Improving Robustness to Compressed Speech in Speaker Recognition

Mitchell McLaren[1], *Victor Abrash*[1], *Martin Graciarena*[1], *Yun Lei*[1], *Jan Pešán*[2]

[1]Speech Technology and Research Laboratory, SRI International, California, USA
[2]Speech@FIT group, Brno University of Technology, Czech Republic
{mitch, victor, martin, yunlei}@speech.sri.com, ipesan@fit.vutbr.cz

Abstract

The goal of this paper is to analyze the impact of codec-degraded speech on a state-of-the-art speaker recognition system and propose mitigation techniques. Several acoustic features are analyzed, including the standard Mel filterbank cepstral coefficients (MFCC), as well as the noise-robust medium duration modulation cepstrum (MDMC) and power normalized cepstral coefficients (PNCC), to determine whether robustness to noise generalizes to audio compression. Using a speaker recognition system based on i-vectors and probabilistic linear discriminant analysis (PLDA), we compared four PLDA training scenarios. The first involves training PLDA on clean data, the second included additional noisy and reverberant speech, a third introduces transcoded data matched to the evaluation conditions and the fourth, using codec-degraded speech mismatched to the evaluation conditions. We found that robustness to compressed speech was marginally improved by exposing PLDA to noisy and reverberant speech, with little improvement using trancoded speech in PLDA based on codecs mismatched to the evaluation conditions. Noise-robust features offered a degree of robustness to compressed speech while more significant improvements occurred when PLDA had observed the codec matching the evaluation conditions. Finally, we tested i-vector fusion from the different features, which increased overall system performance but did not improve robustness to codec-degraded speech.

Index Terms: speaker recognition, speech coding, codec degradation, speaker verification.

1. Introduction

Compressed audio plays a significant role in mobile communications, Voice Over Internet Protocol (VOIP), voicemail, archival audio storage, gaming communications, and internet streaming audio. In most of these there is a widespread use of lossy speech coders. The purpose of speech coders is to compress the speech signal by reducing the number of bits needed for transmission while maintaining the intelligibility of speech once decoded. The distortion introduced by speech coders may have a significant impact on the performance of speaker recognition systems. Of interest, therefore, is the analysis of codec-related degradation and the development of robustness techniques against this degradation. Not only the speaker recognition research community can benefit from this analysis, but also the communications, gaming, and forensics/law enforcement industries.

There have been a number of papers investigating the effect of codecs on speaker recognition performance. In [1] and [2] the effect of GSM coding is examined. Codec mismatch in model training and testing is investigated in [3]. In [4], four standard speech coding algorithms [GSM (12.2 kbps), G.729 (8 kbps), G.723 (5.3 kbps) and MELP (2.4 kbps)] were used for testing the mismatch influence for speaker recognition, and the effect of score normalization was discussed. In [5], two approaches were proposed to improve the performance of Gaussian mixture model (GMM) speaker recognition which were obtained from the G.729 resynthesized speech. The first one explicitly uses G.729 spectral parameters as a feature vector, and the second calculates Mel-filter bank energies of speech spectra built up from G.729 parameters. In [1], the effect of the codec in GSM cellular telephone networks was investigated, and performance of the text-dependent speaker verification system trained with A-law coded speech and tested with GSM coded speech and that of the system trained with GSM coded speech and tested with GSM coded speech were compared.

In this paper we analyzed the impact of speech degraded by several widely used codecs on a state-of-the-art speaker identification system. The robustness of several acoustic features was analyzed, including MFCC, PNCC, and the recent MDMC. The latter two approaches were designed for noise robustness. We analyzed the effect of codec distortion on the probabilistic linear discriminant analysis (PLDA) compensation module. Finally, we explored i-vector combination from the different feature systems as a way to increase robustness to codec distortions.

2. Speaker Identification System

A state-of-the-art speaker identification system based on a standard i-vector extraction process and PLDA modeling framework was used for this study [6, 7]. A Universal Background Model (UBM) with 512 diagonal covariance Gaussian components was trained using the female speech data from the PRISM dataset [8]. An i-vector extractor of 400 dimensions was trained on the same data and i-vector dimensions were further reduced to 250 by linear discriminant analysis (LDA), followed by length normalization and PLDA.

The i-vector fusion [9] 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.

2.1. Codecs Description

We selected a number of codecs representative of those currently in widespread use. First, the codecs were chosen according to what software was available, and would work for

The research by authors at SRI International was funded through a development contract with Sandia National Laboratories (#DE-AC04-94AL85000). The views herein are those of the authors and do not necessarily represent the views of the funding agencies.

the 8 kHz data in our evaluation set. Codecs that operated only with higher sampling rates were excluded as their assumptions regarding higher frequency and any corresponding dependence on this content for correct operation would have resulted in an unfair comparison with the selected codecs. A variety of tools was used to encode and decode (transcode) waveforms. For all data, we first used the National Institute of Standards (NIST) w_decode tool to make sure that data was single channel pulse coded modulation (PCM). Then we used sox (ver. 14.3.2) to convert the waveforms from SPHERE format to the standard WAV format. We use the following codecs in this paper.

Advanced Audio Coding (AAC) is a standardized, lossy compression and encoding scheme. Designed to be the successor to the MP3 format, AAC has been standardized as part of the MPEG-2 and MPEG-4 specifications. It is widely used in YouTube, iPhone, iOS, and Android-based phones. For waveform transcoding we used neroAacEnc and neroAacDec from http://www.nero.com to transcode AAC8 and AAC16.

The Adaptive Multi-Rate (AMR) audio codec is an audio data compression scheme optimized for speech. AMR speech codec consists of a multi-rate narrowband speech codec that encodes narrowband signals at variable bit rates ranging from 4.75 to 12.2 kbps. For AMR transcoding, we used the GSM AMR-NB speech codec (26073-800, 12/12/2001) encoder and decoder programs from http://www.3gpp.org.

GSM (Global System for Mobile communications) 6.10 [10] is a Regular-Pulse Excitation Long-Term Predictor (RPE-LTP) based codec. It was designed for speech applications, and compression is done based on signal prediction and signal correlations. GSM is the standard for the vast majority of cellular communications in the world and is optimized for real-time compression. For transcoding, we used sox, which can be found at http://sox.sourceforge.net.

MPEG-2 Audio Layer III (MP3) is a patented encoding format for digital audio that uses a form of lossy data compression. It is a common format for consumer audio storage and for most digital audio players. The compression works by reducing accuracy of certain parts of sound based on psychoacoustic criteria. For transcoding, we used http://lame.sourceforge.net.

RealAudio is a proprietary audio format developed by RealNetworks, released in April 1995. It uses a variety of audio codecs, ranging from low bit-rate formats for dialup modems, to high-fidelity formats for music. We used http://www.ffmpeg.org for transcoding to Real Audio.

Speex bases its compression on Code Excited Linear Prediction (CELP) [11], which is a traditional technique first proposed in 1985. By 1991, a U.S. Department of Defense standard was established for very low bit-rate communications based on CELP [12]. We used speexenc and speexdec from the http://www.speex.org/software software package.

Windows Media Audio (WMA) is a compression technology developed by Microsoft. We used http://www.ffmpeg.org for transcoding WMA.

3. Experimental Setup

As speech material we used a subset of female data from the NIST Speaker Recognition Evaluation (SRE) 2008 and 2010 corpora. We transcoded the full waveforms by passing the clean waveforms through a coding and decoding step. This exposes the speech to the distortion effects of the codec. The evaluation dataset was constructed from 24 SRE08 and 542 SRE10 segments to produce 559 target and 159,336 impostor trials. Matched-codec trials are reported throughout (i.e., the speaker is enrolled and tested on audio transcoded using the same codec). Speaker models were enrolled using a single segment. The baseline system was trained using 30,675 clean segments from the PRISM data set [8]. PLDA exposed to noise and reverb was trained to additionally include the corresponding noise and reverb degraded data from the PRISM data set. Transcoded data for PLDA retraining was sourced from a subset of 479 and 103 microphone segments from the SRE08 and SRE10 corpora, respectively.

Experiments were based on speech segments found by a voice activity detector (VAD) time alignments extracted from the clean speech and applied to the transcoded data. Therefore, all the experiments, clean and transcoded, contained the same sample durations. This process bypasses the problem of speech detection in transcoded data to provide an unbiased view of codec degradation on speaker recognition performance. It should be noted that some codecs change the length of the file during the encoding or decoding phase. We accommodated this by cutting the waveform appropriately to minimize the offset in speech segmentation and preliminary analysis revealed that this did not pose a significant problem.

Three sets of acoustic features were extracted from the evaluation data. The features used in the experiments are the standard MFCC features, medium duration modulation cepstrum (MDMC) and power normalized cepstral coefficients (PNCC), both described below.

The MDMC feature is obtained using a modified version of the algorithm presented in [13]. In MDMC feature generation, the digital speech signal is pre-emphasized (using a pre-emphasis filter of coefficient 0.97) and then analyzed using a 51.2 ms Hamming window with a 10 ms frame rate. The windowed speech signal is split into 34 channels using a gammatone filterbank, spaced equally between 250 Hz and 3750 Hz in the ERB scale. Amplitude Modulation (AM) signals are estimated from the subband signals using Teagers nonlinear energy operator. The AM power for each subband is estimated at a 100 Hz sampling rate. Medium duration power bias subtraction is performed on the resulting power signal, which is then power normalized using 1/15th root. Discrete Cosine Transform (DCT) was performed on the root compressed power signal and the first 20 coefficients (including the C0) were retained. These 20 coefficients along with their deltas and double-deltas resulted in a 60-dimension feature set.

PNCCs are a noise-robust acoustic feature based on work by Kim and Stern [14]. In PNCC, the acoustic digital signal is pre-emphasized (using a pre-emphasis coefficient of 0.97) and then analyzed using a 25.6 ms Hamming window with 10 ms frame rate. Then a short-time Fourier analysis is performed over the Hamming windowed data, followed by gamma-tone filtering in the spectral domain, using a 30-channel gammatone filterbank with cut-off frequencies of 133 Hz and 4000 Hz, where the center frequencies of the gamma-tone bank are spaced equally in the ERB scale. In this implementation of PNCC, small power boosting is supported as explained in [15]. Short-term spectral powers were estimated by integrating the squared gamma-tone responses, and the resultant was compressed using 1/15th root. DCT was performed on the root-compressed power signal, and the first 20 coefficients (including the C0) were retained. These 20 coefficients along with their $\triangle$s and $\triangle^2$s resulted in a 60D PNCC feature set.

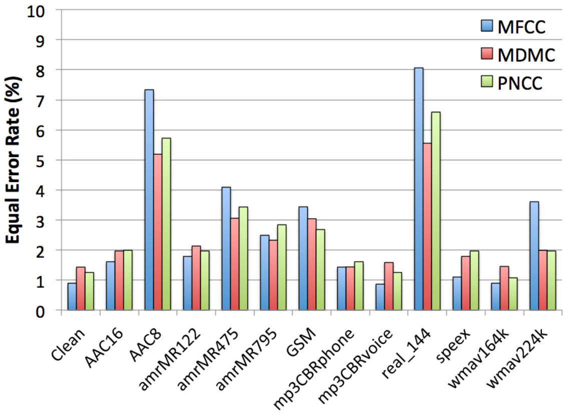

Figure 1: Equal Error Rate (EER) of clean and codec-degraded evaluation data using a clean speech PLDA model.

4. Results

We first analyzed the effect of codecs on speaker identification performance. The main subject of investigation is how the PLDA compensation system is affected when trained with clean data, additional noisy and reverberated speech and finally additional transcoded speech. We evaluate the particular case where the PLDA model is exposed to compression from codecs other than the one being tested. The contribution of i-vector combination to robustness in the unseen codec case is also analyzed.

4.1. Analysis of Codec-Degraded Speech

The first system can be considered the baseline. Here the PLDA model was trained with clean speech data. Results are presented in Figure 1 in terms of Equal Error Rate (EER) when speech is degraded using each codec. Results are presented for each of the three features described in Section 3. From Figure 1 we can conclude that several codecs result in significant EER degradation relative to clean conditions. Specifically, AAC8 and real audio 144 reduce performance by an order of magnitude. Compared to the standard MFCC features, the noise-robust MDMC and PNCC features offer some robustness to codec degradation in the more extreme cases (GSM, AAC8, Real_144, amrMR475 and wmav2_24k), while the opposite trends were observed when codec degradation was less severe (AAC16 and speex for instance). The difference between features can be summarized by the average EERs from transcoded evaluation data MFCC (3.06%), MDMC (2.63%) and PNCC (2.76%).

4.2. Exposing PLDA to Noise and Reverb

We re-trained the PLDA model with the aforementioned clean data and additional noisy and reverberated data. Around 3,000 segments were used for each condition. Noisy data is from babble noise degraded speech at 8, 15, and 20 dB SNRs. The RT60 reverberation time parameters were 0.3, 0.5 and 0.7. The goal of this experiment is to analyze whether a PLDA model exposed to noisy and reverberated speech is more robust to codec distortions than one trained only with clean speech. Results are presented in Figure 2 using the noise and reverb PLDA model. For ease of comparison, results are overlaid on the EERs from the clean PLDA model represented as dashed bars in the Figure.

Figure 2 indicates that a general downward trend in EER can be observed by introducing noisy and reverberated data in the PLDA model. One exception was the Real Audio codec for which the noise and reverb PLDA model degraded performance with respect to the clean model. The average EER from transcoded speech in Figure 2 for MFCC is 2.93%, for MDMC is 2.50% and for PNCC is 2.61%. We found similar feature rank ordering compared to the clean system and therefore it can be concluded that including noisy and reverberated speech in the PLDA model provided some additional system robustness to codec-degraded speech.

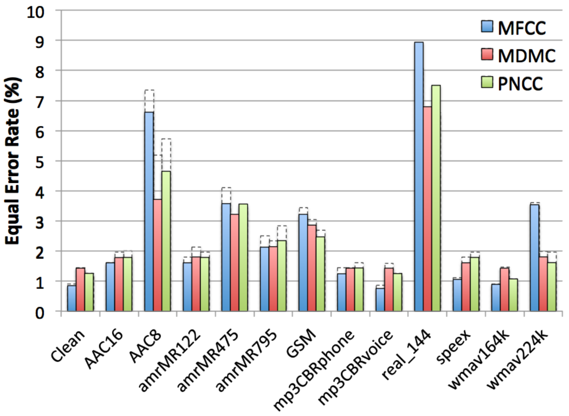

Figure 2: EER of clean and codec-degraded evaluation data using a PLDA model trained on clean, noisy and reverberated speech (overlaid on EER from Figure 1).

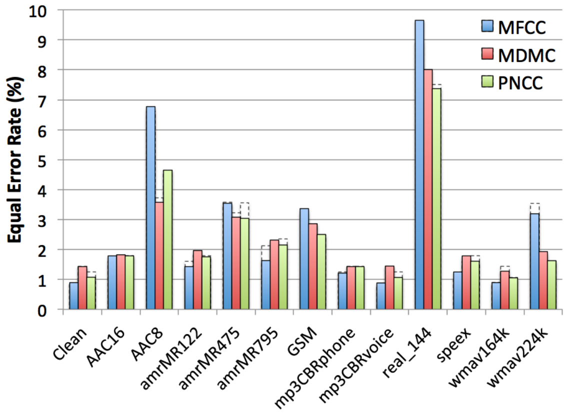

Figure 3: EER of clean and codec-degraded evaluation data using a PLDA models trained on clean, noisy, reverberated, and codec-degraded speech except the evaluated codec (overlaid on EER from Figure 2).

4.3. Unseen Codec Experiments

To observe the benefit of re-training the noise and reverb PLDA model along with codec-degraded data, we included in the training set 582 segments transcoded with each codec except the codec used for the test. The goal is to analyze if a PLDA model exposed to multiple codec degradations other than the one used in the testing case is more robust than a PLDA training set without codec-degraded speech. Specifically we grouped the codecs by classes: aac, amr, mp3, GSM, real, wmav and speex. The codec group from which the enrollment and testing came was excluded during PLDA re-training. Figure 3 illustrates the results from these trials overlaid on the EERs from the noise and reverb PLDA model (represented as dashed bars in the Figure). Despite expectations of additional robustness to compressed speech, it can be seen that including compressed

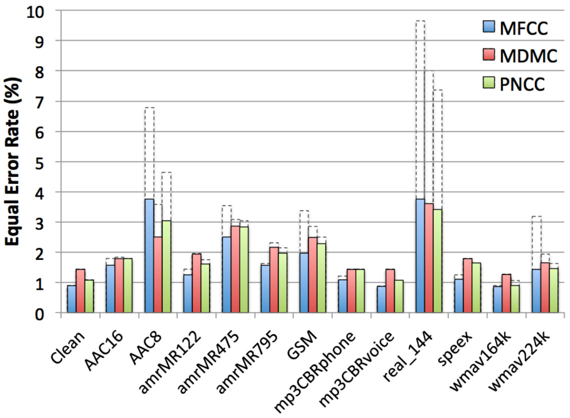

Figure 4: EER of clean and codec-degraded evaluation data using a PLDA model trained on clean, noisy, reverberated, and codec-degraded speech (overlaid on EER from Figure 3).

speech in the PLDA model provided no significant improvement over the noise and reverb PLDA model.

4.4. PLDA Using All Codecs

Here, we explore the case where the PLDA model is exposed to all available codec-degraded speech along with noisy and reverberate speech. This can be considered an optimistic case where the PLDA system has been exposed to multiple codecs including the one used in the enrollment and testing audio. Results when using the codec-aware PLDA model are are overlaid on EERs from the noise and reverb PLDA model (represented by dashed bars) in Figure 4. Figure 4 indicates that including in the PLDA training data speech compressed with the same codec as used for model enrollment and testing significantly lowered EERs. The average EER of transcoded speech for MFCC is 1.74%, for MDMC is 2.02% and for PNCC is 1.88%. Interestingly, the noise-robust features no longer improve over MFCC. In fact, the order of feature performance was reversed, with MFCC improving on PNCC and PNCC improving on MDMC. As previously observed, however, the noise-robust features provided improved performance over MFCC in the high EER codecs: AAC8 and Real Audio. These results indicate that, unless the codec used to transcode enrollment and test data has been observed during PLDA training, the system will offer limited robustness to the degradation that transcoding imparts to the speech.

4.5. I-vector Fusion System

Finally, we explored combining the three feature-specific systems. System combination was performed by combining the i-vectors from the MFCC, MDMC, and PNCC systems per Section 3. To summarize the benefit of i-vector fusion, Table 1 details the average EER across codecs for the individual features and i-vector fusion under the four PLDA modeling conditions. In each case, it can be observed that i-vector fusion provides a relative improvement of 10-14% over the best single system. Robustness to compression from unseen codecs, however, was on average not improved by i-vector fusion, as the noise and unseen PLDA results are comparable and far from the oracle codec PLDA. The improvements of i-vector fusion over the individual features are illustrated for individual codecs in Figure 5 when the test codec was not observed during PLDA training.

Table 1: Comparing Average EERs of features and i-vector fusion across codecs for different PLDA training strategies.

PLDA	MFCC	MDMC	PNCC	ivFusion
Clean	3.06%	2.63%	2.76%	2.37%
Noise & Reverb	2.93%	2.50%	2.61%	2.16%
Unseen Codecs	2.97%	2.63%	2.51%	2.17%
All Codecs	1.81%	2.08%	1.95%	1.56%

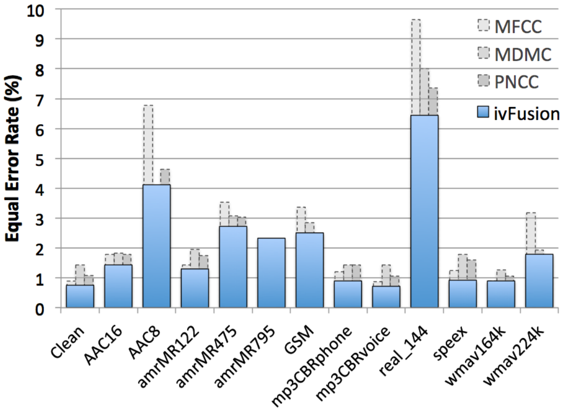

Figure 5: EER of i-vector fusion system overlaid on individual feature EER when including codec training data in the PLDA model that is mismatched to the evaluation codec.

The lack of system robustness to unseen codecs provides motivation for research into techniques tailored to dealing with codec robustness in state-of-the-art speaker recognition systems. Future research will investigate how codec detection can be utilized to improve robustness through metadata information for system calibration, selecting a pre-trained PLDA model that has observed codecs with similar characteristics during training and tailoring features to better deal with compressed speech.

5. Conclusion

We analyzed the impact of codec-degraded speech on a state-of-the-art PLDA-based speaker identification system and proposed mitigation techniques. Baseline MFCC and noise-robust MDMC and PNCC features were analyzed. We compared four PLDA modeling regimes: one trained only on clean data, additionally including noisy and reverberant speech, the inclusion of all codec data including the codec used to degrade enrollment and test speech, and all aforementioned data except speech from the codec used in evaluation. It was found that including noise and reverberant speech in the PLDA model added some robustness to codec-degraded speech with no advantage coming from adding transcoded speech from codecs not used for enrollment and test speech. The optimal solution was to include codec data in the PLDA model that matched the evaluation conditions. The noise-robust MDMC and PNCC features were found to generalize well to codec-degraded speech when the codec used on evaluation data had not been observed during PLDA model training. Finally, we tested i-vector level combination of the different feature subsystems, which improved overall performance of the system by as much as 14% relative but failed to improve robustness to codec-degraded speech.

6. References

[1] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, and F. Pellandini, "Influence of GSM speech coding on the performance of text-independent speaker recognition," in *Proceedings of EUSIPCO*, 2000, pp. 437–440.

[2] M. Kuitert and L. Boves, "Speaker verification with GSM coded telephone speech," in *Proc. Eurospeech*, vol. 97, 1997, pp. 975–978.

[3] M. Phythian, J. Ingram, and S. Sridharan, "Effects of speech coding on text-dependent speaker recognition," in *Proc. TENCON'97*, vol. 1, 1997, pp. 137–140.

[4] T. Quatieri, R. Dunn, D. Reynolds, J. Campbell, and E. Singer, "Speaker recognition using G. 729 speech codec parameters," in *Proc. IEEE ICASSP*, vol. 2, 2000, pp. II1089–II1092.

[5] R. Dunn, T. Quatieri, D. Reynolds, and J. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *Proc. Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, vol. 2, 2001, pp. 1562–1567.

[6] N. Dehak, P. J. 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, 2011.

[7] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE ICCV*, 2007, pp. 1–8.

[8] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proc. NIST SRE Analysis Workshop*, 2011.

[9] M. Kockmann, L. Ferrer, L. Burget, and H. Cernocky, "iVector fusion of prosodic and cepstral features for speaker verification," in *Proc. Interspeech*, 2011, pp. 265–268.

[10] M. Mouly and M.-B. Pautet, *The GSM system for mobile communications*. Telecom Publishing, 1992.

[11] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE ICASSP*, vol. 10, 1985, pp. 937–940.

[12] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The federal standard 1016 4800 bps CELP voice coder," *Digital Signal Processing*, vol. 1, no. 3, pp. 145–155, 1991.

[13] 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.

[14] 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.

[15] C. Kim, K. Kumar, and R. M. Stern, "Robust speech recognition using a small power boosting algorithm," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 243–248.