Exploiting spectral fluctuation in multi-feature HMM-based voice activity detection *

◎ Miquel Espi†, Daisuke Saito†, Nobutaka Ono‡, Shigeki Sagayama†
†Graduate School of Information Science and Technology, The University of Tokyo
‡National Institute of Informatics

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 continues the work started in [1], and intends to address off-line voice activity detection, with direct applications in fields such as surveillance, acoustic event detection, or diarization.

Traditionally, time domain features such as sample energy or zero-crossing rate have been widely used. However, while they provide a good response in high SNR environments, their VAD performance degrades increasingly as the SNR level lowers down, as well as with certain non-stationary. Taking advantage of the fact that most of the speech spectral energy is located in the lower frequencies, features accounting the power in a band-limited region including short-term and long-term temporal dynamics have also been introduced for VAD, along with long used periodicity assumed to be more insensitive to non-periodical background noises [2, 3].

Such features, although increasing the robustness against environmental noise still degrade significantly in the presence of non-stationary noise, and more specifically in the presence on periodical noises, and unvoiced phones. Speech signals have a specific spectral fluctuation behavior regarded as intermediate between harmonic and percussive sounds, fluctuating slow along time, and fast along frequency.

2 Spectral fluctuation of speech

Speech consists of two main components: temporally steady parts of vowels and voiced consonants, and the fluctuations intrinsic to vowels, voiced consonants, stops, fricatives, and affricates. The underlying concept behind this two components is spectral smoothness, and based on this paradigm one can find three different types of signals: signals smooth in time corresponding to stationary sounds, signals smooth in frequency corresponding to transient noises, and a third groups between those two of signals which fluctuate in time and frequency. Speech is in this third group, but requires a more specific characterization to differentiate from other fluctuating sounds. Speech fluctuates fast along frequency, but slow along time.

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.

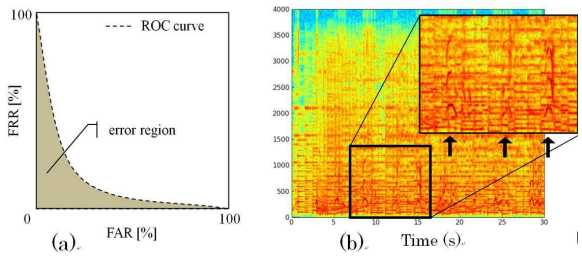

Fig. 1 (a) ROC error region area representation, (b) Spectrogram containing speech and music.

Here, we are able to locate such specific spectral fluctuation in time and frequency of speech by using 2-stage HPSS [4], which was originally introduced for sound separation in music, and is based on HPSS [5]. Its distinctive advantage relies on the ability to decompose a signal into three components based on their spectro-temporal smoothness, or spectral fluctuation, thus representing speech more accurately than periodicity itself when it comes to VAD, because it takes into account not only harmonic components, but all fluctuating components. And with this, not relying on components of speech that are not always present, and avoiding being influenced by the contents that got decomposed in the other decomposed components, with special focus in non-stationary and periodical noises.

3 Features analysis

Being able to separate the components of a signal that fluctuates as speech from the ones that do not, we want to obtain features that characterize this property of speech regarding VAD in a robust way. One would think that spectrum related features such as spectral power, and specially MFCC (non-linearly distributed along the spectrum, in the way speech is), are the most suited features for this matter. However, have analyzed also other conventional features, to see how their performance with regards to VAD increases or decreases when combined with 2-stage HPSS spectral separation, and evaluate their ability to characterize spectral fluctuation or not in VAD.

3.1 ROC error region area evaluation

In order to evaluate each of the features, we have built Likelihood Ratio Test (LRT) VADs for each of them and constructed their associated Receiver-Operating Characteristics (ROC) curves by shifting the LRT threshold. ROC curves report frame-wise analysis, and describe completely the trade-off between rigidness and flexibility of the decision rule, as they are obtained by plotting False Acceptance (FAR) and False Rejection (FRR) error rates for each of the thresholds, accounting for the error introduced when a frame that is not speech is detected as speech,

* スペクトル変動を用いた複数特徴と HMM に基づく音声区間検出、エスピ・ミケル、齋藤大輔（東大情報理工）、小野順貴（国立情報学研究所）、嵯峨山茂樹（東大情報理工）

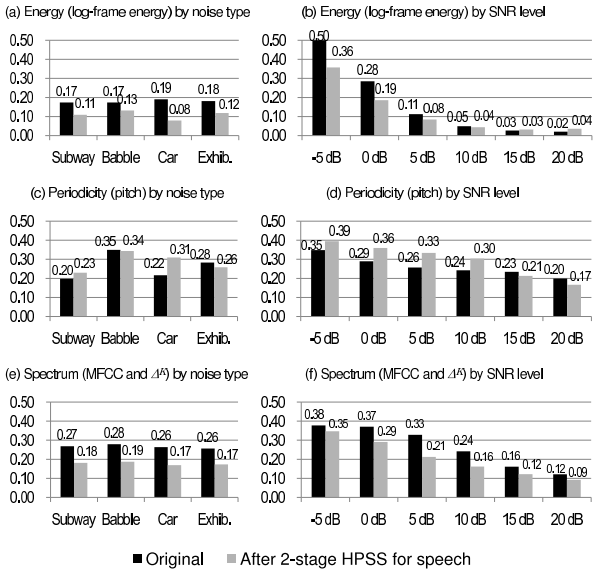

Fig. 2 Error Region Area sizes by noise and SNR level for the analyzed features.

and the error introduced when a frame that is speech is detected as non speech. In this way, one can say that the smallest the region resting under the ROC curve and the FAR axis (Error Region Area, ERA), the better the performance is, as Figure 1 indicates.

3.2 Results

Likelihoods for speech and non-speech were modeled using GMMs with CENSREC-1's noisy training dataset (5, 10, 15, and 20 dB, 4 different noise types, and 52 subjects). CENSREC-1-C's "testa" dataset has been used for the ROC test without hang-over post-processing, for original and 2-stage HPSS separated signals.

The results reveal the fact that energy property based feature (log-frame energy) improves its performance after applying the spectral fluctuation separation, which is related with an improvement in the SNR of the signal, special in the noisiest environments. Such performance improvement can also be observed in the MFCC (spectrum property) analysis, which can be explained by the fact that mostly components of speech remain in the signal, and therefore, the MFCCs of the remaining signal are closer to clean speech than the original. On th other side pitch tracking becomes less robust for VAD, which can be associated with the fact that there is information loss during the decomposition.

4 Evaluation and results

4.1 VAD Classifier

In order to benefit from the robustness of characterizing speech spectral fluctuation and maintain the robustness in commonly supported environments we integrated the spectral fluctuation related features in a multi-feature scheme, summarized in table 1.

Decision rule, and hang-over scheme have been integrated by using a duration explicit Hidden Markov Model (HMM) describing the minimum durations for pause and speech, dealing with short-term energy drops or unexpected peaks by constraining the occurrence of an utterance or a pause upon a certain duration. Explicit duration HMM models this by having in the HMM as much clone states for non-speech as frames in the minimum pause duration with a looped state at the end, and as much clone states for speech as frames in the minimum utterance duration with a looped state at the end as well.

Speech property	Features
Energy	log-frame energy
Periodicity	Pitch
Spectrum	MFCC
Temporal dynamics	Δ^1, Δ^3, Δ^8MFCC
Spectral fluctuation	MFCC, Δ^3, Δ^8MFCC

Table 1 Features extracted

Noise	SNR	Proposed	PAR	Kristj.
Airport	0 dB	22.3	25.8	35.9
	5 dB	15.2	18.1	24.5
	10 dB	11.6	14.4	17.6
Subway	0 dB	22.9	23.1	55.7
	5 dB	15.7	16.2	49.2
	10 dB	12.2	13.7	42.4
Restaurant	0	22.1	26.6	51.5
	5 dB	16.4	18.0	41.9
	10 dB	12.9	14.2	32.4
Street	0 dB	21.8	26.4	47.0
	5 dB	14.2	17.5	35.4
	10 dB	12.5	14.3	25.6

Table 2 VAD performances in terms of Equal Error (when FAR and FRR are the same).

4.2 Results

During the evaluation procedure, frames were set to 32 ms with half overlap, and minimum pause and speech durations were set to 500 ms and 300 ms, respectively. States emissions have been modeled by training a 128 Gaussians GMM for speech and non-speech using the same training dataset of CENSREC-1.

As it can be observed in the results (Table 2), the proposed VAD out-performs similar approaches in several environments, including those which also have periodical noises in it as 'airport' and 'restaurant' in noisy multiple SNR conditions ranging from 0 to 10 dB. However, it has to be noted that PAR [2] does not require training, which could be considered as a robustness property as well.

5 Conclusion

In this research we have analyzed the effect that speech spectral fluctuation separation has in conventional features with regards to their performance in VAD by comparing their ROC curve error region size in GMM-LRT based scheme. Also, a explicit duration HMM based VAD integrating all the features has been compared to related approaches with relevant results.

References

[1] Espi *et al*, 春季研究発表会（春）, 1-5-15, 2011.
[2] Ishizuka *et al.*, Proc. SAPA, 65-70, 2006.
[3] Kristjansson *et al.*, Proc. Interspeech, 2005.
[4] 橘他, 春季研究発表会（春）, 2-8-8, 853-854, 2009.
[5] Ono *et al.*, Proc. EUSIPCO, 139-144, 2008.