ANALYSIS ON SPEECH CHARACTERISTICS FOR ROBUST VOICE ACTIVITY DETECTION

Miquel Espi, Shigeki Miyabe, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama

The University of Tokyo
Graduate School of Information Science and Technology
{espi, miyabe, nishi, onono, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper discusses about effective speech characterization for off-line voice activity detection (VAD), which is an important step prior to speech data mining. Five differents natures of speech are examined; energy, spectral shape, periodicity, phonetic variation, and spectral fluctuation, which is a new a point of view towards VAD. Specific spectral fluctuation patterns of speech have been analysed using multi-stage Harmonic/Percussive Sound Separation algorithm. We compared the performance of the features, and various combinations, to evaluate their robustness in multiple noise environments. The combined approach outperformed the baseline of CENSREC-1-C evaluation framework. The results suggest that the proposed feature extraction approach can improve state of the art VAD methods.

Index Terms— Voice Activity Detection, Speech characterization, Harmonic-Percussive Sound Separation, Delta cepstrum

1. INTRODUCTION

Voice activity detection (VAD) field plays an important role in speech processing including speech recognition, speech enhancement, and speech coding under noisy environments. The method proposed here intends to address off-line voice activity detection, with direct applications in surveillance, and event logging and diarization, among others.

Costs of processing power, storage capacity and network bandwidth are decreasing continually, allowing access to large duration audio archives or spoken documents such as meetings, broadcasts, or personal recordings. Integration of such sources with Human Language Technologies would allow efficient and effective search, and access to the information contained in them. In order to achive such functionality consitently an intermediate stage is necessary to classify and log each the events occuring in the audio source for later specific processing. That is where VAD systems intend to address the task of logging the occurence of speech. From there, speech occurrences can be processed using speech recognition, speaker indexing and identification, depedeing on the type of data required, to enhance the access and search of data along spoken documents. Indeed, VAD already plays an important role as a step prior to speech recognition to peform it in an efficient, speech coding, or telecommunications

VADs are to be used in a wide range of environments with changing background properties, such as stationary an non-stationary noise, burst, or even noiseless signals in the best case scenario. Previously proposed statistical-model based approaches [1], intended to address this issue, providing robust performance in noisy environments. However, although this performance is maintained along stationary noise environments, performance decreases in non-stationary and other kind of noise environments. Other VAD models have been proposed [2, 3, 4] in order to deal better with more kinds of noise environments, but still do not fully solve this issue. Recently proposed scheme based on adaptive integration of multiple speech features and periodic to aperiodic data of the signal [5], has provided great performance in several environments including stationary and non-stationary noise environments.

VAD algorithms consist of two stages: first a set of features is extracted from the signal, and then a classifier outputs a decision depending on the features. Based on this scheme, there are currently three fields of study regarding VAD: those focused on the features, those focused the classifier, and a third group focused on the combination of both features and classifier. This research intends to provide a new point of view on the first case, how speech is characterized and how speech region can be encircled by combining various natures of speech. The accurateness eventually provided by this approach comes with an obvious processing cost which can be accepted when the system is applied to an off-line solution.

The method proposed in this paper has been evaluated using Corpora and Environment for Noisy Speech Recognition-1 Concatenated (CENSREC-1-C) database. CENSREC-1-C is a concatenated speech database specifically intended to validate the performance of VADs. The model proposed in this paper out-performed CENSREC-1-C as well as other VAD standards.

The following sections explain the proposed method with

the following outline: Section 2 intends to consistently define speech from multiple points of view and propose features supporting them, Section 3 describes how the proposal has been evaluated and its results, and Section 4 summarizes the conclusions achieved.

2. NATURE OF SPEECH

To achieve consistent voice activity detection, nature of speech needs to be understood. However, performing this task over different situations is not a simple procedure. Instead, speech requires to be characterized from several points of view, overlapping occasionally, but providing differentiated approaches to cover the speech region as much as possible. This section addresses the matter of speech characterization, and the practical approach using both existing and original features.

Fig. 1 shows a speech sample and each of the proposed features, which offer distinctive performance along different environment circumstances.

2.1. Energy

Time domain analysis is the ground to analyze and locate the boundaries between speech and non-speech activity. Basically, speech is characterized at this level by its energy level, and speech activity can be detected by observation of that fact. Ideally, best results of this feature will come with clean signals, and with the sole presence of speech. On the other hand, it is clearly not robust to signals with low Signal-to-Noise Ratio (SNR), and non-stationary noise.

Fig. 1 shows a sample signal and its Logarithmic frame energy along time for a signal with a 20 dB SNR. In the figure it can be observed how the logarithmic frame energy can be enough to locate speech in the signal in such conditions.

2.2. Spectrum

Frequency domain analysis provides a another point of view to detect the presence of speech activity in the signal, and is widely used for voice activity detection [6]. Frequency domain observation allows to locate speech activity more specifically. Parameterization of frequency domain enables recognition of speech patterns. Compared to the case of time domain observation, frequency domain observation enables the detection of speech in signals where noise dominance is high, but noises do not collide in frequency with speech (e.g. high frequency noises), and enables to differentiate human-voice sounds from other sounds.

Mel-Frequency Cesptral Coefficients (MFCC) are considered as an effective way of observing the frequency domain of of a signal, and are commonly used in the field of speech processing. Since MFCC provide parameterized data, they can be directly used to detect those patterns indicating the presence of speech, even when there are noises in other bands of the frequency spectrum. This explains the applications in speech recognition or speaker identification.

2.3. Periodicity

Periodicity is another intrinsic characteristic of voiced speech. In terms of excitation, voiced speech is characterized, and usually modeled, by having a pitch pulse as its source, as opposed to other kinds of signal. Such source is later modified by vocal filters, which results in what we call speech, but its magnitude can be estimated as it can be observed in Fig. 1. Therefore, the sole presence of a certain pitch in the signal can become enough evidence of voice activity. However, although pitch proves to be a robust feature in many cases, performance may lower down when some kind of noises appear, such as music and other harmonic sounds, with similar properties in terms of periodicity.

2.4. Phonetic variation

Another point of view for speech characterization is to consider speech as a series of phonemes, which change over time. Therefore, the phoneme change rate becomes a meaningful characteristic of speech activity. Dynamic features of speech, first proposed in [7], have been proved to provide good results in order to observe the changes in speech signals, and more specifically to locate the different phonemes present in a signal for speech recognition [8]. In this way, observation of spectral dynamics delimits efficiently the utterances in speech locating the boundaries between speech, and non-speech [9].

Observation of signal dynamics provides fine information even with low SNR signals as observed in Fig. 1, since they reflect the changes in the signal. Still, voice signals which change little over time (e.g. singing voices) might be classified as non-speech. Also the case of non-stationary background noises can cause this feature to incorrectly predict speech activity. We used Delta cepstrum to capture the dynamic properties of the signal [8].

2.5. Spectral Fluctuation

Finally, a fifth way for characterizing speech is spectral fluctuation, first introduced in [10], and introduced here to be used in VAD. Fluctuation should be understood in the spectrum domain and, not just along time, but also along frequency. Speech fluctuates fast along frequency, but slow along time. Such pattern can be isolated and used to identify speech activity, and this can be observed in the spectrum as well as in the fundamental frequency. However, fluctuation observation can be misleading with certain sounds with similar spectral properties to those of speech.

Harmonic Percussive Sound Separation (HPSS) [10] takes advantage of the differences between harmonic sounds

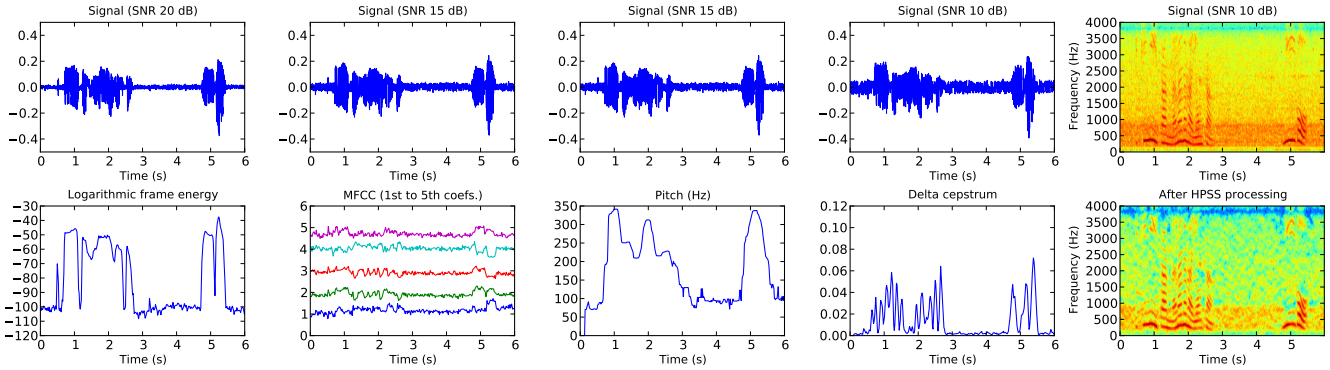

Fig. 1. Observation sample of each of the proposed features (bottom) with their correponding original signal (top). From left to right: Logarithmic frame energy (wave energy), Mel-Frequency Cepstrum Coefficients (Spectrum), Pitch (Periodicity), Delta Cepstrum (Phonetic structure), HPSS spectrogram (spectrum fluctuation)

and percussive sounds in the frequency domain. Basically, it is obtained from the partial differentials of the spectrogram in temporal and frequency direction: harmonic components are smooth in time because they are sustained and for a limited time periodic; on the other hand, percussive components are smooth along frequency, due to their instantaneous and aperiodic properties.

To separate the power spectrogram $W_{i,j}$ of an input signal, where i and j represent frequency and time respectively, into temporally smooth component $H_{h,i}$ and the frequency continuous component $P_{h,i}$ on the spectrogram, L_2 norm of the power spectrogram gradients is examined. That is, $H_{h,i}$ and $P_{h,i}$ are found by minimizing

$$\begin{aligned} J(H,P) &= \frac{1}{2\sigma_H^2}\sum_{h,i}(H_{h,i-1}-H_{h,i})^2 \\ &+ \frac{1}{2\sigma_P^2}\sum_{h,i}(P_{h-1,i}-P_{h,i})^2 \end{aligned} \quad (1)$$

subject to,

$$\begin{aligned} H_{h,i}+P_{h,i} &= W_{h,i} \\ H_{h,i} &\geq 0, \\ P_{h,i} &\geq 0, \end{aligned}$$

where H and P are sets of $H_{h,i}$ and $P_{h,i}$, respectively, and σ_H and σ_P are parameters to control horizontal and vertical smoothness. Minimizing Equation (1) is equivalent to maximum likelihood estimation. Update equations have been omitted and can be found in [10].

Such approach has been successfully used in vocal suppression for music [11], where it was shown that observation of speech feature has a great potential as a decision feature for speech location. The paper [11] explains how voice, due to to its structure, can be isolated by subtracting the long temporally smooth components, and the short frequency continuous components of the signal. This can be achieved by decomposing a mixed signal into both temporally and frequency continuous components with a wide frame first, and then decompose the resulting frequency continuous component again with a short frame.

Voice components of the signal can be found in the resulting temporally smooth component. Fig. 1 shows spectrograms for the original low SNR sample and the result after applying HPSS voice separation process.

Observation of the result after extracting voice through HPSS showed to be robust under low SNR situations, and where there are noises colliding in frequency with speech. This feature has been parameterized in the model by extracting the MFCC of the resulting spectrum.

3. PERFORMANCE EVALUATION

To compare the performance of each of the proposed features separately and in combination, several models have been generated, evaluated and compared. This section describes these models implementation which featured support vector machines, the evaluation framework to evaluate them, and the results obtained.

3.1. Implementation

In order to evaluate the performance of the proposed feature scheme, and since the evaluation of the classifier is out of the scope of this research, the evaluation has been implemented using Support Vector Machines (SVM) [12] since it is a generic classifier widely used. After that, in order to to avoid incorrectly classified frames, a simple post-processing procedure has been applied to the result of the classifier forcing both minimum speech utterance duration (usually between 100 ms and 200 ms), and pause duration (usually between 500 ms and 1000 ms) for frames to be accepted as speech or non-speech, respectively.

3.2. Experiment design

CENSREC-1-C database [13] is a framework maintained by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group, and has been designed to evaluate voice activity detection in noisy environments. The vocabulary of simulated data included in the CENSREC-1-C consist of eleven Japanese digits. The speech data were sampled at 8 kHz, and quantized into 16 bit integers. The details of recording conditions, utterances, and speaking style are same as in CENSREC-1 (AURORA-2J). The simulated speech samples in the database are constructed by concatenating several utterances spoken by one speaker mixed with a certain background noise. The number of utterances in concatenated speech data is nine or ten, and the number of speakers per noise environment is 104 (52 males and 52 females). In this research, simulated noisy samples set contained in CENSREC-1-C have been used. Simulated data consists of mixing speech samples with environmental noise to create samples with different SNR properties (clean, 20, 15, 10, 5, 0, and -5 dB), and four different kinds of noise. A sample spectrogram for each noise can be observed in Fig. 2.

SVM classification model has been trained using the training samples available in CENSREC-1 database. The training set consists of clean data and multi-condition samples with SNRs: 20, 15, 10, and 5. Eight different SVM models have been created to evaluate each of the features performance separately, combining subsets, and the model including all the features. The models tested are: power based model; MFCC based model; pitch based model; delta cepstrum model; HPSS based model; combination of power, MFCC and pitch model referenced in Table 2 as PoMPi; combination of power, MFCC, pitch, and delta cepstrum model; combination of power, MFCC, pitch, and HPSS; and a model including all of the proposed features.

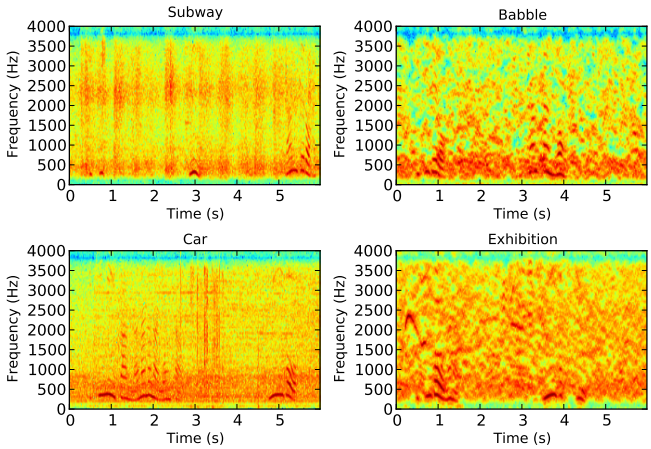

Fig. 2. Spectrogram for each of the environmental noises tested and contained in CENSREC-1-C framework

3.3. Results

3.3.1. Frame based evaluation

Following the evaluation framework guidelines, frame-level performance evaluation of the proposed VAD was based on FRR (False Rejection Rate) and FAR (False Acceptance Rate) and their average values defined by

$$FRR = \frac{N_{FR}}{N_s} \times 100\,[\%] \tag{2}$$

$$FAR = \frac{N_{FA}}{N_{ns}} \times 100\,[\%] \tag{3}$$

where N_s, N_{ns}, N_{FR}, and N_{FA}, are the total number of speech frames, the total number of non-speech frames, the number of speech frames detected as non-speech, and the ones detected as speech frames, respectively.

Fig. 3 presents the frame-level performance results, representing 100-FAR (percentage of correctly accepted frames) and 100-FRR (percentage of correctly rejected frames) for each of the proposed models, with receiver operating characteristic (ROC) curves for each of the noise types. It can be observed how the VAD model featuring all the features out-performs the baseline VAD, however some cases such as HPSS or Delta cepstrum by themselves outperform the full model.

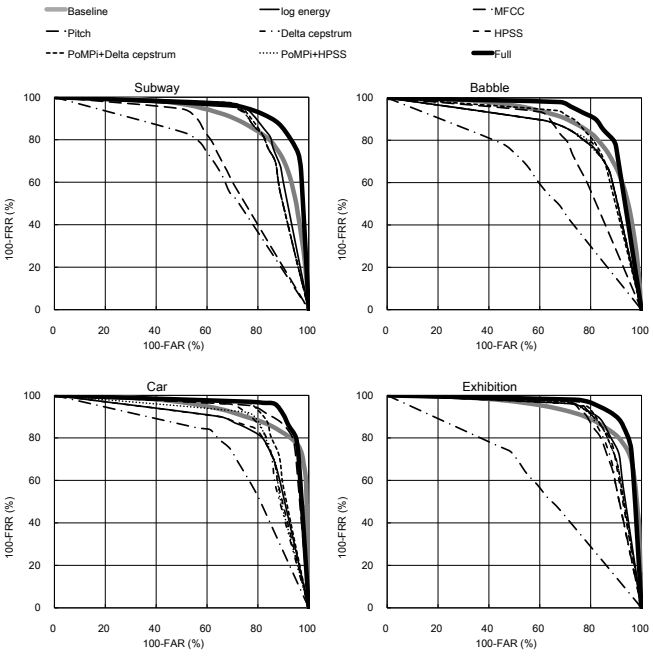

Fig. 3. ROC curves for each of the designed models, along with the curves provided by the CENSREC-1-C baseline

3.3.2. *Utterance based evaluation*

One of the most used applications of VAD is as a front-end for speech recognition systems detecting starting and end points of speech utterances. CENSREC-1-C defines this utterance as each of the connected digit strings. Utterance based performance can be evaluated by obtaining the Correct rate of utterance detection ($Corr$) and the Accuracy of utterance boundary detection (Acc) and their respective average values defined as

$$Corr = \frac{N_c}{N} \times 100\,[\%] \tag{4}$$

$$Acc = \frac{N_c - N_f}{N} \times 100\,[\%] \tag{5}$$

where N, N_c, and N_f, are the total number of speech utterances, the number of correctly detected utterances, and the incorrectly detected ones, respectively.

Tables 1 and 2, summarize the utterance-level performance results by SNR and noise respectively. It can be observed how the proposed VAD out-performs the baseline VAD. The accuracy of the proposed soultion also outperformed other standard VAD algorithms such as ETSI ES 202 050 (63.66%), ETSI ES 202 212 (61.51%) for the same test set [5].

Table 1. Acc and $Corr$ for FULL and CENSREC-1-C's Baseline models for each of the SNR levels contained in the test set and the overall performance for all the tests, and excluding 0, -5 (no included in the training data.

	Acc		*Corr*	
SNR	FULL	Baseline	FULL	Baseline
Clean	98.75	99.83	99.38	99.90
20 dB	96.33	95.25	98.63	96.52
15 dB	91.59	91.33	98.71	94.55
10 dB	84.30	81.87	97.95	90.75
5 dB	71.14	63.59	89.36	83.08
0 dB	23.01	25.04	45.96	57.02
-5 dB	-4.21	-2.60	21.90	36.18
Average	65.84	64.90	78.74	79.71
Except 0, -5 dB	88.42	86.37	96.80	92.96

3.3.3. *Models performance*

By the observation of the frame-based evaluation results in Fig. 3 it can be concluded that, although performances of each of the features by themselves, provide fair results, phonetic variation (Delta Cepstrum) and spectral fluctuation (HPSS) provide a good performance. However, periodicity (pitch) and spectrum (MFCC) observations have low performance, especially in the case of non-stationary noisy environments (Babble and Exhibition) which include also speech in the background noise, making the task more difficult. In the case of models combining features, the difference is wider, and reveals that in separated classifiers were providing information about different sets of frames, therefore, feature combined models "PoMPi+Delta cepstrum" and "PoMPi+HPSS" provide better performance in some enviroments, and especially "Full" improves the performance overall.

Utterance-based results, summarized by Accuarcy and Correctness, reveal the need of a good post-processing engine to enhance the output of the classifier. As explained previously, in this case a basic procedure has been implemented. It is also noticeable the fact that the performance lowers down for lower SNRs 0, and -5 due to the lack of training data for those cases.

4. CONCLUSIONS

SVM has shown to provide fair results, but is still far from being an optimal model for VAD. Its lack of flexibility, and the dependence on the training data, make it a weak model, and more adaptive solutions shall be considered. However, the classifier has been able to show how the proposed features perform along different types of samples, especially in the case of spectral fluctuation observation. Observing the tables and figures shown in the results section, HPSS, as well as the combination of features, provide specifically better performances, which variates over different kinds of noise and SNR levels.

Observing in detail, FAR and especially FRR in the case of the combined model outperform those of the baseline VAD. In conclusion, by accurately defining the nature of speech, we are improving VAD systems from the ground, and therefore, improving the process of diarization.

5. FUTURE SCOPE

Application of these features to more adaptive classifiers, and the inclusion of noise models in order to reduce the dependence on the training set should be the focus in the future. Also, as it has been observed in the results section, some features provided good performance accepting speech, while other features proved to perform better rejecting speech. This means that a combined rejection-acceptance model could provide a good performances on future approaches.

In the long term, automatic transcription of events such as speech from a sound signal and creation of a log with that information is expected to be the next challenge to address. This is, conceive diarization as a complex voice activity detection system integrating not just voice but capable of indexing multiple types of events.

Table 2. *Acc* and *Corr* for all the tested models described in Section 3.2 for each of the environments in the test set

	Acc				*Corr*			
Model	Subway	Babble	Car	Exhibition	Subway	Babble	Car	Exhibition
Power	28.31	35.91	43.30	38.83	57.06	55.16	55.68	58.85
MFCC	53.80	44.24	49.97	57.61	73.11	60.83	62.90	67.82
Pitch	-19.72	15.88	-20.98	-14.93	28.60	18.60	46.42	14.13
Δcepstrum	12.57	27.63	54.52	57.78	52.21	65.12	83.88	**87.73**
HPSS	47.67	40.57	46.71	47.88	74.31	62.72	70.10	76.88
PoMPi	49.49	46.04	51.13	56.84	71.44	63.45	66.37	74.80
PoMPi+Δcepstrum	47.47	51.18	64.27	63.84	74.84	67.73	77.84	80.33
PoMPi+HPSS	49.74	44.06	54.28	52.39	72.80	63.18	70.25	76.51
Full	**65.19**	**63.89**	**74.23**	68.80	**77.91**	72.36	**87.76**	80.92
Baseline	58.31	62.22	73.40	**69.03**	76.12	**75.91**	85.77	85.84

6. REFERENCES

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," in *IEEE Signal Processing Letters*, January 1999, vol. 6, pp. 1–3.

[2] J. Ramirez, J. M. Gorriz, and J. C. Segura, *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness*, 2007.

[3] J. Ramirez and J. C. Segura, "Statistical voice activity detection using a multiple observation likelihood ratio test," in *IEEE Signal Processing Letters*, October 2005, vol. 12, pp. 689–692.

[4] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," in *IEEE Transactions on Speech Audio Processing*, September 2003, vol. 11, pp. 498–505.

[5] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *Proceedings of ICASSP*, April 2008, pp. 4441–4444.

[6] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Proceedings of Interspeech 2005*, 2005, pp. 369–372.

[7] S. Sagayama and F. Itakura, "On individuality in a dynamic measure of speech," in *Proceedings of ASJ Spring Conference, 3-2-7*, June 1979, pp. 589–590.

[8] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Febrary 1986, vol. 34, pp. 52–59.

[9] O. Mizuno, S. Takahashi, and S. Sagayama, "Speech discrimination using dynamic and static spectral features," in *Proceedings of ASJ Fall Conference*, September 1995, pp. 107–108.

[10] N. Ono, K. Miyamoto, and S. Sagayama J. Le Roux, H. Kameoka, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proceedings of EUSIPCO*, 2008.

[11] H. Tachibana, N. Ono, and S. Sagayama, "Vocal sound suppression in monaural audio signals by multi-stage harmonic-percussive sound separation(hpss)," in *Proceedings of ASJ Spring Meeting*, March 2009, pp. 853–854.

[12] Z. Songfeng et al., "Unsupervised clustering based reduced support vector machines," in *Proceedings of ICASSP*, April 2003, vol. 2, pp. 821–825.

[13] N. Kitaoka et al., "Development of vad evaluation framework censrec-1-c and investigation of relationship between vad and speech recognition performance," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2007, pp. 607–612.