論文 / 著書情報
Article / Book Information

Title	A Novel Voice Activity Detection Algorithm using Long-Term Spectral Flatness Measure
Authors	Yanna Ma, AKINORI NISHIHARA
Citation	2013 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, , ,
Pub. date	2013,
Note	この論文はNCSP13で発表したものである。 This thesis was published at NCSP13.

A Novel Voice Activity Detection Algorithm using Long-Term Spectral Flatness Measure

Yanna Ma and Akinori Nishihara

Tokyo Institute of Technology
2-12-1-W9-108,Ookayama, Meguro-ku, Tokyo, Japan
Phone:+81-3-5734-3232
E-mail: mayanna@nh.cradle.titech.ac.jp

Abstract

This paper proposed a novel and robust voice activity detection (VAD) algorithm utilizing long-term spectral flatness measure (LSFM) which is capable of working at very low signal-to-noise ratios (less than 10 dB). The proposed algorithm was evaluated under eleven types of noises from NOISEX-92 and five types of SNR in core TIMIT TEST corpus. Comparisons with three modern standardized algorithms (ETSI AMR option 1&2 and ITU-T G.729 Annex B) demonstrate that our proposed LSFM-based VAD scheme achieved highest speech hit rate among the four algorithms while keeping good hit rate balance between speech hit rate and non-speech hit rate.

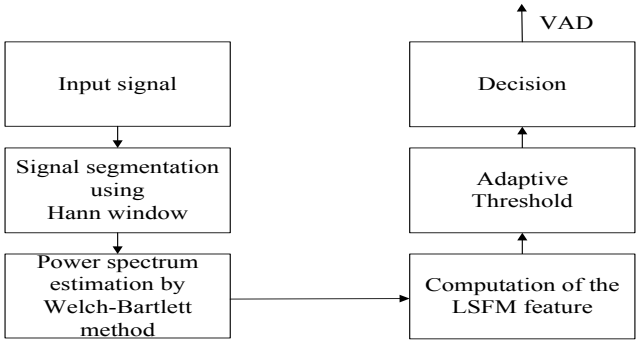

Figure 1: Block diagram of the proposed LSFM-based VAD algorithm

1. Introduction

Voice Activity Detection (VAD) is the method to discriminate speech segments from the input noisy speech. It is an integral part to many speech and audio processing applications and is widely used within the field of speech communication for achieving high coding efficiency and low-bit rate transmission. Examples include noise reduction for digital hearing aid devices [1] and mobile communication services [2].

A typical VAD system consists of two core parts: feature extraction and thresholding. In order to achieve a good detection, the features chosen have to show a discriminative variation between speech and non-speech segments. Researchers have proposed a variety of features exploiting different properties of speech and noise to achieve a better VAD performance. Most features work sufficiently well in stationary noises and high SNR cases such as 10dB and above. However, VAD algorithms based on particular feature or specific set of features are still far from efficient especially when it is operating in adverse acoustic conditions. Therefore, the VAD algorithm which can work efficiently in low SNRs still remains challenging and requires the design of further robust features and algorithms.

The structure of this paper is arranged as follows. Section 2 presents the proposed long-term spectral flatness measure (LSFM)-based VAD algorithm including the choice for proper parameters and adaptive threshold. The experiment conditions and evaluation standard are dicussed in Section 3. The experiment results are provided in Section 4. Finally, a conclusion of this work and the discussion are given in Section 5.

2. The proposed LSFM-based VAD algorithm

Spectral flatness is a measure of the width and uniformity of the power spectrum. The LSFM feature is computed using the spectrum of the last R frames of the input signal $x(n)$ with respect of the current frame of interest. To expand the dynamic range, the spectrum flatness is measured on a logarithmic scale, ranging from 0 to minus infinity. It is calculated by

$$L_x(m) = \sum_k 10log_{10}\frac{GM(m,w_k)}{AM(m,w_k)} \quad (1)$$

where the geometric mean $GM(m,w_k)$ and arithmetic mean $AM(m,w_k)$ of the power spectrum is calculated as

$$GM(m,w_k) = \sqrt[R]{\prod_{m-R+1}^{m} S(n,w_k)} \quad (2)$$

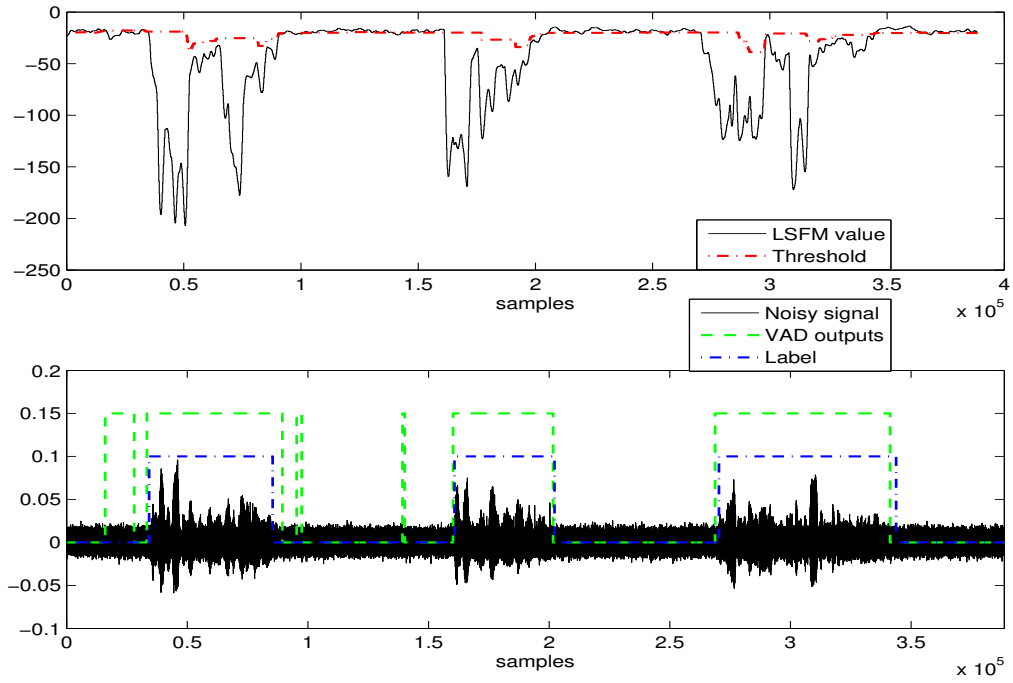

Figure 2: Illustrative example of the adaptive threshold and the VAD output, white noise, SNR=0dB. The upper figure shows the LSFM value and the corresponding adaptive threshold for each frame, the lower figure shows the VAD output decisions and the ground truth namely "Label"

$$AM(m, w_k) = \frac{1}{R} \sum_{n=m-R+1}^{m} S(n, w_k) \tag{3}$$

The short-time spectrum $S(n, w_k)$ used in this research is estimated by Welch-Bartlett method which averages the spectral estimates of M consecutive frames. The expressions are

$$S(n, w_k) = \frac{1}{M} \sum_{p=n-M+1}^{n} |X(p, w_k)|^2 \tag{4}$$

$$X(p, w_k) = \sum_{l=(p-1)N_{sh}+1}^{N_w+(p-1)N_{sh}} w(l-(p-1)N_{sh}-1)x(l)e^{-jw_k l} \tag{5}$$

where $X(p, w_k)$ is the short-time Fourier Transform coefficient at frequency w_k ($w_k \in [500, 4000]$Hz) of the p^{th} frame. $w(i)$ is the short-time Hann window, where $i \in [0, N_w)$. N_w is the frame length and N_{sh} is the frame shift. In this paper, we use 20ms length (N_w=20ms) and 10ms shift (N_{sh}=10ms).

A block diagram of the proposed LSFM-based VAD algorithm is shown in Fig. 1. The algorithm can be described as follows. The input speech signal is decomposed into overlapped frames (20ms length, 10ms shift) by the Hann window. And the spectrum of the segmented signal is estimated by Welch-Bartlett method. At the m^{th} frame, the LSFM feature $L_x(m)$ is computed using the last R frames. The decision about whether there contain speech in the last R frames is made through the comparison with an adaptive threshold.

An illustrative example of the adaptive and the VAD output is shown in Fig. 2. A high spectral flatness indicates that the spectrum has a similar amount of power in all spectral bands, this would sound similar to white noise, and the graph of the spectrum would appear relatively flat and smooth. A low spectral flatness indicates that the spectral power is concentrated in a relatively small number of bands, this would typically sound like a mixture of sine waves which is likely to be speech signal. Hence, the spectral flatness measure is a good feature for VAD.

2.1 Selection of M and R

In Eq. (2-4), M and R are parameters used for computing the LSFM feature L_x. We want to choose appropriate M and R so that the detection error is minimized. The misclassification errors for all combinations of M and R are computed for M=5, 10, 20, 30, 40 and R=5, 10, 20, 30, 40 over eleven types of noise at 0dB. We observed that the best choice of the

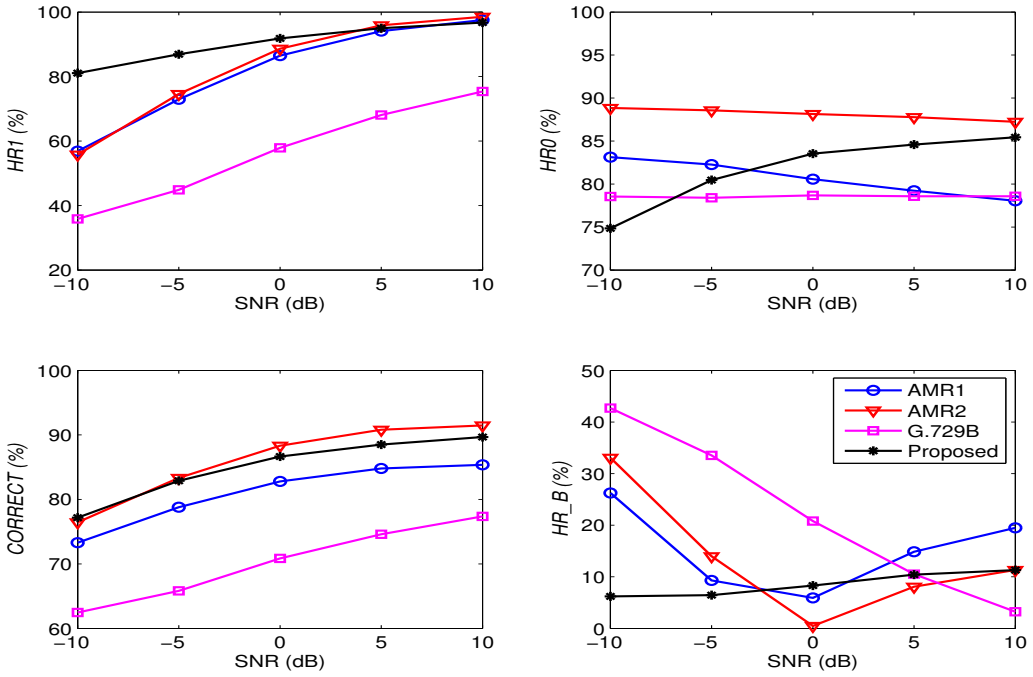

Figure 3: Accuracy rate ($HR1$, $HR0$, $CORRECT$ and HR_B) comparisons of four VAD schemes (AMR1, AMR2, G.729B and the proposed LSFM) averaged over eleven kinds of noise for five SNRs (-10dB, -5dB, 0dB, 5dB and 10dB)

combination of M and R is (10,30) for most noises tested. This fixed combination is then adopted for the following tests.

2.2 Adaptive Threshold

A fixed threshold would lose its efficiency when facing varying acoustic environments. Therefore, it is more suitable to design an adaptive threshold [3]. From Eq. (2-4) we can conclude that $(R + M - 1)$ frames (0.39s for fixed $R = 30$ and $M = 10$) information is needed to acquire the first LSFM feature value. In our implementations, the initial 1.39s of the test signal $x(n)$ is always assumed to be non-speech. From this 1.39s of $x(n)$, 100 realizations of L_N can be collected and saved to ψ_{INL}. The threshold is then initialized to be:

$$THR_{INL} = min(\psi_{INL}). \quad (6)$$

To update the threshold at m^{th} frame, we used two buffers ψ_{S+N} and ψ_N. ψ_{S+N} stored the LSFM measures of the last 100 long window ending at m^{th} frame which were decided as containing speech; similarly, ψ_N stored the LSFM measures of the last 100 long window ending at m^{th} frame which were decided as including non-speech information only. The adaptive threshold for the m^{th} frame is then updated as:

$$THR(m) = \lambda * min(\psi_{S+N}) + (1 - \lambda) * max(\psi_N) \quad (7)$$

where λ is the parameter of the convex combination. We experimentally found that $\lambda = 0.55$ results in maximum accuracy rate in VAD decisions over the TIMIT training set [4].

3. Experiment Conditions and Evaluation method

3.1 Experiment Conditions

In this paper, experiments are carried out on the core TIMIT TEST set [4] consisting of 24 individual speakers (16 male, 8 female) of eight different dialects, each speaking 10 phonetically balanced English sentences. Eleven kinds of noise from NOISEX-92 database [5] have been artificially added to the speech at SNRs of -10, -5, 0, 5 and 10dB.

3.2 Evaluation method

Detection performance was assessed in terms of the speech hit rate ($HR1$) and non-speech hit rate ($HR0$) defined as the fraction of all actual speech frames or non-speech frames that are correctly detected as speech frames or non-speech frames, respectively[6]:

$$HR1 = \frac{N_{1,1}}{N_1^{ref}} \qquad HR0 = \frac{N_{0,0}}{N_0^{ref}} \quad (8)$$

where N_1^{ref} and N_0^{ref} are the number of real speech and non-speech frames in the whole database,respectively, while $N_{1,1}$ and $N_{0,0}$ are the number of speech and non-speech frames correctly classified. The overall accuracy rate ($CORRECT$) and hit rate balance (HR_B) is then defined as:

$$CORRECT = \frac{N_{1,1} + N_{0,0}}{N_1^{ref} + N_0^{ref}} \tag{9}$$

$$HR\_B = |HR1 - HR0| \tag{10}$$

4. Experimental Results

As shown in Fig. 3, the proposed LSFM-based VAD is compared with three modern standardized algorithms (ETSI AMR option 1&2 [7] and ITU-T G.729 Annex B [8]) in terms of accuracy rate. Note that the results in Fig. 3 are averaged values for all eleven noises from NOISEX-92 database. From Fig. 3 we can conclude that G.729B suffers poor $CORRECT$ and $HR1$ with the increasing noise level while keeps high $HR0$ for the whole range of SNRs (78.56% on average) . AMRa performs much better than G.729B for both $CORRECT$ and $HR1$ while suffering degradation of $HR0$ when the SNR level is increased. AMR2 improves considerably over AMR1 in $CORRECT$ mainly because of the high $HR0$ over all SNRs (88.11% on average) while yielding similar $HR1$ with AMR1. Our proposed LSFM-based VAD yields the best $HR1$ from the beginning while keeps continuous increasing for $HR0$ when the SNR increases. Furthermore, our proposed method achieved the best balance (HR_B) between $HR1$ and$HR0$.

5. Conclusion and Discussion

We proposed a robust long-term spectral flatness measure (LSFM)-based VAD scheme which can discriminate speech segments from non-speech segments effectively. Experiments were done on core TIMIT TEST set for eleven kinds of noise from NOISEX-92 database across five different SNRs range from -10dB to 10dB. The performance of our proposed method is compared with three standards, namely G.729B, AMR1 and AMR2. The results are analyzed by accuracy rate. Through extensive experiments, we show that the proposed VAD scheme achieved the best compromise when compared to the three representative VADs analyzed. It acquired high $HR1$ while keeping the best HR_B.

The test database used in the implementations was created to simulate typical conversational speech by inserting 2s silence before and after each utterance from core TIMIT TEST corpus, so that the ratio of speech to non-speech was almost 40% to 60%. While this simulates a conversational speech statistically, this is not very realistic in terms of randomness of pauses, hesitations, etc [9]. Furthermore, it should be noted that according to different applications, the requirement for HR_B maybe variant. For example, HR1 is a crucial factor for speech coding while high HR0 rate is necessary for most speech recognition oriented systems.

References

[1] K.Itoh and M. Mizushima: Environmental noise reduction based on speech/non-speech identification for hearing aids, Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on, Vol. 1, pp. 419–422, Apr. 1997.

[2] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd: The voice activity detector for the Pan-European digital cellular mobile telephone service, Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on, Vol. 1, pp. 369–372, May 1989.

[3] R.V. Prasad, A. Sangwan, H.S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav: Comparison of voice activity detection algorithms for VoIP, Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02), pp. 530–535, 2002.

[4] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue: TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium, Philadelphia, 1993.

[5] A. Varga and H.J. Steeneken: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on recognition systems, Speech Communication, Vol. 12, No. 3, pp. 247–251,1993.

[6] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, and A. Rubio: Efficient voice activity detection algorithms using long-term speech information, Speech Communication, Vol. 42, No. 34, pp. 271–287, 2004.

[7] Digital Cellular Telecommunications System (Phase 2+), Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels, General Description, 1999.

[8] ITU, Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code - Excited Linear Prediction. Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommend. V.7.0, International Telecommunication Union, 1996.

[9] P. Ghosh, A. Tsiartas, and S. Narayanan: Robust Voice Activity Detection Using Long-Term Signal Variability, Audio, Speech, and Language Processing, IEEE Transactions on, Vol. 19, No. 3, pp. 600–613, Mar. 2011.