Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions

Ali Ziaei[1], *Lakshmish Kaushik*[1], *Abhijeet Sangwan*[1], *John H.L. Hansen*[1] *and Doug Oard*[2]

[1]Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas, U.S.A
[2]College of Information Studies and UMIACS
University of Maryland, College Park, MD, U.S.A.
{ali.ziaei, lakshmish.kaushik, abhijeet.sangwan, john.hansen}@utdallas.edu, oard@umd.edu

Abstract

Speech Activity Detection(SAD) is a well researched problem for communication, command and control applications, where audio segments are short duration and solution proposed for noisy as well as clean environments. In this study, we investigate the SAD problem using NASA's Apollo space mission data [1]. Unlike traditional speech corpora, the audio recordings in Apollo are extensive from a longitudinal perspective (i.e., 6-12 days each). From SAD perspective, the data offers many challenges: (i) noise distortion with variable SNR, (ii) channel distortion, and (iii) extended periods of non-speech activity. Here, we use the recently proposed Combo-SAD, which has performed remarkably well in DARPA RATS evaluations, as our baseline system [2]. Our analysis reveals that the Combo-SAD performs well when speech-pause durations are balanced in the audio segment, but deteriorates significantly when speech is sparse or absent. In order to mitigate this problem, we propose a simple yet efficient technique which builds an alternative model of speech using data from a separate corpora, and embeds this new information within the Combo-SAD framework. Our experiments show that the proposed approach has a major impact on SAD performance (i.e., +30% absolute), especially in audio segments that contain sparse or no speech information.
Index Terms: Speech Activity Detection, Long Audio Recordings, NASA, Apollo, Noise Robustness

1. Introduction

Speech Activity Detection (SAD) systems distinguish speech from non-speech in audio, and help input speech-only information to upstream applications such as speech and speaker recognition. SAD is a well researched problem, and good solutions exist for traditional conversational telephony speech (CTS) [3, 4, 5]. More recently, researchers have been focussed on developing noise robust SAD systems, with emphasis on additive noise, channel distortions, *etc.* [2]. For example, the DARPA RATS (Robust Automatic Transcription of Speech) program is investigating SAD problem for channel distorted speech [6, 7, 8]. For the most part, data considered in typical SAD studies tend to be conversational telephony speech where two people talk to each other for short duration. Naturalistic and long duration continuous audio recordings in other hand, are very interesting and challenging in terms of speech activity detection and speech analysis. Prof-Life-Log database [9, 10] is from this group. Therefore, the collection provides the opportunity to address some interesting questions related to speech, speaker, environment and language[11, 12, 13].

In this study, we investigate the SAD problem in the context of long duration audio files which last several days. Particularly, we focus on NASA's Apollo space mission audio data [1]. A key attribute of the Apollo recordings is that they are continuous and natural. This introduces several unique challenges for speech systems including SAD. In what follows, we describe some of these challenges from a SAD perspective.

During the NASA 1960's/70's space missions, the use of head-mounted Plantronics microphones was common, both on the spacecraft and on the ground, but some recordings made on the spacecraft were made using fixed far-field microphones which also picked environmental noise (e.g., glycol cooling pumps and thruster firings) that varied over time. Some of these environments tend to have complex harmonic structure which is readily confused as speech by SAD. Additionally, mission personnel used both push-to-talk and voice-operated-keying. Furthermore, space-to-ground radio communication (and recordings from the recorder on one of the spacecrafts) reached the recording facility in Houston Texas through one of about a dozen ground stations, each of which had a different receiver noise temperatures, and each of which used a different cascade of terrestrial channels to get the signal back to Houston. These factors contributed towards distorting speech, which adds to the challenge of performing effective SAD. Altogether, NASA flew eleven manned Apollo missions that varied in duration between nearly six and more than twelve days. As would be expected, there were periods in which communication was not possible (e.g., during occultation by the Moon and during crew sleep periods) and there were other periods in which radio communication was not initiated (e.g., during meal breaks). The lack of a side channel for text uplink resulted in long sequences of numeric data being read up to the spacecraft. Communications in either direction were typically acknowledged. The net result was an exceptionally broad dynamic range of non-speech durations, ranging from milliseconds to (in the case of sleep periods) nearly half a day. This aspect of the data is in direct contrast with typical speech corpora where speech and pause tends to be balanced in the data capture. In fact, our study shows that long durations of non-speech tends to be the biggest contributer towards SAD false-alarms in the Apollo data.

All these factors make speech activity detection challenging. While some of these problems have been explored in the past, the Apollo data is perhaps unique in the sense that all the

This work was supported by NSF under Grant 1218159.

mentioned challenges can be found in a single audio file. On the positive side, however, the Apollo missions were carefully scripted, often on very tightly constrained timelines that are available to us today. Moreover, three of the recordings have been fully transcribed to an engineering analysis standard (i.e., as coherent interactions, rather than with the timing precision needed for speech recognition). As a result, we have considerable a priori information about the timing of high-activity periods. This information is very valuable as ground truth for SAD (and other speech system) evaluation.

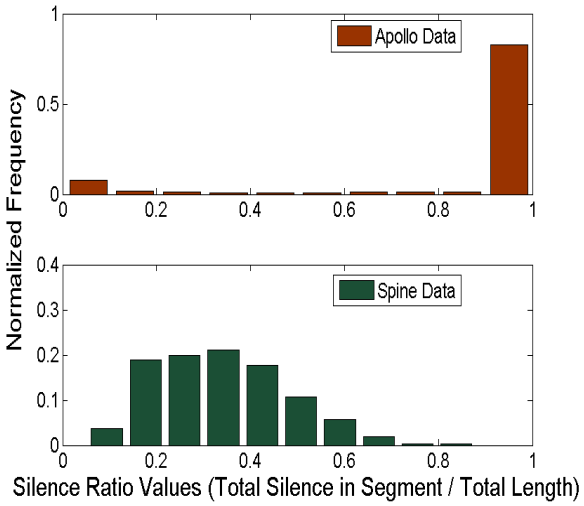

Figure 1: Comparing silence ratio values for Apollo 11 data and SPINE corpus.

2. Data Analysis

In this section, we analyze the speech-pause and SNR (signal to noise ratio) characteristics of the Apollo data, and compare it to SPINE (speech in noisy environment) corpora [14]. The SPINE corpus contains a number of talker-pair conversations where the participants are working on a collaborative battleship-like task.

As mentioned previously, the Apollo data contains long periods of non-speech along with some periods of intense speech activity. This is unlike typical speech corpora where speech and pause tends to be well balanced. To illustrate this fact, we performed the following analysis: we gathered a number of 1-minute cuts from Apollo and SPINE. Next, we computed the silence ratio for the 1-minute cuts, where the silence ratio was defined as proportion of pause duration in the 1-minute cut. Hence, large and small values of silence ratio indicates pause and speech dominants cuts, respectively. In Fig. 1, we compare the distribution of the silence ratios for SPINE and Apollo. From the figure, it is easy to see that SPINE is dominated by speech-pause balanced and speech-dominant cuts. On the other hand, Apollo data is dominated by pause-dominant cuts. This difference in speech-pause distribution has direct impact on SAD design and performance.

In order to analyze the SNR characteristics of the two datasets, we use the NIST-STNR and WADA SNR tools [15]. The STNR and SNR distributions are shown in Fig 2 for both datasets. It can be seen from the figure that the mean STNR and SNR values for Apollo dataset are lower than those for SPINE. This demonstrates that the background acoustic noise characteristics of the Apollo data is far more challenging than SPINE.

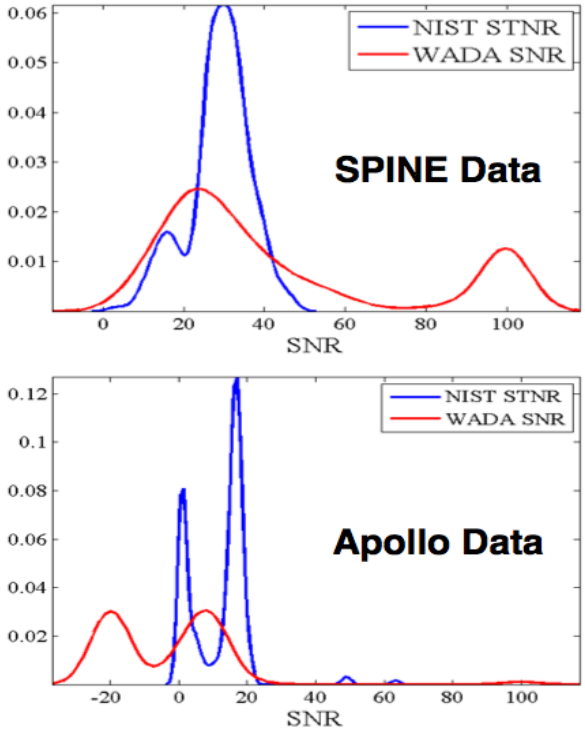

Figure 2: SNR and STNR comparison for SPINE and Apollo data.

For the purpose of SAD evaluation in this study, we selected around 3 hours of data from the Apollo 11 mission for SAD evaluation. We quality checked the available transcripts for the mission to make time adjustments to speech pause boundaries. Additionally, the data was segmented into 1-minute cuts (180 cuts in total). The data selection was performed in a manner such that 1 hour of audio was speech-dominant (more than 40 seconds speech in 1 minute cut), 1 hour was pause-dominant (more than 40 seconds pause in 1 minute cut), and 1 hour was speech-pause balanced (at least 20 seconds of speech and pause in 1 minute cut).

3. Proposed System

In this section, we use the UTDallas Combo-SAD as our baseline system [2]. This unsupervised SAD technique is designed to be noise robust and has been particularly effective in multiple RATS evaluations [2, 8].

For long duration audio recordings, we segment the long recording into contiguous 1-minute cuts, and then run the Combo-SAD independently on each cut. The Combo-SAD method computes several noise robust features at a frame level for each audio cut and projects the combined feature vectors into a single dimension (by using Principal Component Analysis). Let f_i be the Combo feature vector for the i^{th} frame and $\bar{f}_i$ is the normalized feature vector,

$$\bar{f}_i = \frac{f_i - \mu}{\sigma}, \quad (1)$$

where μ and σ are the mean and standard-deviation of the feature vectors for a cut. Now, let X be the principal eigenvector (corresponding to the largest eigenvalue of the feature covari-

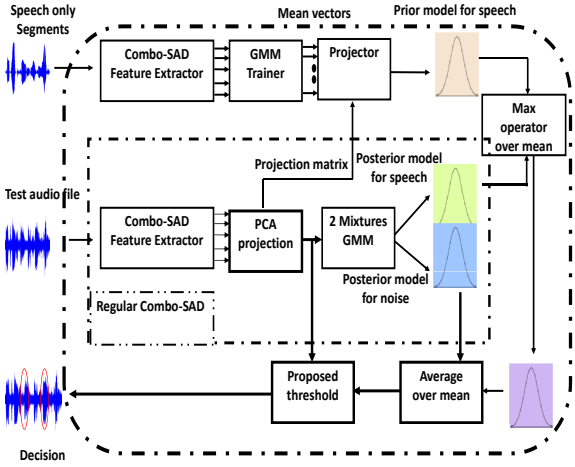

Figure 3: System Block Diagram

ance matrix). Finally, let p_i be the projection of $\bar{f}_i$ on X,

$$p_i = X^T \bar{f}_i, \tag{2}$$

The Combo features are designed to have higher values for speech and lower values for noise/background. Therefore, p_i value will be generally higher for speech than background. The Combo-SAD exploits this principle for decision making. It trains a two-mixture GMM (Gaussian Mixture Model) to automatically determine the speech and background clusters. The mixture with larger mean value is hypothesized to belong to speech and vice-versa. Let μ_{hs} and μ_{hp} be the hypothesized speech and background mixture means of the GMM. In the next step, the mixture means are used to compute the SAD threshold and speech/pause decisions are made. The threshold value is computed using a simple convex combination, *i.e.*,

$$\tau = w\mu_{hs} + (1-w)\mu_{hp}, \tag{3}$$

where w is the weight factor such that $0 \leq w \leq 1$.

The threshold estimation method implicitly assumes that the audio file always contains speech and pause in reasonably balanced proportions. Our experience with long audio recordings has shown that this assumption is frequently wrong, and leads to large number of false alarms. Because long durations of non-speech are very common in Apollo data, a large number of 1-minute segments have no-speech to very-little-speech. Additionally, speech activity tends to be bursty in Apollo data, and a number of 1-minute segments also tend to be speech-dense. In both these cases, the Combo SAD builds relatively poor estimates of the speech and pause models, respectively, which leads to a poor threshold, which in turn leads to errors.

Here, we propose a simple yet effective solution to mitigate the problem of threshold determination. In order to build an effective speech model, we first train a large mixture GMM on speech data extracted from annotated corpora (typical sources are Switchboard, Fisher, etc.). Next, the means of this GMM are projected into the Combo SAD's single-dimension decision making space,

$$\hat{m}_j = X^T m_j, \tag{4}$$

where m_j is the j^{th} mixture mean of the M-mixture GMM, and $\hat{m}_j$ is the corresponding projected value. Furthermore, let μ_{ts} be the mean of projected values $\hat{m}_j$,

$$\mu_{ts} = E[m_j]. \tag{5}$$

It is noted that μ_{ts} can be viewed as prior model of speech (since it was built with speech data from annotated corpora). Also, note that μ_{hs} can be viewed as posterior model of speech (since it is built by Combo-SAD from data). Since higher values on the Combo-SAD's projected dimension are more likely to be speech , if $\mu_{hs} \geq \mu_{ts}$, then we trust the posterior model of speech and use it for decision making. On the other hand, if $\mu_{hs} < \mu_{ts}$, then we trust the prior model of speech and use it for decision making. This new method of threshold estimation can be written as follows:

$$\tau = w \max(\mu_{ts}, \mu_{hs}) + (1-w)\mu_{hp}. \tag{6}$$

The proposed modification has significant impact on the performance of Combo-SAD on speech-sparse and non-speech regions. For both these regions, the estimation of μ_{hs} tends to be poor and leads to large number of false-alarms. Here, the proposed method addresses the problem by defaulting to the prior model of speech for decision making (and discarding μ_{hs} completely). For regions of the audio file where speech is balanced or dense, the decision making continues to rely on μ_{hs} in the proposed method. Fig. 3 captures the entire data flow for the proposed system.

4. Results and Discussion

In this section, we evaluate the proposed method on the Apollo data, and compare performance to a baseline (*i.e.*, Combo-SAD). Fig. 4 compares the operation of Combo-SAD and proposed method of 3 examples files. The three examples were chosen to show the operation on purely non-speech, speech sparse and speech-dense segments. The figure shows the spectrogram of the original audio files, along with the Combo-SAD and proposed SAD decisions. It is useful to note that we have chosen a weight factor w corresponding to the EER (equal error rate) (computed over all files) to make the SAD decisions for the Combo-SAD and proposed SAD method. For comparison, the ground truth SAD decisions are also shown. Finally, the projected value of the Combo-features (p_i) is also shown.

For the non-speech segment, it can be seen that the audio file contains two different noise types and this has been registered in projected values (p_i) (as the value of p_i shifts upwards midway through the audio file). For this file, the speech model (μ_{hs}) built by the Combo-SAD is incorrect, and leads to a high number of false-alarms (as seen in the decision curves). However, in the proposed method, the prior model of speech (μ_{ts}) is employed for threshold estimation, and this leads to perfect decision making (as seen in the decision curve).

In the case of the speech-sparse segment, it can be seen that the speech duration is very small compared to the overall file duration. Because speech information is sparse, the value of μ_{hs} determined by the EM (expectation maximization) algorithm is biased and incorrect (which leads to higher false-alarms). The proposed method solves the problem by employing the prior model of speech (μ_{ts}) (and fewer false-alarms are observed).

Finally, in the case of speech-dense segment, it can be seen that the Combo-SAD has adequate information to build a good estimate of μ_{hs}. Consequently, the SAD decisions are fairly accurate. Moreover, the proposed method also employs the posterior model of speech and the decisions are identical to Combo-SAD.

Fig. 5 shows DET (detection error trade-off) curves for the proposed method and Combo-SAD. The DET curves for overall performance are shown along with curves for (i) speech-sparse, (ii) speech-pause balanced , and (iii) speech-dense data.

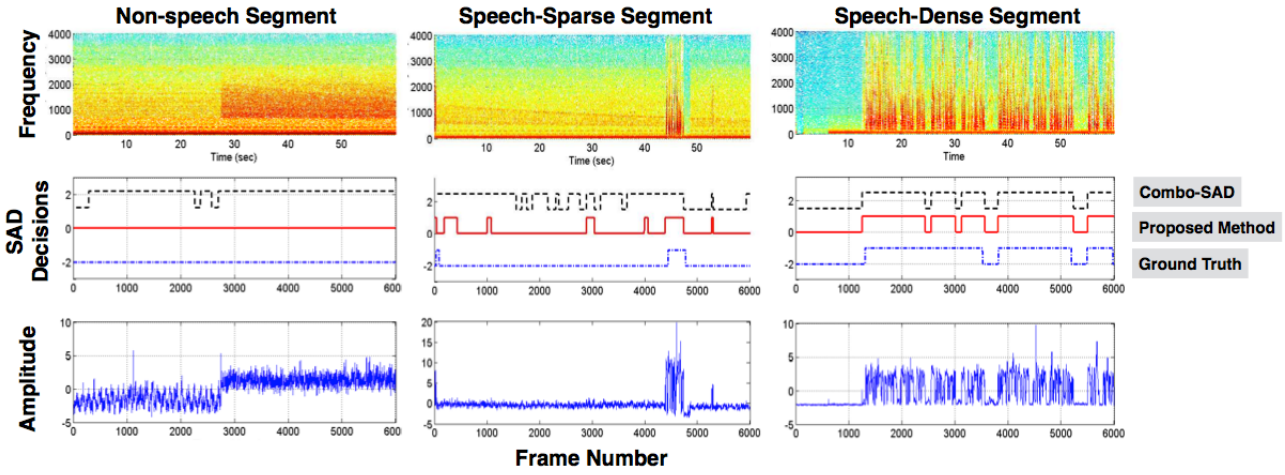

Figure 4: Comparison between proposed method and Combo-SAD on three example files.

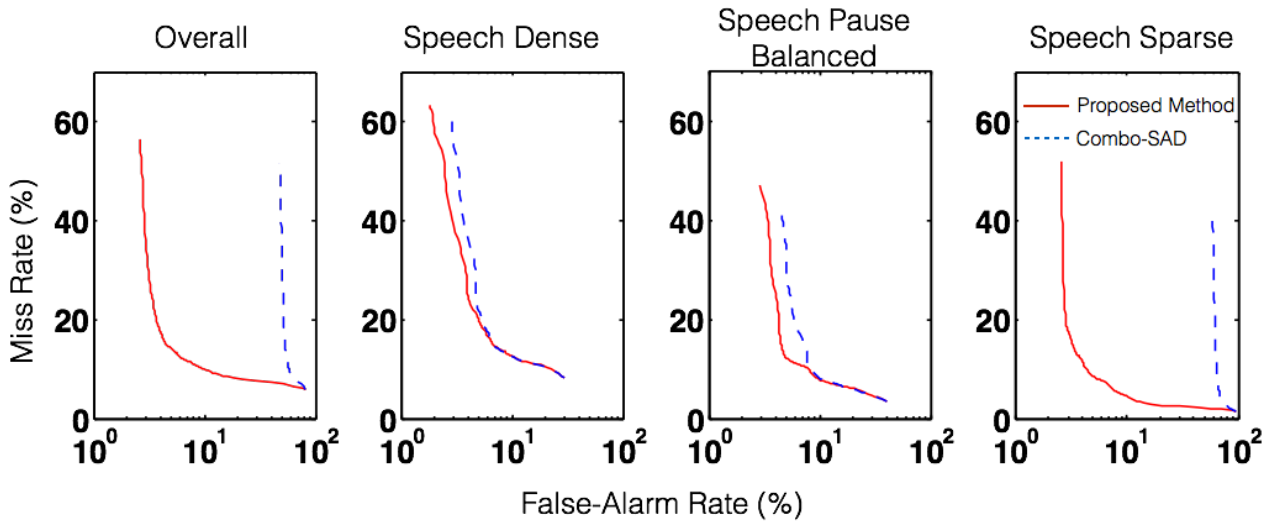

Figure 5: DET curves for APOLLO data utterances.

From the figure, it can be seen that the proposed method obtains a remarkable improvement over the state-of-the-art unsupervised Combo-SAD baseline for speech-sparse data. Due to the formulation of the proposed system, it is very likely that the prior model of speech is frequently chosen for decision making in speech-sparse region (and this explains the huge performance improvement). Interestingly, we also observe moderate improvements for speech-pause balanced and speech-dominant cuts as well. For these two cases, the improvements are seen in terms of false-alarm reduction for higher values of miss-rate. Overall, the EER drops from around 40% to 10%.

5. Conclusion

In this study, we have shown that long duration audio streams such as NASA Apollo missions data have very different speech pause characteristics from typical speech corpora used in the speech community. Specifically, the data has long durations of non-speech combined with durations of intense speech activity (whereas typical speech corpora have more balanced speech-pause profile). Consequently, a dramatic performance drop is seen, even for extremely competitive systems such as the Combo-SAD (as it makes incorrect assumptions about speech-pause distribution characteristics). It is very likely that long duration audio recordings in general may have different statistical properties from typical collections (our experience with another long duration corpus called Prof-Life-Log [**?**] further strengthens this belief). In this study, we proposed a simple yet effective technique which modifies the threshold computation for Combo-SAD. Specifically, the proposed technique first builds a prior model of speech (using external corpora), and dynamically chooses between the prior and posterior speech models (posterior model is built from data by Combo-SAD in an unsupervised fashion). The new method is designed to chose the prior model whenever the posterior model is weak. Therefore, it consistently delivers superior results over Combo-SAD. In our experimental evaluation using actual Apollo 11 data, the proposed method demonstrates a dramatic decrease in EER (equal error rate) from about 40% to about 10% when compared to Combo-SAD.

6. References

[1] A. Sangwan, L. Kaushik, C. Yu, John H. L. Hansen and D. Oard, "Houston, We have a solution: Using NASA Apollo Program to advance Speech and Language Processing Technology," Interspeech 2013.

[2] S. O. Sadjadi and John H. L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, Vol. 20, No. 3, March 2013.

[3] J. Ramirez et al. "Efficient voice activity detection algorithms using long-term speech information." Speech Communication, Vol. 42, No. 3, 2004, pp. 271-287.

[4] A. Davis, S. Nordholm, and R. Togneri. "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold." IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 2, 2006, pp. 412-424.

[5] A. Sangwan, Wei-Ping Zhu, and M. O. Ahmad. "Design and Performance Analysis of Bayesian, Neyman-Pearson, and Competitive Neyman-Pearson Voice Activity Detectors." IEEE Transactions on Signal Processing, Vol. 55, No. 9, 2007, pp. 4341-4353.

[6] T. Ng et al. "Developing a Speech Activity Detection System for the DARPA RATS Program." Interspeech 2012.

[7] G. Saon et al. "The IBM speech activity detection system for the DARPA RATS program." Interspeech 2013.

[8] M. Graciarena et al. "All for One: Feature Combination for Highly Channel-Degraded Speech Activity Detection." Interspeech 2013.

[9] A. Sangwan, A. Ziaei, J. H.L. Hansen, "ProfLifeLog: Environmental analysis and keyword recognition for naturalistic daily audio streams," ICASSP, pp. 4941-4944, Kyoto, Japan, 2012.

[10] A. Ziaei, A. Sangwan, J. H.L. Hansen, "Prof-Life-Log: Personal interaction analysis for naturalistic audio streams," ICASSP, pp. 7770-7774, Vancouver, CA, 2013 .

[11] A. Ziaei, A. Sangwan, J. H.L. Hansen, "Prof-Life-Log: Audio Environment Detection for Naturalistic Audio Streams.," Interspeech 2012.

[12] A. Ziaei, A. Sangwan, J. H.L. Hansen, " Prof-Life-Log: Robust Audio Environment Detection for Naturalistic Audio Streams Using LENA Device," LENA FOUNDATION Conference, Denver, Colorado, USA, April 28-30, 2013

[13] H. Boril, A. Ziaei, J. H.L. Hansen, " Prof-Life-Log: Production of Conversational Speech as a Function of Varying Environment," LENA FOUNDATION Conference, Denver, Colorado, USA, April 28-30, 2013.

[14] Astrid Schmidt-Nielsen, et al., "Speech in Noisy Environments (SPINE) Evaluation Audio," Linguistic Data Consortium, Philadelphia, 2000.

[15] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," INTERSPEECH-2008, pp. 2598-2601.