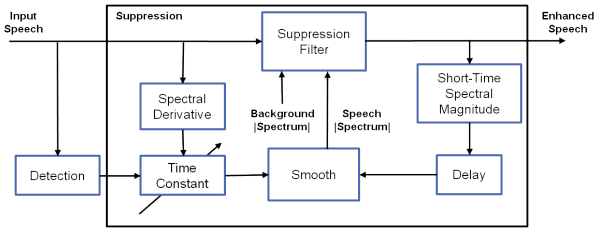

Figure 2: Wiener filter with adaptive estimation of speech and background spectra using a spectral derivative measurement and speech activity detection. Time-varying background spectrum is estimated similarly to the speech spectrum.

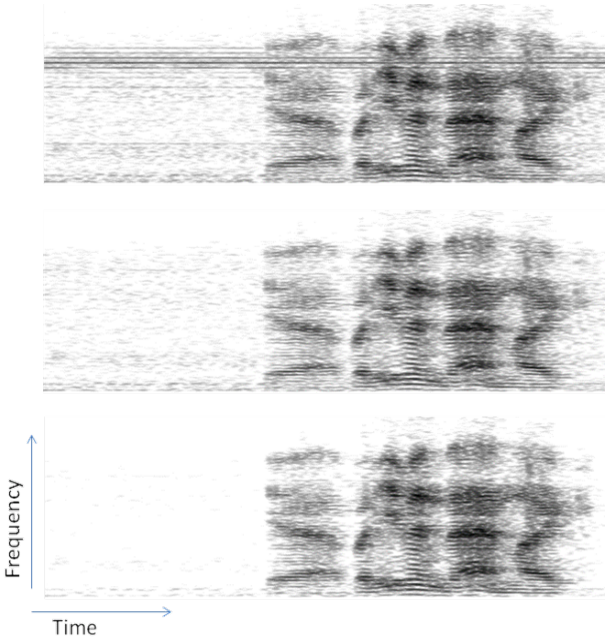

Figure 3: Enhancement of cut from NIST 2006 conversational microphone utterance, pabt.sph (channel B): (top) Original; (middle) Tone suppression; (bottom) Tone suppression and wide-band noise reduction. The duration of the cut above is 2.5 seconds. The frequency range of the spectogram is 4 kHz.

is a sum of the energy across all bands, weighted by the SNR in each band.

The algorithm steps can thus be summarized as:

1. Detect speech or background in each frame using multi-band energy.
2. Estimate the speech signal spectrum by smoothing the enhanced output of the adaptive Wiener filter.
3. Obtain signal change measure, given by a spectral derivative, for controlling smoothing constants in (2).
4. Estimate background spectrum in non-speech regions also with adaptive smoothing using (3).

The entire wideband noise suppression system is shown in Figure 2.

With respect to computation, we have sped up the original version of the algorithm [11, 12] by modifying internal algorithm parameters such as frame rate (10ms from 1ms) and DFT length (256 from 1024) for spectral analysis, and associated smoothing constants, bringing computation to 0.009 times real time. This faster version of the wideband noise reduction algorithm was tested in our current MIT LL SRE system and found to slightly improve performance over prior versions [10].

4.3. Example from the NIST Corpora

Figure 3 shows an example of enhancement of a cut from a NIST utterance. Mid- and high-frequency tones are reduced in the middle panel, while the bottom panel shows the noise suppression from the adaptive Wiener filter.

4.4. Experiments and Analysis

For our initial experiments, we used the experimental setup from Section 3 (8 kHz μ-law). For our second set of experiments, we considered the effect of adding SRE 2008 data to hyperparameter training. We used approximately 13,000 recordings from the 2008 NIST SRE (and followon releases) along with the NIST SRE 2005/2006 data for training UBMs and subspace/Wiener-filter parameters. We used 3000 randomly chosen utterances from NIST SRE 2008 per gender augmented with the NIST 2005/2006 data for z- and t-norm.

Tables 3 and 4 present the results of the noise preprocessing on the front-end system. Table 3 uses only data from microphones in SRE 2005 and 2006 to train hyperparameters (with the same lists as in Section 3). Table 4 uses additional data from SRE 2008 to created a more matched situation for hyperparameter training (the setup of microphones in NIST SRE 2008 and 2010 was similar).

We clearly see that NPP gives the largest gains when used both in SAD and feature processing. However, it is interesting to see that noise preprocessing also gives gains when only used exclusively in SAD or feature generation. For both the IPDF and iVector systems the NPP gave larger gains when applies only to feature processing as compared to being applied only to SAD. This trend indicates the logical conclusion that the most important criterion for better speaker recogniton is an improved clean spectral estimate.

Finally, we note that adding additional 2008 data improves performance for both the IPDF and iVector systems. The results show that even with more matched hyperparameter training, NPP still provides performance improvements.

5. Using Two-Channel Processing for Microphone SAD

5.1. A Statistical Approach to Two-Channel SAD

In this section, we present a solution to speech activity detection for the two-channel microphone interview paradigm used in the NIST SRE 2008 and 2010 speaker recognition evaluations. In this framework, the speaker of interest, referred to as channel A, is recorded using a far-field microphone. For SAD purposes, a near-field microphone channel from the interviewer is provided, which we refer as channel B. The goal of two-channel SAD for this scenario is to extract active speech frames from channel A, while squelching those frames which are corrupted by interviewer leakage.

During previous NIST SRE evaluations, ASR transcripts were provided from a near-field interviewee channel. The use of ASR transcripts, however, does not represent a realistic solution for the two-channel interview framework since it requires oracle knowledge of the close-talking (high SNR) recorded interviewer and interviewee channels. Here, we propose a solution that overcomes this requirement.

We assume a two-state speech activity model with additive background noise, given by

$$\begin{aligned} \mathcal{H}_{A,0}: &\ Y_{A,k} = N_{A,k} \\ \mathcal{H}_{A,1}: &\ Y_{A,k} = X_{A,k} + N_{A,k} \end{aligned} \tag{3}$$

where $Y_{A,k}$, $X_{A,k}$, and $N_{A,k}$ represent DFT coefficients for channel A of observed speech, clean speech, and noise, respectively, and where k denotes frequency channel index. Also, let $\mathbf{Y}_A = \{Y_{A,1}, \ldots, Y_{A,M}\}$, where M is the number of channels used during short-time spectral analysis. A corresponding model with appropriate terms is defined for channel B.

Note that $X_{A,k}$ refers to any speech present in channel A, and may be due to interviewee and/or interviewer. However, due to the relative proximities of the microphones to either speaker, the interviewee speech can generally be expected to appear with a greater amplitude than that of the interviewer. Conversely, in channel B, the interviewer can be expected to appear with a greater amplitude than the interviewee.

As in [13] and [14], we assume real and imaginary DFT components of speech and noise to be independent and normally distributed with variances $\sigma^2_{X,A}(k)$ and $\sigma^2_{N,A}(k)$ for channel A, and $\sigma^2_{X,B}(k)$ and $\sigma^2_{N,B}(k)$ for channel B. We define the *a priori* and *a posteriori* SNRs, respectively, as

$$\gamma_{A,k} = \frac{|Y_{A,k}|^2}{\sigma^2_{N,A}(k)}, \quad \xi_{A,k} = \frac{\sigma^2_{X,A}(k)}{\sigma^2_{N,A}(k)} \tag{4}$$

and

$$\gamma_{B,k} = \frac{|Y_{B,k}|^2}{\sigma^2_{N,B}(k)}, \quad \xi_{B,k} = \frac{\sigma^2_{X,B}(k)}{\sigma^2_{N,B}(k)}. \tag{5}$$

In our implementation, the *a priori* SNR is approximated using the decision-directed approach from [13], and noise estimation is performed according to [15]

To determine the probability of active interviewee speech in channel A which is uncorrupted by interviewer speech, we use as a cost function the joint probability distribution

$$\begin{aligned} \mathcal{L}(\mathbf{Y}_A, \mathbf{Y}_B) &= p(\mathcal{H}_{A,1}, \mathcal{H}_{B,0}|\mathbf{Y}_A, \mathbf{Y}_B) \\ &= p(\mathcal{H}_{A,1}|\mathbf{Y}_A)\, p(\mathcal{H}_{B,0}|\mathbf{Y}_B) \\ &= \left(1 + \Lambda(\mathbf{Y}_A)^{-1}\right)^{-1} (1 + \Lambda(\mathbf{Y}_B))^{-1} \end{aligned} \tag{6}$$

where the likelihood ratio can be derived using (1)-(3) from [14]

$$\begin{aligned} \Lambda(\mathbf{Y}_A) &= \left(\frac{P(\mathcal{H}_{A,1})}{1 - P(\mathcal{H}_{A,1})}\right)^M \prod_{k=1}^{M} \frac{p(Y_{A,k}|\mathcal{H}_{A,1})}{p(Y_{A,k}|\mathcal{H}_{A,0})} \\ &= \left(\frac{P(\mathcal{H}_{A,1})}{1 - P(\mathcal{H}_{A,1})}\right)^M \prod_{k=1}^{M} \frac{1}{1 + \xi_{A,k}} \exp\left(\frac{\gamma_{A,k}\xi_{A,k}}{1 + \xi_{A,k}}\right). \end{aligned}$$

Due to numerical issues, products of probabilities are determined in the log domain.

In our system, the cost function in (6) is generalized as

$$\begin{aligned} \mathcal{L}(\mathbf{Y}_A, \mathbf{Y}_B) &= p(\mathcal{H}_{A,1}|\mathbf{Y}_A)^{\lambda}\, p(\mathcal{H}_{B,0}|\mathbf{Y}_B)^{1-\lambda} \\ &= \left(1 + \Lambda(\mathbf{Y}_A)^{-1}\right)^{-\lambda} (1 + \Lambda(\mathbf{Y}_B))^{\lambda - 1} \end{aligned}$$

Table 3: Performance of systems when comparing noise preprocessing (NPP) effects during feature processing and speech activity detection (SAD) for condition 2 on NIST SRE 2010 short interview train and test data. Hyperparameters and cohorts are trained using NIST 2005 and 2006 data.

SAD Processing	Feature Processing	IPDF EER/oldDCF	iVector EER/oldDCF	Fuse EER/oldDCF
no NPP	no NPP	8.36/0.373	7.37/0.386	7.12/0.323
NPP	no NPP	7.91/0.343	6.82/0.349	6.53/0.293
no NPP	NPP	6.87/0.317	6.53/0.328	5.62/0.270
NPP	NPP	6.62/0.303	5.78/0.295	5.37/0.250

Table 4: Performance of systems when comparing noise preprocessing (NPP) effects during feature processing and speech activity detection (SAD) for condition 2 on NIST SRE 2010 short interview train and test data. Hyperparameters and cohorts are trained using NIST 2005, 2006, and 2008 data.

SAD Processing	Feature Processing	IPDF EER/oldDCF	iVector EER/oldDCF	Fuse EER/oldDCF
no NPP	no NPP	8.32/0.330	5.95/0.274	6.24/0.262
NPP	no NPP	7.03/0.285	5.04/0.244	5.20/0.225
no NPP	NPP	6.28/0.273	4.70/0.237	4.66/0.211
NPP	NPP	5.33/0.244	3.83/0.201	3.91/0.183

Table 5: Performance of Systems when comparing two-channel speech activity detection (SAD) techniques for condition 2 on NIST SRE 2010 short interview train and test data.

Two-Channel SAD	IPDF EER/oldDCF	iVector EER/oldDCF	Fuse EER/oldDCF
ASR	5.33/0.244	3.83/0.201	3.91/0.183
Proposed	4.66/0.200	4.33/0.206	3.70/0.165

where the parameter λ allows control over the relative weight of each channel. This leads to the SAD decision rule

$$\mathcal{L}(\mathbf{Y}_A, \mathbf{Y}_B) \begin{array}{c} (\mathcal{H}_{A,1}, \mathcal{H}_{B,0}) \\ > \\ < \\ \neg(\mathcal{H}_{A,1}, \mathcal{H}_{B,0}) \end{array} \eta.$$

In our system, the parameters λ and η, along with the prior probabilities of active speech, are empirically optimized for speaker recognition performance.

5.2. Experimental Results for Two-Channel Microphone Speech

To assess the effectiveness of the proposed two-channel SAD, we applied it to condition 2 on NIST SRE 2010 short interview train and test data. The expanded hyperparameter training and zt-norm lists using the combined 2005/2006/2008 data set were used as in Section 4.

Results for the proposed SAD along with those using ASR transcripts are provided in Table 5. We see that the new proposed SAD works better than the ASR transcript approach for the IPDF system. No gains were achieved for the iVector system. Fusion results in a modest gain over the baseline system. Overall, we have achieved the goal of creating a SAD system which is independent of unrealistic ASR transcripts and achieves comparable fused performance.

6. Conclusions

We explored multiple issues in the pre-processing of microphone speech data for the NIST SREs–sampling, quantization, enhancement, and SAD. For sampling and quantization, we clearly showed that retaining as much information about the signal as possible resulted in substantial improvements in performance. For enhancement, we showed that improvments occur in both feature and SAD processing. Finally, we demonstrated that the standard "oracle" ASR SAD for the interview data could be replaced by a 2-channel SAD which was not only realistic but outperformed the oracle condition.

Additional work on more detailed qualitative understanding of the results will be pursued. Also, promising improvements from new methods (e.g., SAD) can be incorporated into future systems. Both of these efforts should yield improved baseline systems and point to alternate directions for pre-processing in future NIST evaluations.

7. References

[1] 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.

[2] 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.

[3] W. Campbell and Z. Karam, "Simple and efficient speaker comparison using approximate KL divergence," in *Proceedings of Interspeech*, 2010.

[4] Alan McCree, Doug Sturim, and Doug Reynolds, "A new perspective on GMM subspace compensation based on PPCA and wiener filtering," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[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] Douglas E. Sturim, William M. Campbell, Najim Dehak, Zahi N. Karam, Alan McCree, Douglas A. Reynolds, Fred Richardson, Pedro A. Torres-Carrasquillo, and Stephen Shum, "The MIT LL 2010 speaker recognition evaluation

system: Scalable language-independent speaker recognition," in *Proceedings of ICASSP*, 2011, pp. 5272–5275.

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

[8] 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.

[9] L. Rabiner and R. Crochiere, *Multirate Digital Signal Processing*, Prentice-Hall, 1983.

[10] 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.

[11] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *Proceedings of ICASSP*, 2002, pp. I–257 – I–260.

[12] T. F. Quatieri and R. Baxter, "Noise reduction based on spectral change," in *IEEE Workshop on Appl. Signal Processing to Audio and Acoustics*, New Paltz, NY, 1997.

[13] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[14] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[15] Rainer Martin, "Noise power spectral estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Signal Processing*, vol. 9, no. 5, pp. 504–512, Aug. 2001.