T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	A Robust Voice Activity Detection based on Long-term Mean
著者(和文)	馬 艷娜, AKHTARMUHAMMAD TAHIR, 西原 明法
Authors(English)	Yanna Ma, muhammad-tahir akhtar, Akinori Nishihara
出典(和文)	第26回 信号処理シンポジウム講演論文集, , ,
Citation(English)	26th SIP SYMPOSIUM, , ,
発行日 / Pub. date	2011, 11
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2011 Institute of Electronics, Information and Communication Engineers.

Powered by T2R2 (Science Tokyo Research Repository)

A Robust Voice Activity Detection based on Long-term Mean Power Spectrum

Yanna Ma*, M. T. Akhtar ** and Akinori Nishihara*

* Department of Communications and Integrated Systems, Tokyo Institute of Technology
e-mail: mayanna@nh.cradle.titech.ac.jp

**The Center for Frontier Science and Engineering, The University of Electro-Communications

Abstract: We propose a novel VAD method based on long term mean power spectrum (LMPS). It is estimated by Welch-Bartlett method through averaging periodogram estimate of spectrum of the input speech signal. The proposed scheme was evaluated under eleven types of noises and five types of SNR. Comparisons with three standards demonstrates that the accuracy of the LMPS based VAD scheme averaged over all noises and all SNRs is 16.9% better than that obtained by the best available commercial ETSI AMR VAD option2. The analysis of the experiment results also demonstrate that our proposed LMPS based VAD scheme can achieve high accuracy rate while keep good balance of the speech hit rate (SpHR) and silence hit rate (SiHR) which is very important for VAD.

1: Introduction

Voice activity detection (VAD) is the activity to discriminate speech segments from the input noisy speech. It is widely used within the field of speech communication for achieving high coding efficiency and low-bit rate transmission. Researchers have proposed a variety of features exploiting the different properties of speech and noise to achieve a better VAD accuracy. However, time-domain parameters exhibit dependencies on estimates of background noise and changes of thresholds. Spectral shape, on the other hand, tends to lose its effectiveness when the signal noise ratio (SNR) is low. This would lead to a poorer discrimination between the speech and non-speech regions. Therefore, the VAD problem still remains challenging and requires the design of further robust features and algorithms.

2: Proposed Long-term Mean Power Spectrum Algorithm

The block diagram of the proposed LMPS based VAD is shown in Fig. 1. The input speech signal is first windowed by the Hann window. And the mean power spectrum is estimated by Welch-Bartlett method [1]. Finally the VAD decision results are acquired by comparing the LMPS values with the adaptive threshold.

2.1: Long-term Mean Power Spectrum

The power spectrum $S_x(n, w_k)$ is estimated by Welch-Bartlett method which averages the spectral estimates of M consecutive frames. The expression is as follows:

$$S_x(n, w_k) = \frac{1}{M}\sum_{p=n-M+1}^{n} \left| \sum_{l=(p-1)N_{sh}+1}^{N_w+(p-1)N_{sh}} w_{Hn}(l - (p-1)N_{sh} - 1)x(l)e^{-jw_k l} \right|^2 \quad (1)$$

where $N_w = 320$ and $N_{sh} = 160$ are window length and window shift, respectively.

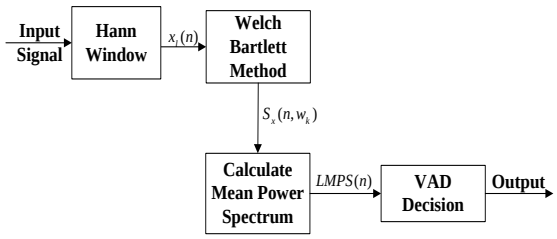

Figure 1: Block diagram of the LMPS algorithm

The long-term mean power spectrum (LMPS) is an asymptotically unbiased and consistent estimate of the power spectral density of the observed signal. It is computed using the last R frames of the input signal $x(n)$ with respect to the current frame of interest. The LMPS, $\mathrm{LMPS}(m)$, at the mth frame across all the chosen frequency is then computed as follows:

$$\mathrm{LMPS}(m) = \sum_{k=1}^{K} \mathrm{MS}(m, w_k) \qquad (2)$$

where the mean value of Welch-Bartlett estimated spectrum is:

$$\mathrm{MS}(m, w_k) = \frac{1}{R} \sum_{n=m-R+1}^{m} S_x(n, w_k) \qquad (3)$$

From experiments we observed that the variance of the LMPS value of the noise signal is half that of the Welch-Bartlett estimated power spectrum, it demonstrate that the fluctuation of the power spectrum of noise is reduced by averaging which makes the discrimination of speech and non-speech segments easier.

2.2: Parameter Selection and Threshold

In the previous description, the detail of how to calculate the LMPS is given. However, there are still several parameters and adaptive threshold waiting to be decided.

2.2.1: Selection $\{w_k\}_{k=1}^{K}$, M and R

It is known that 500 Hz to 4 kHz frequency range is crucial for speech intelligibility [2]. Hence we decide to choose w_k in this range. The exact values of $\{w_k\}_{k=1}^{K}$ are determined by the sampling frequency F_s and the order N_{DFT} of Discrete Fourier Transform (DFT), used to compute the spectral estimate of the observed signal. Thus, $K = N_{DFT}(\frac{4000-500}{F_s})$. In this research, we choose $N_{DFT} = 512$ and $F_s = 16$ kHz which yield $K = 112$. $\{w_k\}_{k=1}^{K}$ are then uniformly distributed between 500Hz and 4kHz.

M and R are parameters used to estimate power spectrum and its long-term mean value, respectively. By experiments, we chose M=20 and R=30 which performs best for most noise types across the all the SNRs.

2.2.2: Adaptive threshold

To update the threshold, we use two buffers $\Gamma_{S+N}(n)$ and $\Gamma_N(n)$. $\Gamma_{S+N}(n)$ stored the LMPS measures of the last 200 frames, which were decided as containing speech; similarly, $\Gamma_N(n)$ stored the LMPS measures of the last 200 frames, which were decided as including noise only. Then the adaptive threshold is given using the following equation:

$$\mathrm{THR}(n) = \lambda * \min(\Gamma_{S+N}(n)) + (1-\lambda) * \max(\Gamma_N(n)) \qquad (4)$$

Where λ is the parameter of the convex combination. By experiments we found that $\lambda = 0.55$ results in maximum accuracy in VAD decisions over the TIMIT test corpus [3].

The initial 2 seconds of the input signal are assumed to be noise only, from which we can get 200 realizations of Γ_N. Let μ_N and σ_N be the mean and standard deviation of these 200 realizations of Γ_N. Since the average LMPS value of noisy speech is larger than that of noise only, the threshold should be more than the mean LMPS value of noise signal, we initialize the threshold to be $\mathrm{THR} = \mu_N + \sigma_N$.

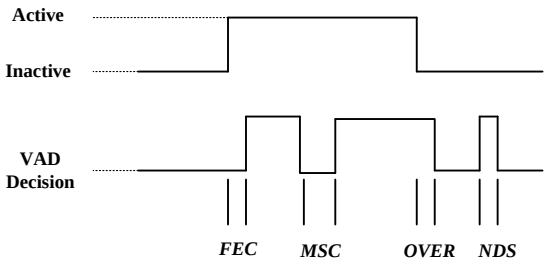

Figure 2: Illustrative example of error rate

3: Evaluation and Results Analysis

To evaluate the performance of our proposed method, some experiments are carried out on part of TIMIT test corpus consisting of 40 individual speakers (20 female, 20 male) of eight different dialects of English. Two-second silence was added before and after each utterance to simulate a typical conversational speech in which the ratio of speech to non-speech was almost 40% to 60%. Then, noise of eleven categories taken from NOISEX-92 database [4] were added at five different SNR levels (-10dB, -5dB, 0dB, 5dB, 10dB) to all test utterances The total test dataset thus obtained consists of 4.75 minutes of noisy speech of which 61.3% was only noise. The decisions obtained by the proposed VAD algorithm are compared against known reference labels which are got from the phonetic transcription of the TIMIT corpus. The beginning and end locations of the speech portions of the silence-padded TIMIT sentences were computed using the start time and end time of the sentence obtained from the TIMIT phonetic transcriptions. The final VAD decisions of proposed method were computed for every 10-ms interval.

3.1: Evaluation Schemes

Two evaluation schemes are adopted for the fair evaluation of the proposed method. The first one is named to be accuracy rate which includes the CORRECT rate, Speech Hit Rate (SpHR) and Silence Hit Rate (SiHR). The other is the testing strategy proposed by Freeman *et al.* [5] As Fig. 2 shows, the error rate includes Front End Clipping (FEC), Mid Speech Clipping (MSC), OVER and Noise Detected as Speech (NDS).

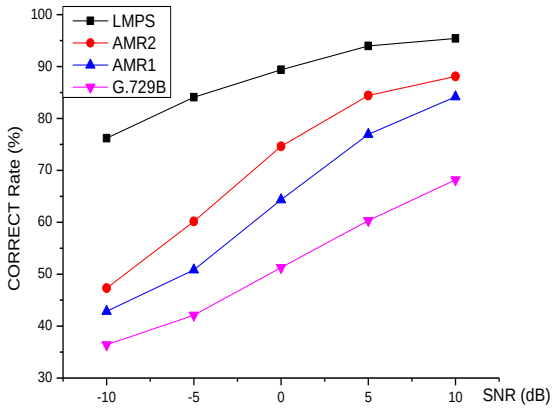

Figure 3: CORRECT Rate comparison for all noises

3.2: Results Comparison

Experiments are done on our proposed method and the three standards. The implementations of G.729B and ETSI AMR VADs (AMR1 and AMR2) were taken from [6] and [7], respectively. Fig. 3 shows the VAD CORRECT Rate comparison in different SNRs between proposed method and the three standards in the aspect of CORRECT Rate. From Fig. 3, we conclude the proposed LMPS VAD algorithm performs the best among the four algorithms across all the test SNRs and especially at as low SNR as -10dB, it can still achieve an average of 76.2% correct rate while other three are only below 50%. As the SNR increases, the correct rate increases steadily for all four schemes. When the SNR reaches 10dB, the LMPS VAD can get as high as 95.4% while AMR2, the best among the three standards, can achieve 88.1%.

In total, the average correct rate across all SNRs is 87.8% for LMPS and 70.9% for AMR2. So the correct accuracy is 16.9% higher than AMR2 and28.9% and 23.9% higher when SNRs are -10dB

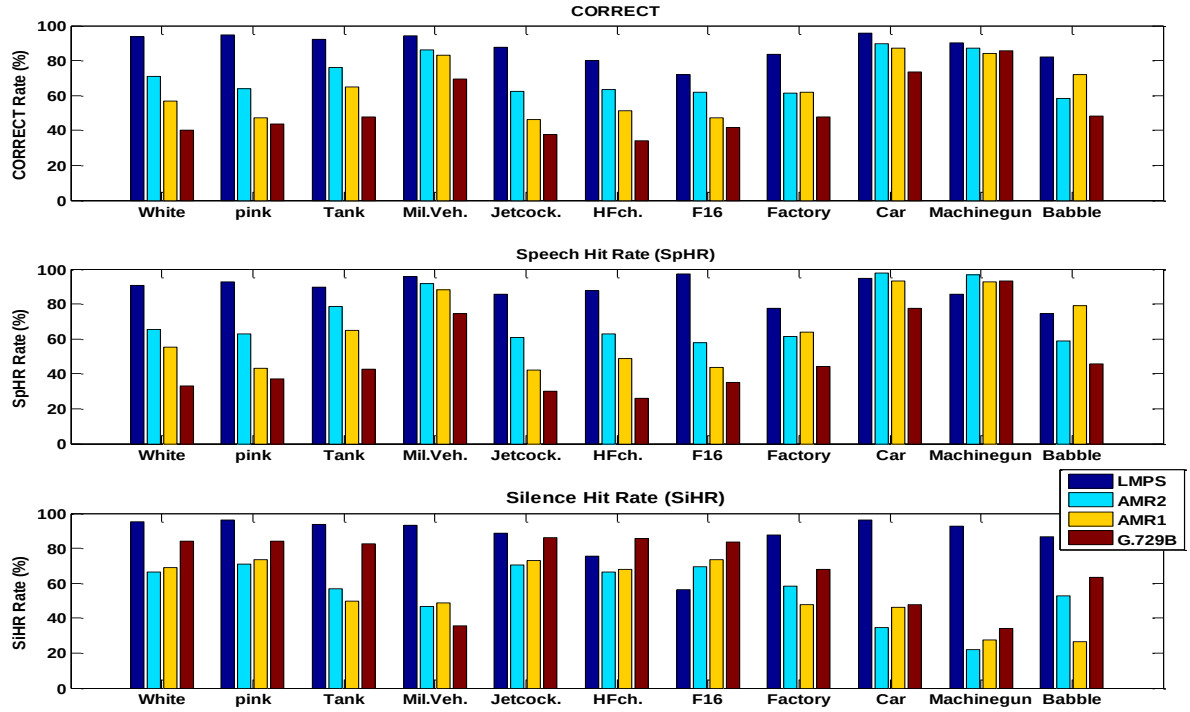

Figure 4: Accuracy Rate comparisons of four VAD schemes for eleven kinds of noises

and -5dB, respectively.

3.2.1: Accuracy Rate Comparison

Only the CORRECT rate is not enough to evaluate a VAD scheme comprehensively. We adopt the Accuracy Rate and Error Rate evaluation standards. Fig. 4 is the comparisons of the four VAD schemes in eleven kinds of noises. For the CORRECT rate we can see from the figure that LMPS method performs best in all eleven kinds of noises. Moreover, the SpHR of LMPS outperforms all others in all the noise conditions except Car Interior and Machinegun. Especially for the stationary noise such as white and pink, LMPS achieves around 30% higher than AMR2 which is the best among all other three. As for SiHR, we can clearly see that the LMPS algorithm also performs better in most noise types with F16 noises to be the exception.

3.2.2: Error Rate Comparison

Fig. 5 shows the Error Rate comparisons of four VAD schemes for eleven kinds of noise, we observe that FEC and MSC errors of LMPS VAD methods are the smallest for most noise types among the four VAD schemes while the OVER and NDS errors of LMPS is obviously larger for noises like HFchannel and F16. However, high value of OVER and NDS implies that noise frames are detected as speech. Depending on the applications, such errors can be tolerable. On the contrary, high FEC and MSC is harmful for any application since it implies that speech frames are decided as noise frame. Generally speaking, the proposed LMPS method outperforms the other three standards even at the aspect of Error Rate.

4: Conclusion

We proposed a robust long-term mean power spectrum based VAD scheme. The experiments are done on TIMIT corpus for eleven kinds of noise from NOISEX92 database across five different SNRs range from -10dB to 10dB. The performance of our proposed method is compared with the three

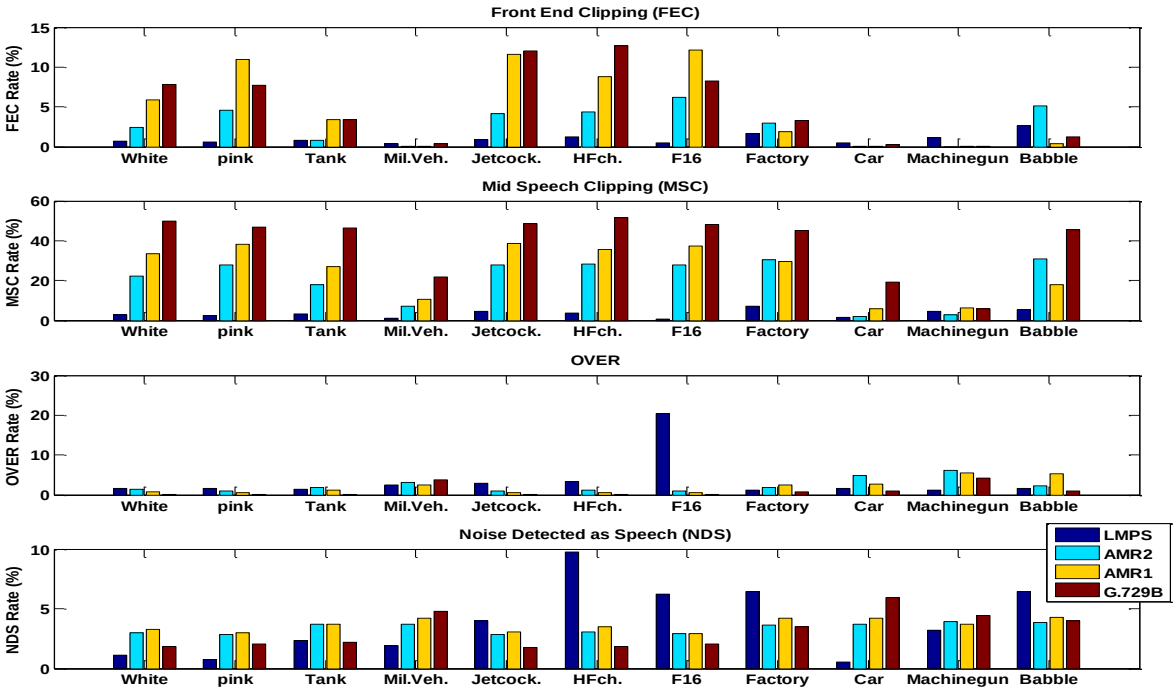

Figure 5 Error Rate comparisons of four VAD schemes for eleven kinds of noises

standards, G.729B, AMR1 and AMR2. The results are analyzed through two evaluation schemes, one is accuracy rate include CORRECT, SpHR and SiHR while the other is error rate which include FEC, MSC, OVER and NDS. Through extensive experiments, we show that the accuracy of the LMPS based VAD performs best among the four VAD schemes by giving 87.8 % average CORRECT rate, 89.7% SpHR and 87.5% SiHR while AMR2, which performs the best among the tree reference standards, gives a 70.9 % average CORRECT rate, 72.3% SpHR and 55.9% SiHR.

The test database was created to simulate typical conversational speech by inserting silence to utterances from TIMIT database, 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. Furthermore, the proposed method lost its efficiency under the noise of HFchannel and F16 because of its high LMPS value compares with other noise types.

References

[1] D. G. Manolakis, V.K. Ingle, and S.M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing.* Norwell, MA: Artech House, pp227-232,Apr.30, 2005.

[2] Bie D. A., and Hansen C. H., "Engineering Noise Control: Theory and Practice", *Edition: 3, illustrated*, Published by Taylor & Francis, 2003, Sec. 4.6 and pp 150.

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

[4] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition:II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*., vol. 12, no.3, pp.247-251, Jul. 1993.

[5] D. K. Freeman, C. B. Southcott, I. Boyd, and G. Cosier, “The voice activity detector for pan-European digital cellular mobile telephone service,” in *proc. IEEE ICASSP,* Glasgow, U,K., 1989, vol. 1, pp.369-372.

[6] ITU, Coding of Speech at 8 Kbit/s using Conjugate Structure Algebraic Code-Excited Linear Prediction. Annex I: Reference Fixed-Point Implementation for Integrating G.729 CS-ACELP Speech Coding Main Body with Annexes B, D and E, Int. Telecommun. Union, 2000.

[7] *Digital Cellular Telecommunication System (Phase 2+); Adaptive Multi Rate (AMR) Speech; ANSI-C code for AMR Speech Codec*, 1998.