IMPROVING SPEAKER IDENTIFICATION ROBUSTNESS TO HIGHLY CHANNEL-DEGRADED SPEECH THROUGH MULTIPLE SYSTEM FUSION

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

Speech Technology and Research Laboratory, SRI International, California, USA
{mitch,scheffer,martin,lferrer,yunlei}@speech.sri.com

ABSTRACT

This article describes our submission to the speaker identification (SID) evaluation for the first phase of the DARPA Robust Audio and Transcription of Speech (RATS) program. The evaluation focuses on speech data heavily degraded by channel effects. We show here how we designed a robust system using multiple streams of noise-robust features that were combined at a later stage in an i-vector framework. For all channels of interest, our combination strategy presents up to a 41% relative improvement in miss rate at a 4% false alarm rate with respect to the best-performing single-stream system.

Index Terms— i-vector, speaker verification, degraded speech

1. INTRODUCTION

The DARPA RATS program aims at designing robust processing methods of speech acquired from highly degraded transmission channels. The four tracks pursued in RATS are speech activity detection, keyword spotting, language identification, and speaker identification — the last of which is the focus of this paper. Audio recordings are severely degraded when telephone conversations are re-transmitted over eight different military transmitter/receiver combinations [1].

The SCENIC (Speech Content Extraction from Noisy Information Channels) team is composed of speech laboratory teams from five institutions: SRI International, International Computer Science Institute, University of Texas Dallas, Carnegie Mellon University, and University of California at Los Angeles. Each team focuses on robust feature extraction and speech activity detection in the context of RATS degraded data. This widespread focus provided considerable strength to the SCENIC SID submission through the complementary nature of novel features and SAD algorithms from each of the team members.

Section 2 of this article describes the five features contributing to the SCENIC team submission. Section 3 outlines two SAD approaches used in the system. The process of combining multiple feature and input streams into a single score is given in Section 4, along with the specifics of score calibration. Section 5 provides the experimental protocol, followed by the presentation of results and analysis of the compounded system in Section 6.

2. ROBUST FEATURE EXTRACTION

This section describes the five features used in the SCENIC submission. These features, selected from a pool of ten through a process of cross-validation of the development set (see Section 5), are as follows:

- Perceptual linear prediction (PLP) features are the standard features used in speech recognition.
- Medium duration modulation cepstrum (MDMC) features extract modulation cepstrum-based information by estimating the amplitude of the modulation. More details can be found in [2].
- Power-normalized cepstral coefficients (PNCC) features use a power law to design the filter bank as well as a power-based normalization instead of a logarithm. More details can be found in [3].
- Mean Hilbert envelope coefficients (MHEC) features [4] utilize a gammatone filter bank instead of the Mel filter bank, and the filter bank energy is computed from the temporal envelope of squared magnitude of the analytical signal obtained using the Hilbert transform. More details can be found in [4].
- Sub-band autocorrelation classification (SACC) [5] provided a pitch estimate from an estimator that is trained using a multilayer perceptron. Used in conjunction with prosodic-based i-vector modeling, the pitch

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. 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 & Property Management Division, Southwest Branch. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. “A” (Approved for Public Release, Distribution Unlimited)

estimates are converted in to a set of features suitable for the i-vector framework [6]. These features are referred to as PROSACC in this article. More details on SACC pitch estimation can be found in [5].

3. SPEECH ACTIVITY DETECTION

Two speech activity detection (SAD) approaches were used in the SCENIC submission: hidden Markov model (HMM) and Gaussian mixture model (GMM) approaches. Instead of combining the SAD outputs in an effort to obtain a more reliable set of speech labels, both SAD approaches were applied independently to each of the five features, resulting in ten different systems for use in the subsequent i-vector and score-level fusion process. These SAD approaches were applied to audio recordings of more than 10s. For audio recordings of less than 10s, an energy-based SAD was used in which the frames in the lowest 10th percentile of the energy distribution were dropped.

3.1. HMM SAD

The SCENIC team developed a robust speech detector as part of the SAD track of the RATS project. Referred to as SCENIC SAD in this article, this SAD consists of a feature combination front end from four acoustic features: standard PLP acoustic feature; a GABOR spectrogram long-range representation post-processed by a multilayer perceptron; a voicing estimator which is a PCA-based combination of four basic voicing features; and a spectral flux estimator and a multiband voicing estimator. The backend of this SAD includes a HMM decoder from speech and background HMM models.

The HMM SAD was developed in the context of speech recognition and, subsequently, keyword spotting for the RATS program. The system is based on the modeling of multiple speech models with a decoding backend similar to what one would use in speech recognition. Consequently, low speech energy or pause frames needed to be excluded from the feature stream in order to benefit SID performance.

3.2. GMM SAD

An alternative SAD system was developed that uses a much simpler strategy in that the speech detection is based on the log-likelihood ratio output of two GMMs, one for speech and one for non-speech. These two components were trained using the SID development set, and annotations were provided as part of the RATS data distribution. Speech was detected in an audio stream by first calculating the likelihood ratio between the speech and non-speech models. A median filter of length 31 frames was then applied to smooth the detection output.

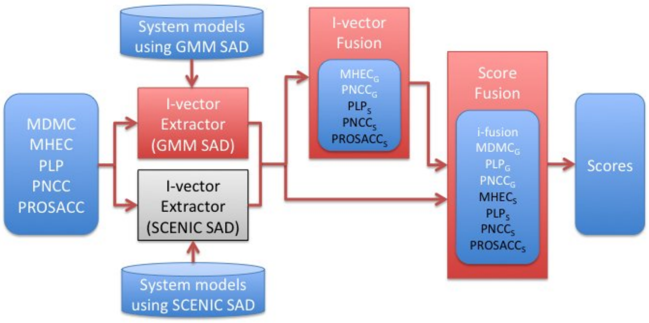

Fig. 1. SCENIC SID system involving five features, two speech activity detectors, i-vector fusion of five feature streams and score fusion of seven feature streams, along with i-vector fused scores.

4. SPEAKER RECOGNITION SYSTEM

4.1. Single-stream System

Each stream of features for both SAD outputs was processed in the same fashion. We used a standard i-vector / probabilistic linear discriminant analysis (PLDA) framework as our speaker recognition system [7, 8]. I-vectors were extracted for each feature+SAD combination, resulting in 10 i-vector streams for possible selection in the fusion process. The SCENIC team employed two styles of fusion: i-vector fusion and score-level fusion. Figure 1 illustrates the data flow through the system and how the different SAD and fusion algorithms and are incorporated.

4.2. I-vector Fusion

I-vector fusion consists of concatenating each i-vector from each stream into a single vector before employing the PLDA backend. The i-vector dimensions are first reduced using LDA, and only after concatenation does a second dimensionality reduction shrink the total dimension to 200. Five out of ten systems were selected for the i-vector fusion process. This selection was based on maximizing SID performance through cross-validation of the development set. The systems selected for i-vector fusion were MHEC_G, PNCC_G, PLP_S, PNCC_S and $\mathrm{PROSACC}_\mathrm{S}$, where the subscript letters G and S indicate GMM SAD or SCENIC SAD, respectively. It is interesting to note that PNCC from both SAD configurations was selected.

4.3. Score Fusion and Calibration

Single system i-vector streams were fused at the score level along with the scores from the i-vector fused system. The selection of streams to be included in the score-level fusion were selected independently of the previous i-vector fusion and included MDMC_G, PLP_G, PNCC_G, MHEC_S, PLP_S, PNCC_S and $\mathrm{PROSACC}_\mathrm{S}$.

	Test (seconds)			
Enroll (seconds)	3	10	30	120
3	X	X	X	
10	X	X	X	
30			X	
120				X

Table 1. *The eight trial conditions evaluated.*

Fusion of systems at the score level was performed using logistic regression and a binary cross entropy objective [9]. This is the standard fusion approach in speaker recognition. The selection experiments were carried out using cross-validation sets where fusion parameters were trained on one and applied to the other.

5. EXPERIMENTAL PROTOCOL AND SYSTEM CONFIGURATION

The RATS SID task was defined as a speaker verification task where each speaker model was trained using six different sessions. A trial was designed using one speaker model and one test session. The transmission channels of the six different sessions were picked randomly to have speaker models trained on multiple transmission types. Some of the trials were thus performed on channels seen in enrollment, while others were not.

The primary metric was defined as the percentage of misses at the 10% false alarm rate. Multiple duration configurations for the enrollment and tests were of interest in this evaluation. A total of eight conditions were formed with durations of 3, 10, 30 and 120 seconds for the input files (Table 1). Note that an enrollment duration of 10 seconds denotes that speaker models were trained using six sessions, each with 10 seconds of nominal speech activity.

Regarding data used for our development, data from LDC releases LDC2012E49, LDC2012E63 and LDC2012E69 under the RATS program were divided by the SCENIC team into training and development sets. Table 2 presents the distribution of languages across the datasets. A major factor that influenced this distribution was that speakers of the dev set were required to have at least seven original (not re-transmitted) recordings. For PLDA training, segments of the train set had 10-, 30- and 120-s cuts taken from each segment in the train set to better represent the i-vector distribution of evaluation data.

For the i-vector framework used by all feature streams, we used universal background models (UBMs) with 2048 diagonal covariance Gaussian components trained in a gender-independent fashion. The PROSACC systems used 1024-component UBMs. The i-vector dimensions of 400 were further reduced to 200 dimensions by LDA (100 for PROSACC), followed by length normalization and PLDA.

Language	Train Set	Dev Set
Levantine	6056	1532
Farsi	1086	359
Dari	18	270
Pushto	3291	2630
Urdu	0	494

Table 2. *Language distribution of recordings in Train and Dev sets.*

	I-vector Fusion		Score+IV Fusion	
Eval. Cond.	m4FA	EER	m4FA	EER
3-3	62.60	21.72%	57.32	20.04%
3-10	39.15	14.61%	32.63	12.99%
3-30	25.89	10.96%	20.39	9.50%
10-3	46.82	17.64%	43.52	16.95%
10-10	21.45	10.01%	18.72	9.37%
10-30	9.74	6.30%	8.15	5.80%
30-30	6.50	5.13%	5.83	4.87%
120-120	2.04	2.82%	1.94	2.76%

Table 3. *Performance of SCENIC SID system across different enrollment and test conditions based on the development set.*

6. RESULTS

Table 4 presents the performance of individual feature streams for matched enroll and test durations. Both GMM and SCENIC SAD results are provided. Verification performance is reported in terms of miss rate at 4% false alarm (m4FA) and equal error rate (EER). Compared to other features, MDMC and PNCC were consistently the best performers across all duration conditions, illustrating their robustness to degraded conditions. PNCC was in particular found to be a major contributor in the SCENIC system, with both SAD alternatives being utilized in both fusion stages. Interestingly, the prosodic system was able to find speaker-discriminative information even in limited audio. Despite the generally lower performance from PROSACC, this system was highly complementary in the fusion process. In contrast to other features, PROSACC used in conjunction with SCENIC SAD outperformed the alternative GMM SAD. It is believed that SCENIC SAD provided a more continuous transition between high-energy speech frames facilitating pitch estimation, as opposed to the more rigid detection strategy of the GMM SAD.

Table 3 provides the results of i-vector fusion and the subsequent score fusion involving the i-vector fused results for the development dataset. While i-vector fusion provides significant gains over any single system, its combination with our systems at the score level brings even further improvements. This was particularly the case for shorter durations. Results from the final fused system (Score+IV Fusion) in Table 3 provided a considerable relative improvement of up to

		GMM SAD		SCENIC SAD	
Eval. condition	**Feature**	**m4FA**	**EER**	**m4FA**	**EER**
3-3	MDMC	68.25	24.18%	68.46	24.34%
	MHEC	70.33	24.93%	71.57	25.40%
	PNCC	68.90	24.43%	69.04	24.82%
	PLP	78.17	28.68%	78.21	28.66%
	PROSACC	78.50	27.80%	77.77	27.63%
10-10	MDMC	28.26	12.10%	30.32	12.89%
	MHEC	30.26	12.41%	32.33	13.38%
	PNCC	28.33	12.14%	30.19	12.93%
	PLP	33.48	13.53%	35.01	14.35%
	PROSACC	59.05	20.13%	58.38	19.90%
30-30	MDMC	9.46	6.36%	10.65	6.84%
	MHEC	10.61	6.82%	11.64	7.10%
	PNCC	9.73	6.38%	10.60	6.78%
	PLP	13.36	7.71%	15.08	8.23%
	PROSACC	38.88	14.27%	37.37	13.92%
120-120	MDMC	3.30	3.60%	3.97	3.98%
	MHEC	3.66	3.82%	4.25	4.15%
	PNCC	3.48	3.71%	3.73	3.85%
	PLP	4.81	4.39%	5.49	4.74%
	PROSACC	20.47	9.54%	17.72	8.76%

Table 4. *Comparison of the individual features and SAD labels across matched enrollment and test durations*

41% in m4FA over any individual feature in Table 4, thus demonstrating the strength of the fusion approach employed in the SCENIC SID submission.

7. CONCLUSIONS

The RATS program presents a highly challenging task for speaker recognition where speech has been heavily degraded by transmission effects. The SCENIC approach is to bring robustness to these degradations to all components of the pipeline. We showed in this paper how this approach can be successful as the final systems use multiple speech detectors, multiple feature streams, and a robust modeling and fusion approach that shows impressive improvements and complementarity in this task.

8. RELATION TO PRIOR WORK

The proposed work is related to already published work achieved during the RATS program. To our knowledge, this is the first paper that comprehensively describes and analyzes the speaker recognition task in this program. Other work in the same program includes speech activity detection, keyword spotting, and the noise-robust feature extraction used in this paper. For noisy speaker verification, we cite [10], which inspired the authors for this work.

9. ACKNOWLEDGMENTS

The authors would like to thank all SCENIC team members for their valued input in the SID system development, Vikramjit Mitra from SRI International for producing the features for the SCENIC SID submission; Omid Sadjadi from UTD for assisting in system development and providing the MHEC feature extractor; Byung-Suk Lee and Dan Ellis (working in collaboration with ICSI) for providing the SACC pitch tracker and relevant updates; and Mark Havilla from CMU for providing the PNCC feature extractor.

10. REFERENCES

[1] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[2] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4117–4120.

[3] C. Kim and R.M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4101–4104.

[4] S.O. Sadjadi and J.H.L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5448–5451.

[5] Byung Suk Lee and Daniel P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classication," in *Proc. Interspeech*, 2012.

[6] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. Cernocky, "Recent progress in prosodic speaker verification," in *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4556–4559.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[8] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[9] Niko Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, Software available at http://www.dsp.sun.ac.za/∼nbrummer/focal/index.htm.

[10] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4253–4256.