# Using Spectral Fluctuation of Speech in Robust Voice Activity Detection [*]

○ Miquel Espi, Shigeki Miyabe, Takuya Nishimoto, Nobutaka Ono, Shigeki Sagayama
Graduate School of Information Science and Technology, The University of Tokyo

## 1 Introduction

Speech signals have a specific spectral fluctuation behaviour 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 characeristic spectral behaviour applied to robust Voice Activity Detection (VAD). VAD is considered here as a mark-up process to use speech processing technologies in an efficient and effective manner, with special attention to environments with changing background properties. Recent developments based on adaptive integration of multiple speech features and periodic to aperiodic data of the signal [1] have shown robust performance including non-stationary noises.

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. In [2], we analyzed a new proposal of speech characterization adding spectral fluctuation information. Here, we implement a VAD based on GMM likelihood ratio test making use this proposal and and evaluate its performance.

## 2 Feature Extraction

### 2.1 Speech Characterization

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 convetional features allow.

Additionally, we also considered the following convetional features in the system:

- Energy: time domain analysis of amplitude.
- Spectrum: analysis of similarity/dissimilarity with speech spectral envelope.
- Periodicity: pitch analysis.
- Phonetic dynamics: analysis of variation in spectral envelope.

The features associated with each of the natures are summarized in Table 1.

### 2.2 Spectral Fluctuation

Speech signals can be regarded 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.

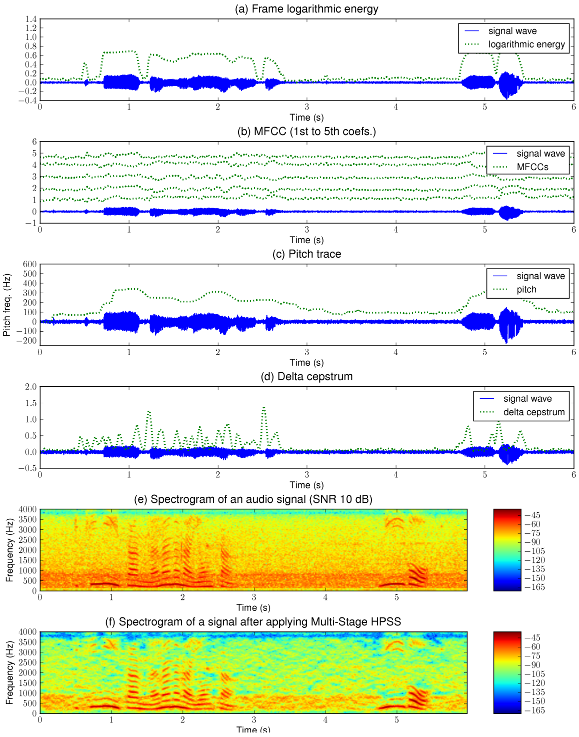

Fig. 1 Features characterizing speech: (a) frame energy, (b) MFCCs, (c) Pitch (d) Delta Cepstrum, (e) Original Spectrogram, and (f) Spectrogram plot after Multi-Stage HPSS.

In order to do this we used HPSS [5], 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 [5].

By conducting two different HPSSs with long and short analysis windows, referred to as multi-stage HPSS [6], we can obtain the desired components

[*] スペクトルの動的特徴を用いた頑健な発話区間検出の検討、エスピ・ミケル、宮部滋樹、西本卓也、小野順貴、嵯峨山茂樹（東京大学 情報理工）

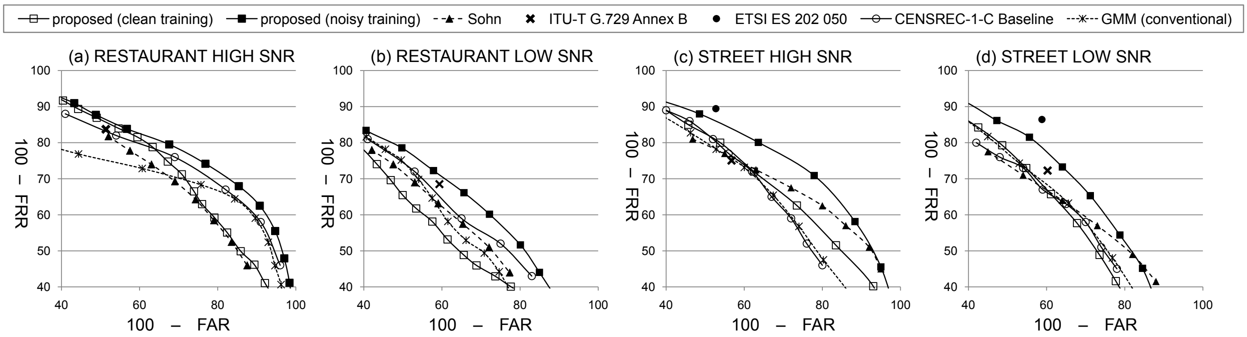

Fig. 2 ROC curves for each environment. ETSI ES 202 050's performance is out of the graph's range in RESTAURANT's low SNR (FAR = 87.14%, FRR = 0.51%), and high SNR (FAR = 84.3%, FRR = 1.7%).

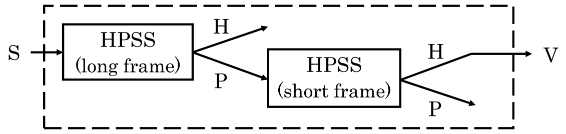

Fig. 3 Multi-Stage HPSS flow with a spectrogram $S$ as input, and $V$ as the output containing most of the speech.

Table 1 Speech natures and features in use.

| **Speech Nature** | **Feature** |
|---|---|
| Energy | Frame logarithmic energy |
| Spectrum | MFCCs |
| Peridicity | Pitch |
| Phonetic dynamics | Delta MFCCs |
| Spectral fluctuation | Multi-Stage HPSS (MFCCs) |

with speech-like spectral fluctuation. First, to discard components regarded stationary for relatively long term, we apply HPSS with a wide analysis window. Then, we apply HPSS again with a short window to discard transient components. Taking the resulting smooth-along-time component as the spectrogram where most of the speech remains. This process is illustrated in Fig. 3.

Observation of the results after extracting the voice component through multi-stage HPSS showed to be robust under low SNR conditions, and where there are noises colliding in frequency with speech. Fig. 1 (e) and (f) illustrates this fact with spectrograms for a low SNR speech sample and the result after applying multi-stage HPSS. This feature has been parameterized in the model by extracting the MFCC of the resulting spectrum.

## 3 Experimental results

The VAD classifier has been implemented using two Gaussian Mixture Models (GMM) for: speech, and noise. Each of the GMMs are fed with the presented features, and trained to fit a 128 gaussians mixture. Test has ben conducted using the dataset 'remote' of CENSREC-1-C, which consists of real environment recordings, in two different sites: RESTAURANT and STREET. Both environments present strong non-stationary background noise.

Three models for speech have been generated: clean speech triaining, noisy speech training, noisy training without the spectral fluctuation feature (referenced in the figures as 'clean training', 'noisy training', 'GMM conventional', respectively). We also compared the results with the following VADs: Sohn VAD [7], ITU-U G.729 Annex B, and ETSI ES 202 050.

The proposed VAD using noisy speech training outperformed reference VADs like Sohn's VAD and ITU-T G.729 Annex B, and also outperformed ETSI ES 202 050 performance in the subset test of RESTAURANT which includes speech in the background, among other types of non-stationary noise. Also note that the performance of the clean speech training model outperforms some of the comparison VADs (Sohn and G.729 Annex B) in high SNR scenarios.

## 4 Conclusions and future scope

As it can be observed on the results section, the proposed approach to speech characterization including spectral fluctuation, along with the likelihood ratio test classifier proves to be effective providing better performance than conventional VAD methods, and also robust to non-stationary noises.

In the future, a more VAD-oriented model needs to be developed, in terms of integration of hang-over schemes in the model by using HMM. Also the development of better noise models such as the use of longer frames for noise detection should be considered.

## References

[1] Fujimoto *et al.*, , Proc. ICASSP, 4441-4444, 2008.
[2] Espi *et al.*, Proc. IEEE SLT, 139-144, 2010.
[3] Kristjansson *et al.*, Proc. INTERSPEECH, 369-372, 2005.
[4] 水野他, 春季研究発表会（秋）, 3-2-1, 107-108, 1995.
[5] Ono *et al.*, Proc. EUSIPCO, 139-144, 2008.
[6] 橘他, 春季研究発表会（春）, 2-8-8, 853-854, 2009.
[7] Sohn *et al.*, IEEE Signal Processing Letters, 6, 1-3, 1999.