Recent Developments in Voice Biometrics: Robustness and High Accuracy

Nicolas Scheffer, Luciana Ferrer, Aaron Lawson, Yun Lei, Mitchell McLaren
Speech Technology and Research Laboratory (STAR)
SRI International
Menlo Park, CA
{nicolas.scheffer, luciana.ferrer, aaron.lawson, yun.lei, mitchell.mclaren}@sri.com

Recently, researchers have tackled difficult voice biometrics problems that resonate with the defense and research communities. These problems include non-ideal recording conditions that are frequently found in operational scenarios, such as noise, reverberation, degraded channels, and compressed audio. In this article, we highlight SRI's innovations that resulted from the IARPA Biometrics Exploitation Science & Technology (BEST) and the DARPA Robust Automatic Transcription of Speech (RATS) programs, as well as SRI's approach for codec-degraded speech. We show how these advancements support the case for the biometrics community adopting the use of speaker recognition.

Keywords: voice biometrics, speaker recognition, forensics

I. Introduction

During the previous few years, voice biometrics technology (aka speaker recognition) has overcome many obstacles that prevented its wide, trusted use. In particular, the research community has tackled difficult speaker biometrics problems in the context of defense and intelligence research programs. In this paper, we review the latest innovations in speaker recognition that resulted from two programs and show how these advancements support the case for adopting the use of speaker recognition by the community of biometrics users.

SRI has a long and successful track record in speaker recognition, both for its academic performance and for its many pioneering innovations, like using higher-level features such as prosody or employing speech content to improve the accuracy of speaker recognition.

In this work, we highlight SRI's innovations achieved during recent defense- and intelligence-related programs. For example, we review the robustness of systems to noisy and reverberant environments, as well as total language independence as demonstrated under the IARPA BEST program. Further, we show how speaker recognition technology has overcome the very hard challenges posed by the DARPA RATS program, which is focused on achieving high-accuracies for speaker- and language-identification problems in extreme channel-degraded environments, such as that of push-to-talk radios. We then highlight our pioneering work in mitigating the effect of audio compression on speaker recognition, showing results on a variety of codec families.

We conclude by giving a peek at the future of speaker recognition, the challenges, and the technologic advancements that will enable broadly using speaker identity in biometrics and other possible applications.

II. Speaker Recognition

A. Overview

The core speaker recognition task is usually defined as a detection problem (i.e., determining whether a specified target speaker is speaking during a given segment of speech). More explicitly, one or more samples of speech data from a speaker (referred to as the "target" speaker) are provided to the speaker recognition system. These samples are the "training" data. The system uses these data to create a "model" of the target speaker's speech. Then a sample of speech data is provided to the speaker recognition system. This sample is referred to as the "test" segment. Performance is judged according to how accurately the test segment is classified as containing (or not containing) speech from the target speaker.

Metrics that reflect accuracy are related to a typical hypothesis test (i.e., based on false positives (referred to as false alarms) and false negatives (misses)). A specific version of Receiving Operator Curve (ROC) is usually used, called Detection Error Tradeoff curve (DET) [15]. In this work, we report equal error rates (EER), where false alarm and miss rates are equal, or the false alarm rate at a particular miss rate.

B. Challenges

As for any detection task, the main challenge of speaker recognition is extracting features that will represent a speaker in the same manner independently of variations that can occur in the observations. Minimizing the intra-class variability while maximizing the inter-class variability is our goal.

Speech is a complex signal, and many possible variations of that signal exist for the same individual. During the previous few years, the community has tackled the problem of extrinsic variability and how to factor out extrinsic variability from the speaker model (sometimes referred to as channel compensation in articles). This kind of variability is detrimental to high accuracy speaker recognition. Indeed, recorded speech varies as a function of many factors that are not a function of the speaker's identity, including:

- Acoustic environment (e.g., background noise)
- Channel (e.g., microphone, handset, recording equipment)
- High signal-to-noise ratio (SNR)
- Audio degradation through compression
- Speaker's physical condition (emotion, intoxication, illness)
- What is said (text-independent versus text-dependent)
- Speaking context (level of formality, planning, language)

C. *SRI Approach for Mitigation and High Robustness*

SRI's approach to these challenges is to handle the problem at every step of the speaker recognition pipeline, and to make each pipeline stage robust to undesired variations.

SRI's system uses multiple types of features extracted from speech, which are then modeled using advanced machine learning. The systems are then optimally fused by also accounting for meta-information automatically extracted from the audio signal. We briefly present these steps below, but their combined use is what achieves maximum accuracy, as is demonstrated later in this document.

1) *Feature Diversity*

A successful approach to speaker verification is to combine different knowledge sources by separately modeling them and by fusing them at the score level to produce the final score that is later thresholded to obtain a decision. Combinations of systems are most successful when the individual systems being combined are significantly different from each other.

Prosody—the intonation, rhythm, and stress patterns in speech—is not directly reflected in the spectral features. SRI has pioneered the use of this information, showing great effect in combination with traditional features [16]. The state-of-the-art approach to extracting prosodic features is to compute the pitch and energy contour in the signal using Legendre polynomial coefficients.

We also use spectral-based features, many of which were developed specifically for noise-robustness under the RATS program. These include perceptual linear prediction (PLP) features and mel-frequency cepstrum coefficients (MFCC)—the standard features used in speech recognition. In addition, we use medium duration modulation cepstrum (MDMC) features [2], which extract modulation cepstrum-based information by estimating the amplitude of the modulation. 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 logarithmic one. Mean Hilbert envelope coefficient (MHEC) features [4] use 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. Subband autocorrelation classification (SACC) [5] provides a pitch estimate from an estimator that is trained using a multilayer perceptron. These features are referred to as PROSACC in this article.

Please note that not all features are used in the experiments to follow in the next section.

2) *Advanced Modeling*

Recently, the speaker-verification community has enjoyed a significant increase in accuracy from the successful application of the factor analysis framework. In this framework, the i-vector extractor paradigm [1] along with a Bayesian backend is now the state-of-the-art in speaker verification systems. An i-vector extractor is generally defined as a transformation where one speech utterance with variable duration is projected into a single low-dimensional vector, typically of a few hundred components.

The low rank of the i-vector itself opened up new possibilities for the application of advanced machine-learning paradigms that would have been otherwise too costly with the very high dimensionality used by most earlier systems. Probabilistic linear discriminant analysis (PLDA) [2, 3] has proved to be one of the most powerful techniques for producing an acceptable verification score. In this model, each i-vector is separated into a speaker and a channel part, analogous to the formulation in the Joint Factor Analysis framework [4].

SRI uses this state-of-the-art technology in its standard pipeline, but also pioneered its use for robustness against highly degraded conditions, such as additive noise [17].

3) *Metadata Extraction*

SRI's pipeline can account for metadata information about the audio recording. Instead of relying on annotated data, or developing specific systems, we proposed a universal audio characterization system that extracts metadata information based on the i-vector [18]. We can thus detect if an audio recording contains certain kinds of noise, channels, or the speaker's gender. The fusion process will adapt to the detected conditions in making its verification decision.

4) *System Fusion and Calibration*

Fusion of systems is performed either at the score level or at the i-vector level. At the score level, system fusion is performed using logistic regression with a cross entropy objective [6], the standard fusion approach in speaker recognition. This approach offers the benefit of producing calibrated scores, treatable as log-likelihood ratios, which are ideal for forensic comparisons and decisions.

As mentioned in [18], we developed a component that takes into account the metadata extracted from the universal audio characterization system. A modified version of the logistic regression fusion algorithm is used so that log-likelihood ratios are still produced but are biased depending on the mismatch in metadata between the enrollment and test utterances.

III. Robustness to Degraded Audio Recordings

In this section, we highlight the impact of SRI's approach for different types of degraded audio conditions and other extrinsic variations. We look at experimental results first in the IARPA BEST program for noise and reverberant speech as well as cross-language trials. We then show results on highly degraded channels from the RATS program. We conclude by showing our systems performance on compressed audio waveforms.

A. *Noise, Reverb, and Cross-Language Verification*

The IARPA BEST program[1] significantly advanced the state-of-the-science for biometrics technologies. Under this program, the speech track was focused on substantially improving the accuracy of speaker recognition under non-ideal conditions.

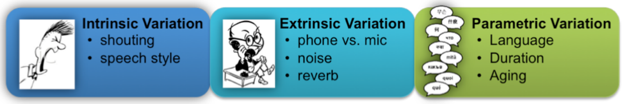

Figure 1. Variations of interest in the IARPA BEST program.

The variations of interest are depicted in Figure 1, and fall into three categories:

- Intrinsic: Vocal effort, Speech style.
- Extrinsic: Channel, Noise, Reverb, ...
- Parametric: Language, Age, Duration, …

To evaluate speaker recognition accuracy on multiple variations, SRI created the PRISM dataset [19], building on data previously collected by LDC. The PRISM corpus emulated conditions of interest to the BEST program by creating trials from waveforms degraded by adding noise or reverberation[2].

In Figure 2, we show the benefit of SRI's comprehensive approach by showing the increase in speaker recognition accuracy for every step of the pipeline.

The conditions defined in the PRISM set and represented in the horizontal axis of the figure are:

- *telphn:* telephone calls
- *intmic:* microphone recordings in an interview setting
- *telall:* telephone calls over landline but also microphones
- *voc:* vocal effort: low and high
- *lang:* Trials made of languages other than English
- *noise:* Clean signals degraded with real noise samples at different SNR levels ranging from 20 dB to 6 dB.
- *reverb:* Clean signals degraded with artificial reverb at reverb times (RT) of 0.3, 0.5, and 0.7 seconds

The baseline system is SRI's standard recognition pipeline without the mitigation mechanism for the variations of interest.

The robust system uses an enhanced i-vector PLDA model designed to be robust to the variations of interest in BEST. Improvements are highly significant, reducing error by a factor of 10 times on the noise condition while also improving results for "cleaner" conditions like telephone calls.

The robust + prosody system is a system fusion of low-level features used in the baseline and robust systems with a speaker recognition system based on prosodic information. We

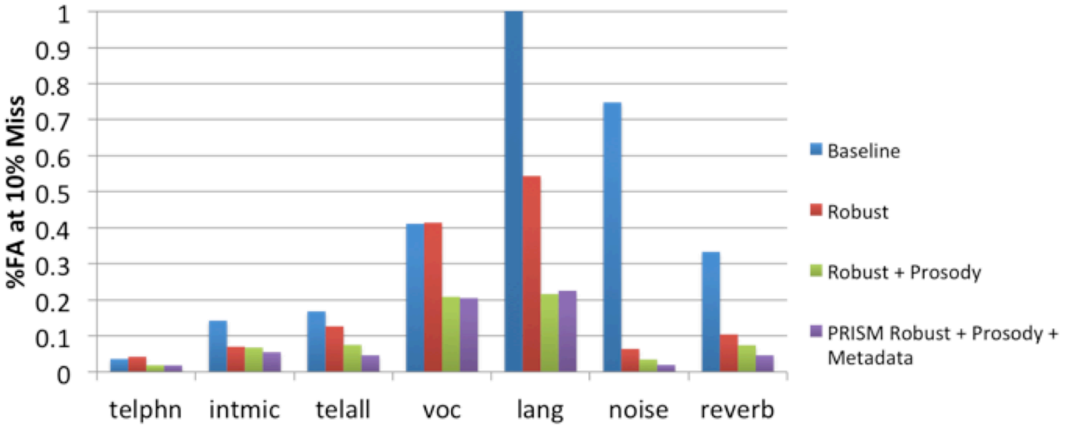

Figure 2. SRI speaker recognition system results on the IARPA BEST variations of interest. Up to 10x imporvements can be observed after enabling multiple SRI's innovation compared to the baseline.

[1] http://www.iarpa.gov/Programs/sc/BEST/best.html

[2] The PRISM data set is available online at https://code.google.com/p/prism-set/

see that an additional improvement can be observed, especially in the language and vocal effort condition.

Finally, we enable metadata extraction and handling in the robust + prosody + metadata system to obtain additional improvements for the noise and reverb condition (note that the metadata extractor was designed to extract only the noise and reverb characteristics in the signal).

To summarize, we have shown how SRI tackled the challenging problem in the BEST program of mitigating the effect of multiple, undesired variations. We succeeded by making every step of our pipeline more robust so that the compounding effect would benefit all conditions in ensemble.

B. *DARPA RATS: Speaker Identification in Noise and Channel Degraded Audio Conditions*

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 (1) speech-activity detection, (2) keyword spotting, (3) language identification, and (4) speaker identification—the last of which is described in this section. Successful speaker identification in this environment required a robust system using multiple streams of noise-robust features that were combined at a later stage in an i-vector framework [20].

The audio recordings [3] used in the RATS program are severely degraded with additive noise, channel-convolved noise, bandwidth limitations, and frequency shifting. Telephone conversations are re-transmitted over eight different military transmitter/receiver combinations. All the data was retransmitted across all the channels and re-recorded, resulting in more than 100,000 files. The core languages from which speakers are selected are Levantine Arabic, Farsi, Dari, Pashto, and Urdu.

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 the 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. 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.

TABLE I. ENROLL AND TEST DURATION COMBINATIONS.

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

The system was composed of five different features: PLP, MDMC, MHEC, PNCC, PROSACC. 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.

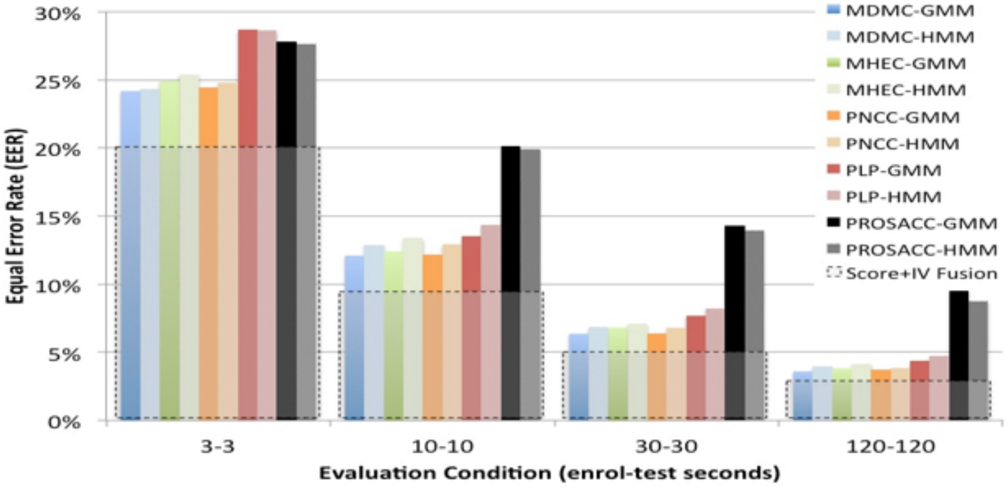

Figure 3. SRI speaker recognition system results on the DARPA RATS development set.

The 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. Fusion of systems at the score level was performed using logistic regression and a binary cross-entropy objective [6], the standard fusion approach in speaker recognition.

Results from four core conditions are provided in Figure 3 above, showing the relative performance of the five acoustic features with both HMM and GMM SAD, as well as the gain from the final score plus i-vector fusion system (in dashed lines). For all durations, the MDMC and PNCC features with GMM SAD had the least errors. The fusion system was always significantly better than any single system, benefiting in particular from the PNCC features and substantially from the inclusion of PROSACC, despite the system's low accuracy on its own.

Codec-degraded speech is commonplace in contemporary communications. The effect of transcoded speech on speaker identification and the mitigation thereof is necessary to sustain high identification performance. SRI has recently conducted some initial work to address these aspects [1].

Codec experiments involved transcoding clean microphone speech from the NIST 2008 and 2010 Speaker Recognition Evaluation (SRE) dataset using seven different codecs with a range of coding parameters. The codecs included Advanced Audio Coding (AAC); the Adaptive Multi-Rate (AMR) codec; Global Systems for Mobile communications (GSM) 6.10; MPEG-2 Audio Layer III (MP3); RealAudio; Speex; and Windows Media Audio (WMA). Readers are directed to [1] for more details on the codecs and experimental configuration.

In addition to evaluating the effect of transcoded speech on the state-of-the-art MFCC system, we evaluated two noise robust features—Power Normalized Cepstral Coefficients (PNCC) [2] and Medium Duration Modulation Cepstrum (MDMC) [3]—to observe whether noise-robustness generalizes to codec-robustness for speaker identification.

As an initial experiment, we evaluated the effect of transcoded speech on a system developed using only clean speech data. Figure 4 illustrates the considerable degradation to speaker identification performance attributed to the transcoded speech. Interestingly, the noise-robust features provided improved performance compared to MFCC on the particularly detrimental codecs, where the average EER was MFCC (3.06%), MDMC (2.63%), and PNCC (2.76%).

Next, noisy and reverberant data was added to the PLDA training dataset, and the PLDA model was retrained. Noisy data consisted of adding babble noise to 3000 segments at SNR levels of 8, 15, and 20 dB, while the reverberation RT60 times were 0.3, 0.5, and 0.7 seconds.

This experiment was designed to explore whether model robustness to noise and reverberant data generalized to robustness to transcoded speech. This was supported with results indicating a trend similar to that found in the noise experiments, with a general downward trend in average EER with MFCC (2.93%), MDMC (2.50%), and PNCC (2.61%).

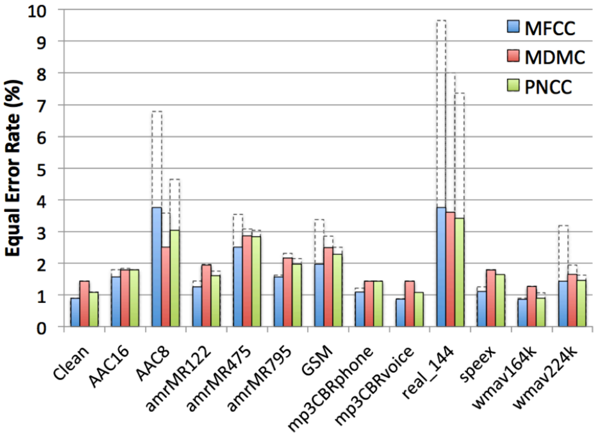

Figure 4. EER of clean and codec-degraded evaluation data using a PLDA model trained on clean, noisy, reverberated and codec-degraded speech (overlaid on EER from the "unseen codec" PLDA).

We then exposed the PLDA model to transcoded data degraded using codecs *not* used for the test sample. This mimics the real-world condition of evaluating test data degraded by unseen codecs. Interestingly, no additional robustness was observed by this technique, thus indicating that degradations from each codec are not closely correlated to alternate codecs.

Finally, the PLDA model was retrained to include all transcoded training data, as in the optimistic case in which the test codec has already been observed by the system during development. Figure 5 illustrates results from the

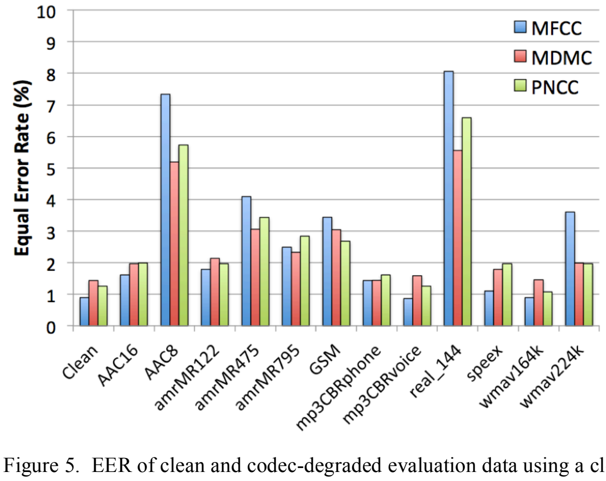

Figure 5. EER of clean and codec-degraded evaluation data using a clean speech PLDA model.

retrained PLDA overlaid on the “unseen codec” PLDA model.

Significant improvements were found when examples of the codec used for model enrollment and testing were observed during system development. The average EER was for MFCC (1.74%), MDMC (2.02%), and PNCC (1.88%). In this case, MFCC was superior to the noise-robust features (with the exception of the severely degraded EERs from the AAC8 and RealAudio codecs).

Results from these experiments suggest that current state-of-the-art speaker identification technology is not sufficiently robust to codecs not observed during system training. Particularly severe degradations such as those caused by RealAudio and AAC8 codecs can be considerably reduced by including examples of the transcoded speech in system training data; however, knowing the codecs that will be encountered is often not possible. Given the ever-changing nature of codec availability, SRI is currently researching techniques to improve speaker identification robustness to unseen codecs.

IV. Conclusion

In this work, we show the success of SRI’s approach to tackling non-ideal recording conditions for voice biometrics in multiple instances during the previous few years. We demonstrate that our comprehensive method can bring significant improvements in accuracy, whether dealing with noisy and reverberant conditions in IARPA BEST, highly degraded channels in DARPA RATS, or codec-degraded speech. SRI’s robust pipeline leverages feature diversity, advanced modeling, and system fusion based on audio metadata—key enablers of those accuracy improvements. Improvements in accuracy are seen with each approach employed in SRI’s pipeline but even more so when systems are combined and these techniques are used together.

Acknowledgment

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (#DE-AC04- 94AL85000), the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views of the funding agencies.

References

[1] M. Mclaren, V. Abrash, M. Graciarena, Y. Lei, and J. Pesan, “Improving Robustness to Compressed Speech in Speaker Recognition,” to be presented in *Proc. Interspeech,* 2013.

[2] V Mitra, H. Franco, M. Graciarena, and A. Mandal, “Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition,” in *Proc. IEEE ICASSP,* 2012, pp. 4117–4120.

[3] K. Walker and S. Strassel, “The RATS Radio Traffic Collection System,” in *Odyssey 2012—The Speaker and Language Recognition Workshop,* 2012.

[4] 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.

[5] 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.

[6] S.O. Sadjadi and J.H.L. Hansen, “Hilbert Envelope-Based Features for Robust Speaker Identification under Reverberant Mismatched Conditions,” in *Proc. IEEE Inter- national Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5448–5451.

[7] B.S. Lee and D.P.W. Ellis, “Noise Robust Pitch Tracking by Subband Autocorrelation Classication,” in *Proc. Interspeech*, 2012.

[8] 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.

[9] C. Kim and R. M. Stern, “Feature Extraction for Robust Speech Recognition Based on Maximizing the Sharpness of the Power Distribution and on Power Flooring,” in *Proc. IEEE ICASSP,* 2010, pp. 4574–4577.

[10] “NIST SRE12 Evaluation Plan,” http://www.nist.gov/itl/iad/mig/upload/NIST SRE12 evalplan- v17-r1.pdf

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Frontend Factor Analysis for Speaker Verification,” *IEEE Trans. ASLP,* vol.19, May 2010.

[12] S.J.D. Prince, “Probabilistic Linear Discriminant Analysis for Inferences about Identity,” in *ICCV-11th. IEEE,* 2007, pp. 1–8.

[13] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *Odyssey 2010—The Speaker and Language Recognition Workshop. IEEE,* 2010.

[14] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A Study of Inter-Speaker Variability in Speaker Verification,” *IEEE Trans. ASLP,* vol. 16, July 2008.

[15] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance.” In *Proc. Eurospeech,* 1997, pp. 1899-1903.

[16] 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.

[17] Y. Lei, L. Burget, and N. Scheffer, “A Noise Robust i-Vector Extractor Using Vector Taylor Series for Speaker Recognition,” in *Proc. IEEE ICASSP 2013.*

[18] 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.

[19] L. Ferrer et al., “Promoting Robustness for Speaker Modeling in the Community: The PRISM Evaluation Set,” in *Proc. of SRE11 Analysis Workshop.*

[20] M. McLaren et al., “Improving Language Identification Robustness to Highly Channel-Degraded Speech through Multiple System Fusion,” in *Proc. ICASSP 2013.*