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 Automatic 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 developing robust processing methods for 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 team is composed of speech laboratory teams from five institutions: SRI International, the International Computer Science Institute, the University of Texas Dallas, Carnegie Mellon University, and the University of California at Los Angeles. Each team focuses on robust feature extraction and speech activity detection in the context of degraded RATS data. This widespread focus provided considerable strength to the SCENIC SID submission through the complementary nature of features and speech activity detection (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 coefficient (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 coefficient (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 the squared magnitude of the analytical signal obtained using the Hilbert transform. More details can be found in [4].
- Sub-band autocorrelation classification (SACC) [5] provides a pitch estimate from an estimator that is trained using a multilayer perceptron. The resulting pitch signal and an energy signal obtained using get f0 (the pitch tracker software widely used in speech processing) are then modeled over overlapping windows

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)

of 20 ms shifted by 5 ms as described in [6] except that the approximation is done using Legendre polynomials instead of the discrete cosine transform. 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 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 longer than 10s. For audio recordings shorter 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 program. Referred to as HMM SAD in this article, this SAD consists of a feature combination frontend 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 an 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.

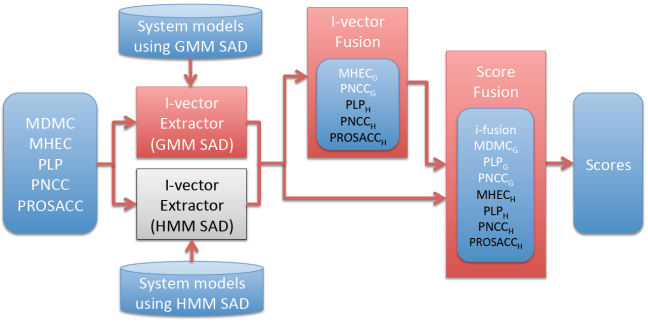


Fig. 1. SCENIC SID system involving five features, two speech activity detectors, i-vector fusion of five feature streams and score fusion of seven feature streams, along with i-vector fused scores.

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 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 PROSACC_S, where the subscript letters G and S indicate GMM SAD or HMM 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 single-system 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_H, PLP_H, PNCC_H and PROSACC_H.

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 out and applied to the other. The two sets for cross-validation contained a unique subset of speakers for both enrollment and testing.

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 a 4% 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.

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 in the dev set were required to have at least seven original (not re-transmitted) recordings. For PLDA training, segments in 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 (in the case of PROSACC, 200D i-vectors were reduced to 100D), followed by length normalization and PLDA.

6. RESULTS

Figure 2 presents the performance of individual feature streams for matched enrol and test durations. Both GMM and HMM 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 durations, illustrating their robustness to degraded conditions. PNCC in particular was 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, offering a 10% relative improvement in m4FA when added to the score level fusion of the alternate features in the 30-30 condition. In contrast to other features, PROSACC used in conjunction with HMM SAD outperformed the alternative GMM SAD. It is believed that HMM SAD provided a more continuous transition between high-energy speech frames. Since the PROSACC system only defines uniform regions of extraction (20ms long) over speech segments, more regions are defined for the HMM SAD than for the GMM SAD probably explaining the observed results.

Table 1. *The eight trial conditions evaluated.*

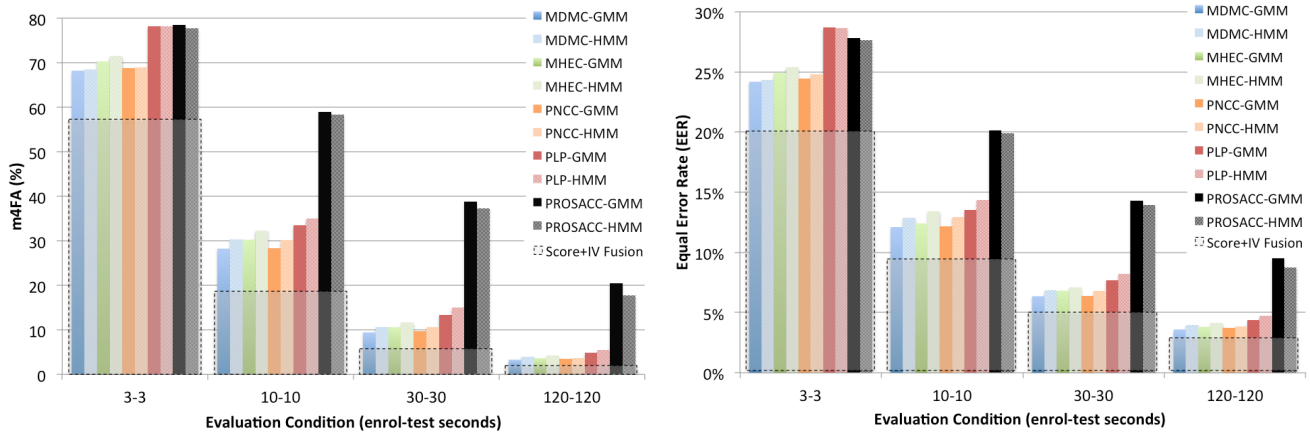
	Test (seconds)			
Enrol (seconds)	**3**	**10**	**30**	**120**
3	X	X	X	
10	X	X	X	
30			X	
120				X

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

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

Table 3. *Development set performance (m4FA / EER) of SCENIC SID system across different enrol-test conditions. m4FA: miss rate at 4% false alarm, EER: equal error rate.*

Eval. Cond.	IV Fusion	Score Fusion	Score+IV Fusion
3-3	62.6 / 21.7%	58.4 / 20.3%	57.3 / 20.0%
3-10	39.2 / 14.6%	32.9 / 13.3%	32.6 / 13.0%
3-30	25.9 / 11.0%	21.3 / 9.8%	20.4 / 9.5%
10-3	46.8 / 17.6%	44.1 / 17.2%	43.5 / 17.0%
10-10	21.5 / 10.0%	19.4 / 9.7%	18.7 / 9.4%
10-30	9.7 / 6.3%	8.8 / 6.1%	8.2 / 5.8%
30-30	6.5 / 5.1%	6.3 / 5.1%	5.8 / 4.9%
120-120	2.0 / 2.8%	2.5 / 3.1%	1.9 / 2.8%



(a) m4FA (b) EER

Fig. 2. Comparison of the individual features with GMM or HMM SAD across matched enrol and test durations overlaid with results from Score+IV fusion. m4FA: miss rate (%) at 4% false alarm, EER: equal error rate.

Table 3 provides the results of i-vector fusion, score fusion and the 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 41% in m4FA over any individual feature in Figure 2, thus demonstrating the strength of the fusion approach employed in the SCENIC SID submission.

Figure 3 illustrates the effect of adding the "next best" feature to the score-level fusion process starting with the best single feature: MDMC with GMM SAD. PROSACC was found to be the second best feature, demonstrating the considerable benefit high-level prosodic information provided in the context of degraded speech data. Three-way fusion additionally included PLP and obtained results comparable the best fusion. This selection represents three considerably different feature extraction techniques of which PROSACC and PLP were shown to provide the lowest individual performance in Figure 2.

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. Despite the numerous score streams available for fusion, competitive performance was achievable through score-level fusion of three diverse features.

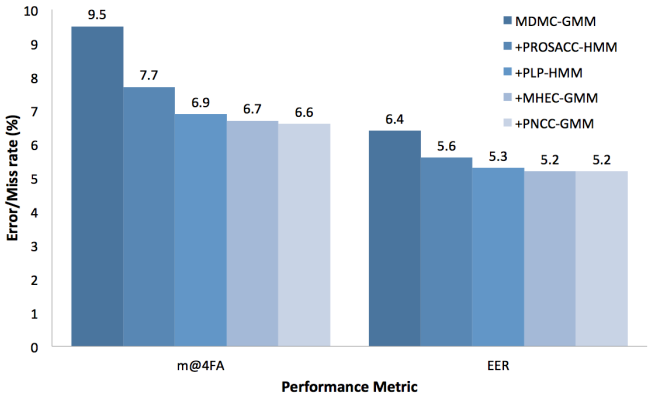


Fig. 3. Illustrating the effect of adding the "next best" feature to score-level fusion on the 30-30 evaluation condition.

8. RELATION TO PRIOR WORK

This 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 of this work.

9. ACKNOWLEDGMENTS

The authors would like to thank all SCENIC team members for their valued input; 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 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] B.S. Lee and D.P.W. Ellis, "Noise robust pitch tracking by subband autocorrelation classication," in *Proc. Interspeech*, 2012.

[6] M. Kockmann, L. Ferrer, L. Burget, and J. Cernocky, "iVector fusion of prosodic and cepstral features for speaker verification," in *Proc. Interspeech*, Florence, Italy, Aug. 2011.

[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 *Proc. IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[9] N. 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.