A Noise-Robust System for NIST 2012 Speaker Recognition Evaluation

Luciana Ferrer, Mitchell McLaren, Nicolas Scheffer, Yun Lei, Martin Graciarena, Vikramjit Mitra

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.

{lferrer,mitch,scheffer,yunlei,martin,vmitra}@speech.sri.com

Abstract

The National Institute of Standards and Technology (NIST) 2012 speaker recognition evaluation posed several new challenges including noisy data, varying test-sample length and number of enrollment samples, and a new metric. Target speakers were known during system development and could be used for model training and score normalization. For the evaluation, SRI International (SRI) submitted a system consisting of six subsystems that use different low- and high-level features, some specifically designed for noise robustness, fused at the score and iVector levels. This paper presents SRI's submission along with a careful analysis of the approaches that provided gains for this challenging evaluation including a multiclass voice-activity detection system, the use of noisy data in system training, and the fusion of subsystems using acoustic characterization metadata.
Index Terms: Speaker recognition, noise-robustness, PLDA, iVector

1. Introduction

NIST's 2012 speaker recognition evaluation posed several new challenges: clean and noisy test samples of varying lengths, a varying number of enrollment sessions, and knowledge of the target speakers during development and the permission to use them for system training and score normalization [12]. Further, a new metric was introduced involving two operating points and separate weighting of false alarms for test samples corresponding to a target speaker or an unknown speaker.

SRI's approach to tackle these challenges included: (1) a careful design of a development set matching the evaluation data description as closely as possible, which was used for model training and system tuning and calibration; (2) a multiclass, noise-robust voice-activity-detection (VAD) system with cross-talk removal; (3) the use of metadata aimed at representing the acoustic characteristics found in the enrollment and test samples; (4) a set of six features, some of them specifically developed for noise robustness; (5) the iVector fusion of these feature-specific subsystems with metadata used in the fusion; and (6) a final transformation of the scores to take advantage of knowledge of the target speakers. The system design was simple: all six features were modeled with an identical iVector/probabilistic linear discriminant analysis (PLDA) approach with some small differences in its parameters.

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024 and by Sandia National Laboratories (#DE-AC04-94AL85000). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch or Sandia National Laboratories. "A" (Approved for Public Release, Distribution Unlimited)

This paper presents an analysis of the explored approaches and shows which of these approaches gave significant gains for the evaluation data.

2. System Description

This section describes the development set; the VAD system; the individual subsystems; and the fusion strategy used to build the evaluation system.

2.1. Development Set

This year NIST released the list of target speakers more than a month in advance of the evaluation. Target speakers were most of the speakers available in the 2008 and 2010 evaluation data, including a total of 1818 speakers with a large variance in the number of available sessions. We chose to use these same target speakers as enrollment speakers in our devset, holding out 168 speakers to be used as "unknown test speakers" (that is, speakers for which no target model is trained). Additionally, 200 speakers from the 2004 through 2006 evaluation data were chosen as unknown test speakers. For each target speaker, up to six sessions were kept for testing and the rest were used for speaker enrollment.

No "summed" data was used for enrollment, testing, or system training. SRE10 microphone data of 16kHz was used with all other data sampled at 8kHz. Interview data was used only when both the interviewee and interviewer recordings were available and of the same length. This facilitated the use of cross-talk removal by VAD as described later.

The evaluation set had around 10k male and 15k female segments for model enrollment, and 8k male and 10k female test segments. Test segments were cut to contain active speech of random durations between 15 and 200 seconds. Up to five cuts per segment were produced.

In addition to the original segments of the dataset, noise was added to each segment to produce a noisy version of each. The noise conditions were created from the clean data set through artificial degradation at different signal-to-noise ratio (SNR) levels, using different samples of heating, ventilation, and air conditioning (HVAC) noise taken from freely available online resources and speech spectrum noise formed by summing hundreds of telephone conversation sides for each noise sample. The train and speaker enrollment portion of the development set was duplicated and degraded to around a 6 or 15 dB signal to noise ratio (SNR), randomly choosing one noise type, using a version of the publicly available tool FaNT modified to account for the C-weighting specification. In contrast, test segments were duplicated twice by renoising at both 15dB and 6dB SNR. Different noise signals were used for training, enrollment and testing.

A large set of trials was developed under a number of constraints based on the evaluation plan provided by NIST. These

constraints include same-gender trials; English-only and normal vocal effort test segments; and a preference for different-number phone-call trials (trials were discarded if both numbers were known and different). Trials were created by pairing every target model against all test samples, creating a large number of impostor trials and the largest possible number of target samples under the aforementioned constraints. Trials involving two signals recorded during the same session (i.e., two different microphone recordings of the same interview) were excluded.

The total number of trials obtained was around 14,000 target and 20 million impostor male trials, and 19,000 target and 38 million impostor female trials. Around half of the impostor trials were from unknown speakers.

Training data were extracted from Fisher 1 and 2, Switchboard phase 2 and 3, and Switchboard cellphone phase 1 and 2, along with all available Mixer speakers except the unknown test speakers (target speakers are included in the training data). A total of 11,971 speakers were used from the Fisher data; 1,950 from Switchboard data; and 2,937 from Mixer data, for a total of 38k male and 51k female segments.

2.2. Voice-Activity Detection

We used a multi-class Gaussian mixture model (GMM)-based VAD system including cross-talk removal for interview segments. The multi-class VAD involved first training speech/non-speech GMMs for both clean and noisy classes using mel-frequency cepstral coefficients (MFCCs) of 10 dimensions plus energy and deltas, double deltas, and triple deltas. GMMs were trained using data from the training part of the development data set with bootstrapped annotations from our previous VAD approach involving a speech/non-speech hidden Markov model (HMM) decoder and various duration constraints. All audio used for VAD training and evaluation was first Wiener filtered. VAD setup was tuned to optimize speaker recognition performance on the development set described above.

Frame-level likelihoods were obtained from each of the four GMMs and the log likelihood ratio of the speech versus non-speech models was found. Finally, a median filter of 41 frames was used to smooth the obtained scores. Frames with a smoothed score above 0.1 were declared speech.

For the interview recordings, we used a more complex algorithm to suppress cross-talk due to interviewer speech. The algorithm is the following: (1) segment the interviewee channel as per the method described above; (2) segment the interviewer channel with a stricter threshold of 2.1; and (3) remove segments found in (2) from segments found in (1). If more than 50% of the speech from (1) was removed, the threshold in step (2) was revised to limit the cross-talk removal to 50%.

2.3. Subsystem Description

Six different subsystems are included in the system, corresponding to different feature sets extracted from the speech. An iVector/PLDA approach is used for modeling all features.

2.3.1. Features

The following is a description of the six sets of features used in the subsystems. MDMC, PNCC, and MHEC are features specifically designed to be robust under noisy conditions.

MFCC (Low-Level) These features use a 200-3300 Hz bandwidth front end consisting of 24 Mel filters to compute 19 cepstral coefficients plus energy and their delta, and double delta coefficients over windows of 20ms shifted by 10ms, producing a 60-dimensional feature vector.

PLP (Low-Level) The perceptual linear prediction (PLP) features use a 100-3760 Hz bandwidth front end consisting of 24 Mel filters to compute 12 cepstral coefficients plus energy and their delta, double delta, and triple delta coefficients, producing a 52 dimensional feature vector.

MDMC (Low-level) Medium duration modulation cepstral (MDMC) features extract cepstra from amplitude modulation spectrum by using a modified version of the algorithm described in [1]. Audio was sampled every 10 ms using a 51.2 ms Hamming window and analyzed by a 30 channel gammatone filter bank spaced equally from 250 Hz to 3750 Hz in the ERB scale. The AM power signal from each subband was power normalized using 1/15th root, followed by DCT, after which only the first 20 coefficients were retained with deltas and double deltas appended.

PNCC (Low-level) Power-normalized cepstral coefficient (PNCC) features use a frequency domain 30-channel gammatone filter bank that analyzes the speech signal [2] at 10 ms with a 25.6 ms Hamming window, where the filterbank cutoff frequencies were at 133Hz and 4000Hz. Short-term spectral powers were estimated by integrating the squared gammatone responses, and the resultant was compressed using 1/15th root, followed by DCT. The first 20 DCT coefficients were retained with deltas and double deltas appended.

MHEC (Low-level) Mean Hilbert envelope coefficient (MHEC) features [3] utilize a 24-channel gammatone filter bank with cutoff frequencies at 300 Hz and 3400 Hz, where filter bank energies were computed from the temporal envelope of the squared magnitude of the analytical signal obtained through the Hilbert transform. The estimated temporal envelope is low-pass filtered with a 20 Hz cutoff frequency, which was then analyzed using a 25 ms hamming window with a 10 ms frame rate. Log compression was performed on the resulting followed by DCT to generate 20 cepstral features. Deltas and double deltas were then appended.

PROS (High-level) Prosodic features are extracted from overlapping uniform regions of a length of 20 frames shifted with respect to each other by 5 frames. The feature vector is composed of the coefficients of the Legendre polynomial approximation of order 5 of the pitch and energy signals over the region [4]. Pitch and energy signals are obtained using the get_f0 code from the Snack [5] toolkit. The waveforms are preprocessed with a bandpass filter (200 Hz to 3300 Hz).

2.3.2. Modeling

All subsystems included in our submission use the iVector/PLDA framework for modeling [6, 7]. The iVectors are transformed using linear discriminant analysis (LDA) and log-likelihood ratios for each trial are estimated using probabilistic linear discriminant analysis (PLDA). All models were gender-dependent.

Background models were trained using only 8k samples from Mixer data, while the iVector extractor was trained using every training session available in the training set. The LDA and PLDA models were trained using all training data corresponding to speakers who participated in at least six sessions and any speaker data used in enrollment. Noisy data was used in combination with the clean segments only in the LDA/PLDA stage [8] and for enrollment.

With the exception of the PROS system, features obtained after VAD were mean and variance normalized over the utterance. For the five low-level systems, the feature vectors were modeled by a 2048-component, gender-dependent GMM with diagonal covariances. The dimension of the iVectors for

these systems was 600, further reduced to 150 by LDA. For the high-level PROS system, the feature vectors were modeled by a 1024-component gender-dependent GMM with diagonal covariances and the dimension of the iVectors was 200, further reduced to 100 by LDA. Mean and length normalization were performed on the iVectors after LDA.

2.4. System Fusion and Compound Score Transformation

Two system combination and calibration strategies are used: (1) iVector fusion and (2) score-level fusion or calibration using metadata. Fused scores are further transformed to account for the given prior probability of the test sample coming from a known target speaker.

iVector Fusion: The iVectors produced by individual systems (after LDA) were concatenated, and the final vector was further reduced to 150 dimensions via LDA. The fused iVectors were modeled and scored using PLDA.

Score-level Fusion: For score-level fusion, the fused scores were a linear combination of scores from individual systems where weights and bias are learned using linear logistic regression. A single set of fusion parameters was learned on all development data, both clean and noisy. This procedure is also used for calibration of individual systems and iVector fusions.

Acoustic Characterization Metadata: Given that the NIST SRE evaluation data was designed to contain many different types of variabilities, with only a few of them available as labels, we used our 'universal audio characterization' approach [9] to generate metadata for the fusion. The system was trained to predict the acoustic characteristics available in the training data using the MFCC iVectors. To this end, training signals were grouped into six classes: clean/low SNR/high SNR, for telephone data and microphone data. A Gaussian model was trained for each class with covariances tied across classes. Given an acoustic sample, this system produced a six-dimensional vector of posterior probabilities for each of the six classes. A single metadata vector was obtained for each speaker model by averaging the vectors from enrollment segments. During fusion, the verification scores were obtained as a linear combination of scores from the individual systems plus a value obtained from evaluating the bilinear form $q_1^T W q_2$, where W is a symmetric matrix learned during training and q_1 and q_2 are the metadata vectors corresponding to the enrollment segments and the test segment [10].

Compound Scores: The scores resulting from fusion were further transformed to account for the probability of test segments coming from known target speakers. This probability is 0.5 for the core and extended test conditions. This was done using Bayes' rule to transform the raw likelihood ratios output by the system into posterior probabilities using the prior probabilities for the target speakers (assumed to be uniform across speakers) and an unknown target class. These posteriors were finally converted back into likelihood ratios. This procedure was proposed by Niko Brummer in [11].

3. Results and Analysis

We show results on the SRE 2012 evaluation conditions 1 through 5 [12] in which test samples are restricted to: interview speech (C1); telephone speech (C2); interview speech with added noise (C3); telephone speech with added noise (C4); telephone speech collected under noisy conditions (C5).

All results shown in this paper correspond to (1) pooled gender trials; (2) the core training condition in which all available data for each target speaker is used for enrollment; (3) calibrated scores with parameters learned by linear logistic regression on the development set trials; (4) the extended test condition; and (5) compound scores as described in Section 2.4. The C primary metric is used for all results. This metric (described in detail in [12]) is an average of two costs given by a weighted sum of miss and false-alarm error probabilities with the thresholds given by the theoretically optimal thresholds assuming the scores are proper likelihood ratios. Further, the false-alarm errors are weighted differently depending on whether the test sample comes from a target speaker or not.

Note that NIST advised participants not to compare performance across conditions but only within them. For example, C3 is significantly easier than C1 even though C1 is clean and C3 is noisy, because C3 involves only tests of longer durations, while C1 contains a mix of durations.

3.1. Effect of Voice-Activity Detection

The left plot in Figure 1 shows a comparison of the results on the MFCC system when using the described VAD algorithm with different sets of models from which the likelihood ratio of speech versus non-speech are obtained: (1) one GMM for speech and one for non-speech both trained only on clean data (this VAD is called *clean* in the figure); (2) one GMM for speech and one for non-speech both trained on clean and noisy data (*clean+noi*); (3) two GMMs for speech and two for non-speech, trained separately on clean and noisy data (*clean&noi*); and (4) approach (3) without cross-talk removal (*clean&noi noxtalk*). We see that the third approach provides the most robust solution.

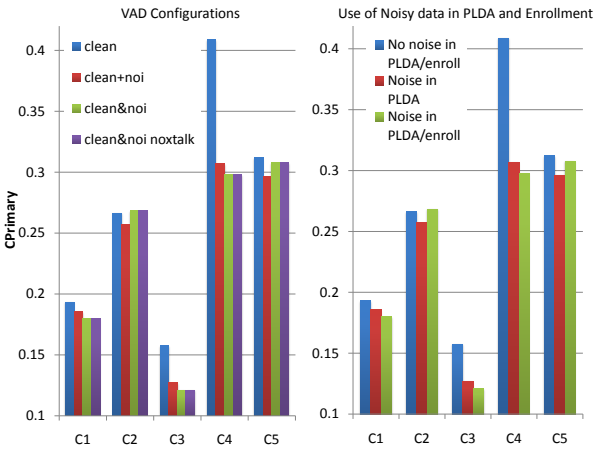

Figure 1: Use of noisy data for system training and enrollment for the MFCC system. Left: Comparison of performance using GMMs trained with different data for VAD (noise in PLDA and enrollment is used for these experiments). Right: Comparison of performance when adding noisy data in PLDA training and enrollment (clean&noi VAD is used for these experiments).

3.2. Effect of Data Used for PLDA and Enrollment

The 2012 SRE was the first time that a variable number of enrollment samples was available for the target speakers within a single evaluation condition. Under these conditions, the current PLDA approach does not behave well. The reasons for this are yet to be discovered, but the current solution is to simply take the average of the enrollment iVectors and then use standard PLDA to compute a score between this average iVector and the test iVector. In our experiments, this approach leads to signif-

icant gains for the low-level systems and the score-level and iVector fusions that range from 25% to 50% on all evaluation conditions, except C3 where no consistent gains are observed. The PROS system does not benefit from averaging enrollment iVectors. We submitted three systems to the evaluation, two of them using separate enrollment iVectors during PLDA scoring and one using the average iVector. In the rest of this paper, we only show results using the latter approach.

Three of the five common conditions in the evaluation contained noisy data. Our development set included renoised data with characteristics similar to those in the evaluation test data. We explored the use of this data during PLDA training and as additional enrollment data. The right plot in Figure 1 shows three sets of results on the MFCC system: (1) no renoised data in PLDA or enrollment, (2) renoised data in PLDA only, and (3) renoised data in PLDA and enrollment. The figure shows gains in the noisy conditions of up to 25% from adding noisy data in PLDA training with no losses on the clean data. Adding noisy data in enrollment does not lead to consistent gains. On the other hand, gains from using noise in enrollment were consistent and large for the system that uses separate iVectors for enrollment (not shown here). Based on those results, we decided to use noise in enrollment for all evaluation systems. Results in the next section use noisy data in both PLDA and enrollment.

3.3. Subsystem and Fusion Results

Figure 2 shows the results for the individual subsystems. The figure shows that the PNCC system is the best system overall, always better than the more standard MFCC system.

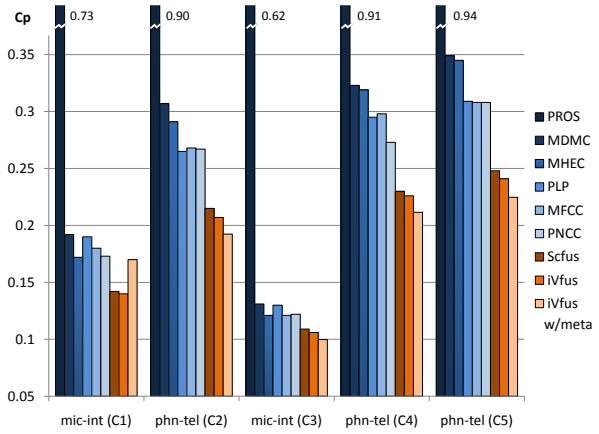

Figure 2: Performance of individual systems and different system fusion techniques. PROS performance is indicated on the bars since showing it to scale would obscure the differences between the other systems.

Figure 2 also shows a comparison of fusion results: (1) the score-level fusion of the six individual systems (*Scfus*); (2) the iVector fusion of PLP, PNCC, MFCC and PROS systems calibrated using logistic regression as for all score-level fusions (*iVfus*); and (3) the fusion in (2) but with the addition of the acoustic characterization metadata during fusion (*iVfus w/meta*). The selection of systems used in 1 and 2 was based on an exhaustive search on the development set. We can see that the iVector fusion is always better than the score-level fusion. Finally, the use of metadata during fusion gives significant gains in all conditions except C1. This was not the case in our development set, where we saw gains of approximately 10% on the condition corresponding to C1. This might point to some difference in the nature of the interview data in the evaluation versus the development data that warrants further study.

The system we submitted to the evaluation was a score-level fusion of all six individual systems plus the iVector fusion, calibrated using metadata. The addition of the individual systems to the iVector fusion does not bring any consistent gains in the evaluation conditions (the gain on the development set was only marginal). We do not show these results in the figure, to reduce clutter.

All results in this paper correspond to compound scores as explained in Section 2.4. The gain obtained on the average Cprimary from the use of this transform on the *iVfus w/meta* system is 15%, being from 11 to 18% on the individual conditions.

An interesting question, given the variety of features available for fusion, is how much is the system gaining from each feature. This is a hard question to answer since, for each number of systems being fused, several combinations give similar performance. Table 1 shows, for n between 1 and 4, the n-way iVector fusions (calibrated without metadata) for which the average Cprimary over the five evaluation conditions is within 2% relative of the top n-way fusion. The five-way and six-way fusions are not better than the four-way fusions and, hence, are not shown in this table. Interestingly, a pattern arises where most n-way fusions are formed by some top (n-1)-way fusion plus one additional system. Both the PLP and the PROS systems are necessary to reach the best performance of 0.183. These are the systems that provide the most new information to the fusion once two low-level systems are already present in the mix.

Table 1: Top n-way fusions along with the best average Cprimary for each n (in parenthesis). The * indicates "the (n-1)-way fusion in the same line".

1-way	2-way	3-way	4-way
(0.227)	(0.201)	(0.189)	(0.183)
PNCC	*+MFCC	*+PROS	* + PLP
	PLP+MDMC	*+PROS	* + MFCC
	PLP+PNCC	*+PROS	* + MDMC
	PLP+MHEC	*+PROS	* + MFCC
		PLP+MFCC+PROS	
		PLP+MFCC+MDMC	

4. Conclusions

We present a description of the system submitted to the 2013 NIST speaker recognition evaluation by SRI International. This system was among the top performers in the evaluation. The system includes several aspects that make it noise-robust. A multi-class speech activity detection system trained with clean and noisy data and the use of noisy data in PLDA result in gains on noisy conditions of up to 20% and 25%, respectively. The fusion of several systems based on low- and high-level features improves performance on both clean and noisy data between 15 and 20% relative to the best individual subsystem — a system based on power-normalized cepstral coefficients. The use of metadata during fusion describing the acoustic characteristics of the enrollment and test data gives additional gains in noisy conditions.

5. References

[1] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Mar. 2012.

[2] C. Kim and R. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Mar. 2012.

[3] S. Sadjadi and J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, May 2011.

[4] N. Dehak, P. Dumouchel, and P. Kenny, "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, Sep. 2007.

[5] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Beijing: China Military Friendship Publish, Oct. 2000.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, may 2011.

[7] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proceedings of the Speaker and Language Recognition Workshop, Odyssey 2010*, Brno, Czech Republic, Jun. 2010, keynote presentation.

[8] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise robust speaker recognition using probabilistic linear discriminant analysis," in *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Mar. 2012.

[9] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Proceedings of the Speaker and Language Recognition Workshop, Odyssey 2010*, Brno, Czech Republic, Jun. 2010.

[10] N. Brummer, L. Burget, P. Kenny, P. Matejka, E. de Villiers, M. Karafiat, M. Kockmann, O. Glembek, O. Plchot, D. Baum, and M. Senoussauoi, "ABC system description for NIST SRE 2010," in *Proceedings of NIST 2010 Speaker Recognition Evaluation*. National Institute of Standards and Technology, 2010, pp. 1–20.

[11] N. Brummer, "LLR transformation for SRE'12." [Online]. Available: http://sites.google.com/site/bosaristoolkit/sre12/llrTrans.pdf

[12] "NIST SRE12 evaluation plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf.