Using Spectral Fluctuation of Speech in multi-feature HMM-based voice activity detection

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

Graduate School of Information Science and Technology, University of Tokyo, Japan

{espi, miyabe, nishi, onono, sagayama}@hil.t.u-tokyo.ac.jp

Abstract

Observation of speech spectrum leads to the fact that speech has a specific spectral fluctuation pattern both along time and frequency. In this paper, we integrate the usage of this nature in a multi-feature approach for voice activity detection. The effect of separating such specific spectral fluctuation using multi-stage HPSS (Harmonic-Percussive Sound Separation) has been analyzed over conventional features in voice activity detection, reducing frame-wise detection error by up to 78%, depending on the SNR conditions and noise type. The multi-feature approach has been tested using Hidden Markov Models to model the features stream as a sequence, which has out-performed standard and similar VAD proposals in utterance-based tests intended for automatic speech recognition.

1. Introduction

In speech processing, voice activity detection (VAD) plays an important role as a front-end in multiple fields 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, acoustic event detection, or diarization.

Speech signals have a specific spectral fluctuation behavior regarded as intermediate between harmonic and percussive sounds, fluctuating slow along time, and fast along frequency. Here, we exploit this fact by separating such intermediate components using multi-stage HPSS (Harmonic-Percussive Sound Separation), to track such a characteristic spectral behavior applied to robust VAD.

The decrease of costs derived from processing power, storage capacity per network bandwidth has caused a change of paradigm when it comes to access to large duration audio archives 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. That is where an intermediate stage is necessary to classify and log each of the events occurring in the audio source for later specific processing. VAD does this by logging the occurrence of speech.

Real environments have changing background properties, such as stationary and non-stationary noise, burst, and robustness is an important feature of VAD. 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] 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 [4], has provided great performance in several environments including stationary and non-stationary noise environments.

In this paper we present an in-depth analysis of the effects of speech spectral separation in conventional features and how this affects the performance of a VAD. The algorithm has been developed using HMM, since the features in use also contain information as a sequence. The proposed approach in this paper has been evaluated using Corpora and Environment for Noisy Speech Recognition-1 Concatenated (CENSREC-1-C) database [5]. CENSREC-1-C is a concatenated speech database specifically intended to validate the performance of VADs. The model proposed in this paper out-performed baseline as well as similar VAD proposals.

Section 2 introduces speech spectral fluctuation and the extraction method. Sections 3 describes the feature scheme and evaluates the effects of separating speech spectral fluctuation, and finally the VAD is evaluated in section 4.

2. Spectral fluctuation of speech

Speech spectral fluctuation occurs slow along time, and fast along frequency. This is because speech consists of pitches and slowly changing phonemes, defining a distinguishing behavior different of other sounds such as music, or noises stationary or non-stationary. Isolating such components of the signal enables to track speech activity from a different point of view to what conventional features allow. Speech signals can be characterized as intermediate between harmonic and percussive sounds. Such intermediate signal components can be extracted by first discarding the long components smooth in time, and then discarding the short components smooth in frequency. The remaining signal would be such intermediate component with similar level of spectral fluctuation to that of speech.

In order to do this, we used HPSS [6], which takes advantage of the differences between harmonic sounds and percussive sounds in the frequency domain. Basically, it is obtained from the partial differentials of the spectrogram in temporal and frequency directions: harmonic components are smooth along time because they are sustained and for a limited time periodic; and percussive components are smooth along frequency, due to their instantaneous and aperiodic properties. This is, separation of a power spectrogram S in two spectrograms: H containing the components smooth along time and P containing the components smooth along frequency. More detailed description of the algorithm can be found in [6].

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 $H_{i,j}$ and the frequency contin-

28 – 31 August 2011, Florence, Italy

uous component $P_{i,j}$ on the spectrogram, norm of the power spectrogram gradients is examined.

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

where H and P are sets of $H_{i,j}$ and $P_{i,j}$, respectively, and σ_H and σ_P are parameters to control horizontal and vertical smoothness. Note that ultimately adding H and P will result in obtaining the original spectrum. Minimizing Eq. (1) is obtained by an iterative update rule with a similar manner to expectation-maximization algorithm [6].

By conducting two sequential HPSSs with long and short analysis windows, referred to as multi-stage HPSS [7], we can obtain the desired components with speech-like spectral fluctuation. First, we apply HPSS with a wide analysis window (512 ms) to discard components regarded as stationary for relatively long term, which will be in the resulting smooth-in-time component referenced as H. Then, we apply HPSS again to the component P with a short window (32 ms) to discard transient components which will remain in the smooth-in-frequency component referenced as P. This results in the smooth-in-time (the component referenced as H) component containing most of the speech. The flow can be observed in Fig. 1.

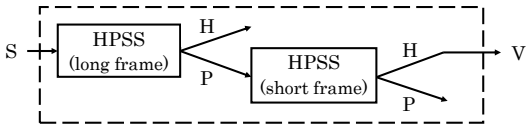

Figure 1: Flow of Multi-stage HPSS for speech.

3. Feature Analysis and Extraction

3.1. Feature extraction

In order to track spectral fluctuation, we have analyzed the effects of the Multi-Stage HPSS processing in conventional features regarding VAD. We have considered the following properties of speech for its characterization:

- Energy: time domain analysis of amplitude, which indicates the presence of speech in noiseless scenarios.
- Spectrum: analysis of similarity/dissimilarity with speech spectral envelope. Robust to noises present in spectral regions other than speech, but weak with noises in the same spectral regions of speech.
- Periodicity: pitch analysis. Speech has a periodic pulse as source, generating a periodicity in the signal. However, is weak with signals that include music or any other harmonic sounds with similar properties to those of speech.
- Phonetic dynamics: analysis of variation in spectral envelope. Speech can be considered as a series of phonemes which change over time, and this can be observed in the spectral dynamics delimiting efficiently the utterances [8].

We also included the deltas of the features with several window lengths [9] to observe their temporal information information and analyze its usefulness. We obtained the deltas using the following equation:

$$\Delta^K c_t = \frac{\sum_{k=1}^{K} k\cdot(c_{t+k}-c_{t-k})}{2\cdot\sum_{k=1}^{K}k^2} \quad (2)$$

where c_i and K are the value of feature c at the i-th frame, and the length of the delta (number of frames), respectively. The delta lengths we applied are:

- Immediate: usually referred as delta, with a window length of 1 frame (16 ms).
- Short term: window length of 3 frames (48 ms).
- Long term: window length of 8 frames (128 ms).

3.2. Feature performance analysis

In order to compare the effects of using the Multi-Stage HPSS we build a likelihood ratio test based VAD which compares the likelihoods of a 'speech' and a 'non-speech' GMM (16 gaussians) for each frame. We built the corresponding ROC (Receiver Operating Characteristic) curves, and measured the size left below the False Alarm and False Rejection errors. Therefore, the smaller the size of this region, the closer to an ideal VAD. The measure is explained in Fig. 2.

We obtained this error region size measure for each of the conventional features presented: before Multi-Stage HPSS (referenced as 'original' in the figures), and after Multi-Stage HPSS.

The GMMs for 'speech' and 'non-speech' have been trained using the 'train' dataset contained in CENSREC-1 [5], a Japanese version of AURORA-2, which contains samples in 4 different environment types (Subway, Babble, Car, and Exhibition), 4 SNR levels (5, 10, 15, and 20 dB), and 110 different speakers (55 female and 55 male). For the test we used the 'testa' dataset contained in CENSREC-1-C, which contains the same environments as in the training set, 6 SNR levels (-5, 0, 5, 10, 15, and 20 dB) and 104 different speakers (52 female and 52 male). We used the entire training data set to train the GMMs for each of the feature so there is no matched training neither by noise type, nor by SNR level.

The results of measuring the error region size are shown in Fig. 3 and Fig. 4, by feature, noise type, and SNR level. There we can observe how the error size for energy reduces almost half in low SNR, although the improvement is not so in high SNRs. This is specially significant in Subway and Car environments, where the only non-stationary sound in the signal is the targeted speech itself. On the other side, in Babble and Exhibition environments, although there is an improvement, is not so significant due to the existence of untargeted speech in

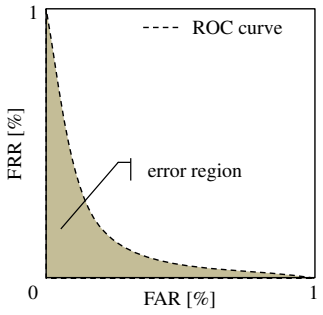

Figure 2: Representation of the region of the ROC curve, which is taken to measure the error size marked in grey.

the background noise, as well as other sounds. In the case of pitch the result is the opposite, speech spectral fluctuation separation affects negatively to lower SNRs. We can conclude that there is no observable improvement on using the spectral fluctuation separation in none of the noise types, or SNR levels when using pitch feature. This might be because although pitch is part of the spectral fluctuation, separating such components includes some loss of information, and therefore, pitch tracking becomes more difficult. In the case of the MFCC characterizing the spectrum, the improvement is again notable in general in all the noise types in the tests, especially in the case of higher SNRs as it can be observed in Fig. 4. In the case of the dynamic features, we tested immediate, short-term, and long-term dynamic features for all of the features in analysis. However, Fig. 3 and Fig. 4 include only most relevant ones for analysis purposes. Here we can see that the improvement is still there for the features which are improved in their instantaneous case (energy, spectrum), and provide no distinctive improvement in the case of periodicity.

4. VAD evaluation

Using the information we obtained by analyzing the performance of each of the features separately, we built 3 different VAD based on the performance of the features separately, to analyze and compare the effects of including the features after the speech spectral fluctuation separation. The three feature extraction schemes are as follows:

- Basic: include the 4 properties introduced above without further processing. These are: energy (Logarithmic Frame Energy), Spectrum (MFCC), Periodicity, (pitch), and Spectral dynamic features (immediate, short and long term deltas of the MFCCs).
- SFSEP: which includes the features described in 'Basic' and the most salient features after separation speech specific spectral fluctuation with Multi-Stage HPSS processing: Energy, Spectrum, and Spectrum long-term dynamics.
- ExSFSEP: extends 'SFSEP' scheme with immediate energy dynamics after Multi-Stage HPSS for speech, and the immediate pitch dynamics of the original signal.

The VAD has been built using a 2 states HMM as shown in Fig. 5, being state 0 for 'non-speech' and state 1 for 'speech', and using initially $\{0.45, 0.55, 0.90, 0.10\}$ as state transition matrix. The emission probabilities for states H_0 and H_1 have been obtained from 16 mixtures GMMs.

4.1. Experimental setup

To test the performance of the resulting VAD we conducted an open test. GMMs to define the emission probabilities have been trained using the 'train' dataset from CENSREC-1, as in the feature analysis. The test has been done over the 'remote' dataset contained in CENSREC-1-C. This dataset is characterized by being recorded live in the corresponding environment. It includes 2 types of noises: 'restaurant' and 'street', the first one being highly non-stationary; both with high (about 53.4 dBA in 'restaurant' and 58.4 dBA in 'street') and low (about 69.7 dBA in 'restaurant' and 69.2 dBA in 'street') SNR conditions. The dataset features 10 different speakers (5 female and 5 male).

4.2. VAD results

Voice activity detection most common application is to be used as a front end for a speech recognition system, detecting the endpoints of speech utterances. In the case of CENSREC-1-C, utterances are formed by connected digit strings. To evaluate

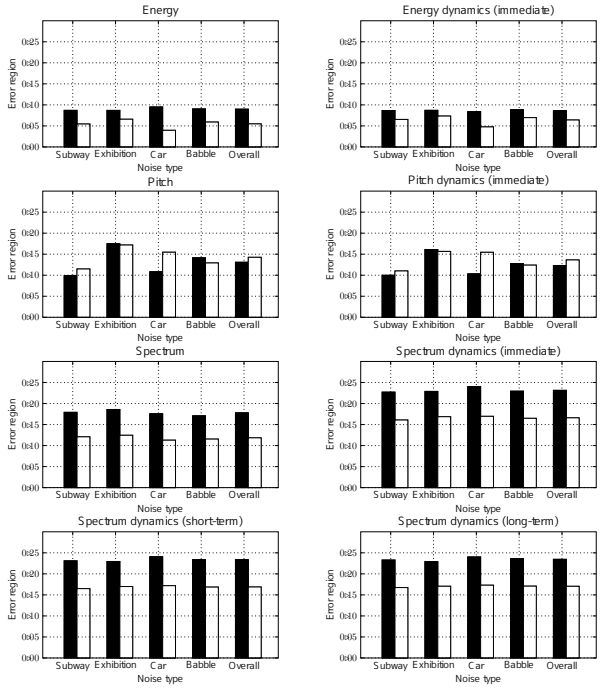

Figure 3: Error region sizes comparative between the original feature (black) and after applying Multi-Stage HPSS for speech (white), separated by noise type.

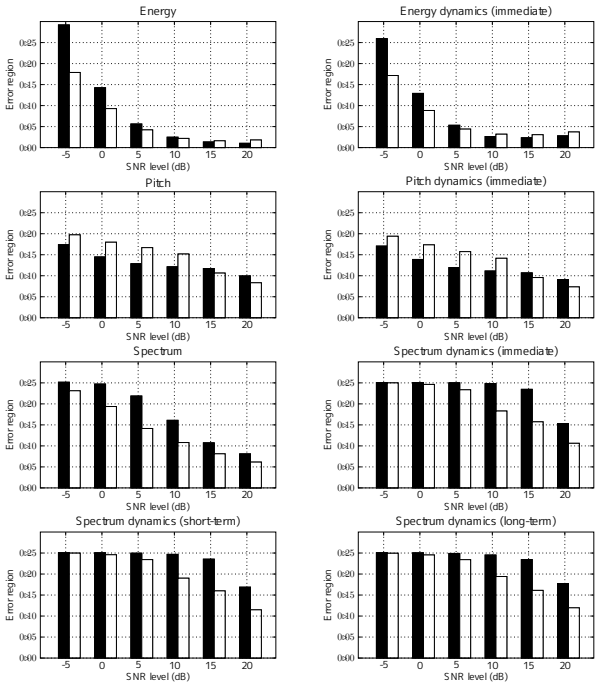

Figure 4: Error region sizes comparative between the original feature (black) and after applying Multi-Stage HPSS for speech (white), separated by SNR level.

Table 1: VAD results (%)

	$Corr$					Acc				
	Restaurant		Street			Restaurant		Street		
Dataset	High	Low	High	Low	Overall	High	Low	High	Low	Overall
Baseline	74.20%	56.62%	39.42%	41.45%	52.90%	21.45%	-43.48%	-15.65%	-33.91%	-17.90%
Sohn	72.75%	57.10%	97.39%	78.55%	76.45%	45.51%	-6.38%	94.49%	57.39%	47.75%
PAR	71.72%	57.10%	87.25%	80.58%	73.91%	24.35%	-6.67%	64.35%	54.49%	34.13%
Basic	64.34%	51.20%	71.21%	66.92%	63.41%	22.61%	-10.34%	62.10%	47.23%	30.36%
SFSEP	**73.63%**	**62.35%**	**89.12%**	**78.34%**	**75.83%**	**35.82%**	**15.20%**	**76.27%**	**64.30%**	**47.14%**
ExSFSEP	**76.75%**	**63.02%**	**92.44%**	**79.64%**	**77.96%**	**41.13%**	**16.78%**	**80.51%**	**66.36%**	**51.19%**

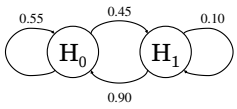

Figure 5: HMM states representation and the transition probabilities.

the utterance- level VAD performance, two evaluation measures have been used: the rate of correctly detected utterance boundaries ($Corr$), and the accuracy in utterance boundary detection (Acc), which can be defined as follows:

$$Corr = \frac{N_c}{N} \times 100\,[\%] \quad (3)$$

$$Acc = \frac{N_c - N_f}{N} \times 100\,[\%] \quad (4)$$

where N is the total number of speech utterances, N_c is the number of correctly detected utterances, and N_f is the number of incorrectly detected utterances. $Corr$ assesses how many speech utterances can be detected by VAD algorithms, but Acc also takes into account the number of over-detected utterances.

If a detected speech segment is shorter than an actual speech segment, phoneme information at the beginning or the end of the utterance is missing, resulting in a speech recognition error. On the other hand, even if a detected speech segment has additional non-speech ranges before and after the speech utterance, it would not cause significant damage to speech recognition performance. This measures assume that, if a detected speech segment includes all speech intervals of an utterance without overlapping either the preceding or succeeding speech utterance, it can be a candidate for correct detection.

The results show how the proposed approach, denoted as 'SFSEP' and 'ExSFSEP' in Table 1, outperforms significantly the results of 'Basic', which is the same except for lacking the Multi-HPSS processed features. From this, we can conclude the fact that including such features provides a significant improvement. This improvement is not so notable between 'SFSEP' and 'ExSFSEP' which only includes the Energy and Pitch immediate dynamic features (Δ^1). The table also shows the results of the Baseline included in the CENSREC-1-C framework and Sohn VAD, as well as the results obtained by Periodic to Aperiodic Ratio (PAR) algorithm [10], which is based on estimating the periodic to aperiodic ratio of the signal for end-point detection. Compared to this methods, the proposed approach also provides better results specially in the case of the 'restaurant' environment, where there is more presence of non-stationary noise.

5. Conclusions

This paper presented a noise robust VAD technique based on the integration of multiple speech features with spectral fluctuation enhanced features. A set of features and the effect of speech specific spectral fluctuation separation has been analyzed, resulting in a set of features which improve their performance. The evaluation results show that the proposed method improves VAD accuracy compared with other standard and recently proposed VADs. In the future, we plan to include new features to enable better tracking of spectral fluctuation in the model.

6. References

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

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

[3] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Transactions on Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, September 2003.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] T. Fukuda, O.Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 834–844, October 2010.

[10] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in *Proceedings of SAPA*, Sept. 2006, pp. 65–70.