Towards Improving Statistical Model Based Voice Activity Detection

Ming Tu, Xiang Xie, Yishan Jiao

School of Information and Electronics, Beijing Institute of Technology

tuming90@gmail.com, xiexiang@bit.edu.cn, alicechiao13@gmail.com

Abstract

Statistical model based voice activity detection (VAD) is commonly used in various speech related research and applications. In this paper, we try to improve the performance of statistical model based VAD via new feature extraction method. Our main innovation focuses on that we apply Mel-frequency subband coefficients with power-law nonlinearity as feature for statistical model based VAD instead of Discrete Fourier Transform (DFT) coefficients. This proposed feature is then modeled by Gaussian distribution. Performances of this method are comprehensively compared with existing methods. Meanwhile we also test power-law nonlinearity on existing methods. Experimental results prove that with proposed subband coefficients the performance of statistical model based VAD could be improved a lot. Power-law nonlinearity on DFT coefficients could also bring some improvement.

Index Terms: voice activity detection, Mel-frequency subband coefficients, power-law nonlinearity, likelihood ratio test

10.21437/Interspeech.2014-370

1. Introduction

Voice activity detection (VAD) has always been an important module of automatic speech recognition (ASR) system, speech codec and some speech enhancement systems which need to track background noise [1]. For applications sensitive to VAD system, it becomes more demanding that VAD's accuracy should be as high as possible. Also, with speech technologies being applied to more complex scenarios, VAD system should maintain efficiency in these situations, especially when background noise is nonstationary.

Statistical model based VAD method, first proposed in [2], is a commonly used VAD algorithm because of its speaker and noise independence. Inspired by the speech enhancement algorithm in [3], statistical model based VAD method uses discrete fourier transform (DFT) coefficients as feature and likelihood ratio test (LRT) as decision-making strategy. Since then, several variants have been put forward to get better VAD performance. Instead of the Hidden Markov Model (HMM) based hang-over scheme in [2], [4] uses multiple observation likelihood ratio test (MOLRT) to smooth VAD output. It is proved that both the performance of VAD and ASR system using this method are improved. [5] considers the harmonic structure of speech into the calculation of log likelihood ratio (LLR) and also improves VAD performance compared with [4]. Experiments on ASR system verify its efficiency. However, if the background noise also contains harmonic structure, such as babble and music noise, this method brings little improvement or even degrades the performance. In [6], the authors propose to combine a double threshold energy detection algorithm with MOLRT based VAD. Mel-frequency cepstral coefficients (MFCC) are used as feature and Gaussian Mixture Model (GMM) is employed for modeling and calculating multiple observation log likelihood ratio (MOLLR). Though superiority over previous methods is obvious, it needs to train different models for different background noise and a speech model, which limits its application.

In this paper, we study VAD along the direction of MOLRT based statistical model methods and aim to bring some improvement by introducing new features. We propose to use novel Mel-frequency subband coefficients instead of DFT coefficients as feature. Usually, Mel-frequency subband coefficients could be obtained during the calculation of MFCC, just after the step of logarithm operation and before the step of discrete-cosine transform (DCT). We make two points of change to this calculation procedure to make it better than DFT coefficients used in existing statistical model based VAD methods. First we remove the preemphasis module before calculating DFT. The second point, also the most important point, is that we replace logarithm operation with cubic root, which is inspired by power-law nonlinearity. Different from logarithm nonlinearity, Stevens' power law believes the relationship between perceived intensity and stimulus intensity is power law and 0.33 is the best choice of power exponent [1]. Based on this assumption, it is reasonable to apply cubic root after perception-related filter bank like Mel-frequency filter bank. Given extracted subband coefficients, single Gaussian distribution is used to model every filter bin of each time frame, and we also use the framework of MOLRT to do decision-making. TIMIT database [7] mixed with three types of noise from Noisex-92, two of which are nonstationary, is used for performance evaluation. In order to test the effect of power-law nonlinearity on DFT coefficients, we also directly apply cubic root on DFT coefficients and then do verification using algorithms in [4] and [5]. Experimental results show that the proposed Mel-frequency subband coefficients with power-law nonlinearity could improve the performance of MOLRT based VAD method a lot, even better than the recently proposed non-negative matrix factorization based VAD algorithm in [8]. Also, directly applying cubic root on DFT coefficients could also bring some improvement.

The rest of this paper is organized as follows. Section 2 explains the relation of our method to prior work. In section 3 we review statistical model based VAD algorithm and further demonstrate our proposed method in section 4. Section 5 shows experimental settings and results analysis and the conclusion is made in section 6.

2. Relation to prior work

To the best of our knowledge, this is the first work to use Mel-frequency subband coefficients with power-law nonlinearity for statistical model based VAD. Also, applying power-law nonlinearity directly on DFT coefficients could elevate VAD performance of existing methods is a new finding in this research. We draw the inspiration mainly from the following two aspects.

First, subband coefficients are widely used in computational auditory scene analysis, where a perception-related filter bank is often used as front-end processing of audio signal. [9] uses a Mel-frequency filter bank to first bandpass filter time-domain signal into different channels, and then DFT is applied to obtain a new time-frequency representation to estimate binary mask. The theory in [10] proposes to use gammatone filter bank as preprocessing module, which is uniformly spaced on equivalent rectangular bandwidth (ERB) rate scale. The obtained subband coefficients called cochleagram function as time-frequency representation for speech and noise separation. These methods are believed to improve intelligibility of enhanced speech. Here, we extend this time-frequency representation to VAD and obtain the Mel-frequency subband coefficients according to the scheme of calculating MFCC.

Second, inspired by research on audiology and psychoacoustics, power-law nonlinearity has already been used in many speech-related research, especially in feature extraction research. In fact, in the theory of [10] cubic root is standardly applied to the time-frequency representation before separation algorithm. In the literature of ASR, the well-known Perceptual Linear Prediction (PLP) coefficients [11] and recently proposed Power-Normalized Cepstral Coefficients (PNCC) [12] both employ power-law nonlinearity before DCT procedure. These two features have been proved to be noise robust when being applied to robust ASR, especially PNCC. In the task of speaker identification using MFCC, [13] concludes that performance could be improved a lot by replacing logarithm with cubic root in the calculation of MFCC. Based on that, in our research power-law nonlinearity is also applied to the feature extraction module of VAD.

3. Statistical model based VAD algorithm review

Statistical model based VAD algorithm uses Gaussian distribution to model DFT coefficient in each frequency bin. Given a frame of magnitudes of DFT coefficients $\mathbf{X}$, two hypotheses H_1 and H_0 which respectively represent speech active and speech absent could be formulated. Considering the independence of different frequency bins, the probability density functions of $\mathbf{X}$ conditioned on H_1 and H_0 are as follows:

$$p(\mathbf{X}|H_0) = \prod_{k=1}^{N} \frac{1}{\pi\lambda_N(k)} \exp\left\{-\frac{X_k^2}{\lambda_N(k)}\right\} \tag{1}$$

$$\begin{aligned} p(\mathbf{X}|H_1) = \prod_{k=1}^{N} & \frac{1}{\pi\left[\lambda_N(k)+\lambda_S(k)\right]} \\ & \cdot \exp\left\{-\frac{X_k^2}{\lambda_N(k)+\lambda_S(k)}\right\}, \end{aligned} \tag{2}$$

where N is the number of effective DFT points, $\lambda_N(k)$ and $\lambda_S(k)$ are noise and speech variance separately, X_k is the magnitude value of corresponding frequency bin of $\mathbf{X}$.

Then the log likelihood ratio of the given frame could be calculated by

$$\log\Lambda = \frac{1}{N}\sum_{k=1}^{N}\log\Lambda_k \tag{3}$$

where

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\}. \tag{4}$$

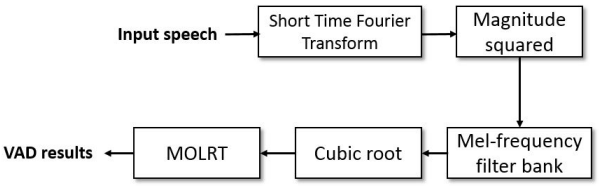

Figure 1: Diagram of proposed VAD algorithm.

$\xi_k = \lambda_S(k)/\lambda_N(k)$ and $\gamma_k = X_k^2/\lambda_N(k)$ are called a priori and a posteriori signal-to-noise ratio (SNR) respectively [3].

$\lambda_N(k)$ could be calculated using static estimation procedure. But here we use the noise variance update rule proposed in [5]. Then, the core problem becomes the estimation of a priori SNR ξ_k. [2] employs a decision-directed (DD) method to calculate ξ_k. Finally, the likelihood ratio of each frame is obtained. However, in order to smooth the output likelihood and give reasonable decision results, MOLRT is applied with the assumption that the contextual DFT coefficients of current frame are independent [4]. MOLLR is calculated as

$$\Lambda_i = \frac{1}{2m+1}\sum_{t=i-m}^{i+m}\log\Lambda_t. \tag{5}$$

Λ_i is the MOLLR of current frame i and m decides the window length. The decision rule is to simply threshold on Λ_i as

$$\mathbf{o}_i = \begin{cases} H_1, \text{ when } \Lambda_i > \eta \\ H_0, \text{ when } \Lambda_i \leqslant \eta \end{cases} \tag{6}$$

η is the threshold which can be tuned considering the tradeoff between hit rate and false alarm.

4. Proposed method

Based on the description above, a diagram as in figure 1 could be drawn to illustrate the procedure of our proposed method. The feature extraction section follows the calculation of MFCC. First Short Time Fourier Transform (STFT) is calculated using proper window settings without preemphasis. After obtaining magnitude of spectrogram, a multiple channel filter bank, whose center frequencies are uniformly distributed on Mel scale, is employed to integrate across frequencies according to the bandwidth of filters. This process is believed to mimic auditory filtering. When calculating MFCC the next step is to apply logarithmic nonlinearity to the output of filter bank. In our proposed method, we change this operation to cubic root based on the power-law nonlinearity. Then the novel Mel-frequency subband coefficients are obtained, which is used as feature in MOLRT based statistical model based VAD. Finally we could get the VAD results.

Another aspect of our work is that cubic root is directly utilized to nonlinearly compress the magnitude of DFT coefficients used in the VAD methods of [4] and [5]. This nonlinearly compressed feature is also compared with proposed subband coefficients in experiments.

5. Experimental settings and results

We use TIMIT database for the evaluation of our proposed VAD method. The first 200 utterances in test folder of TIMIT are used, which cover 20 speakers, 7 males and 13 females. In order to make the evaluation more reasonable, the transcription of

TIMIT is updated. We use a simple energy-based VAD method to relabel the speech and silence section of clean TIMIT utterances. Then the label is combined with existing TIMIT transcription to form the new transcription for our experiments. To balance the percentage of speech and silence, 0.5 second more silence is added to the head and tail of each clean TIMIT utterance, resulting in average length of 4-5s and a speech percentage about 65%. Three types of noise from Noisex-92 database, which include babble, factory and white noise, are added to all clean utterances in three levels of SNR: 0dB, 5dB, 10dB. Then, we have 200 testing utterances for each of the three kinds of noise and three levels of SNR.

For MOLRT based VAD methods mentioned in this paper, parameters of STFT are set as follows: frame length is 32 milliseconds and overlap ratio is 0.5 with Hamming window. Window length when calculating MO-LLR is set to 17 empirically with m in equation (5) equals 8. η in equation (6) is tuned to draw the receiver operating characteristic (ROC) curves. We accumulate the number of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) for each utterances [1] and then true positive rate (TPR) and false positive rate (FPR) of the whole testing utterances of one specific type of noise and one specific SNR level can be obtained through the accumulated TP, FN, FP and TN. In order to quantitatively demonstrate the comparison between different VAD methods, we also employ accuracy at equal error rate (EER) [8] as another form of results. EER is the error rate at which TP and FP sum to 1, and the accuracy at EER could be calculated through TP and TN.

We compare our proposed method with the methods in [4], [5] and [8], respectively called according to authors' names as "Ramírez", "Tan" and "Germain" in result demonstration. The Mel-frequency filter bank in proposed method has 128 channels, each channel of which is a triangular shaped bandpass filter in Mel domain. Note that the sparsity strength in [8] is set to 16 which could give the best performance with our testing utterances. This is also a disadvantage of [8] that the sparsity strength has to be tuned according to the length of test utterance. Realization of [4] also employs the noise variance update method in [5]. Results in 0dB SNR with babble, factory and white noise are shown in figure 2. From ROC curves, it is obvious that our proposed method outperforms other 3 existing methods both in nonstationary and stationary noise. When background is white noise, the superiority of the proposed method is greater. Accuracy at EER of the four methods are shown in table 1, including results under three different types of noise and three levels of SNR. Bold values indicate the accuracy at EER of the proposed method, which is larger than all other methods in all situations. Thus, under all types of noise and all levels of SNR with our experimental settings the proposed method is superior to other three methods, especially when background is white noise. Therefore, it could be firmly proved that our proposed VAD method overmatches the method in [4], [5] and [8], and our proposed Mel-frequency subband coefficients are also better than DFT coefficients in the task of statistical model based VAD.

The other interesting finding during this work is that by simply applying cubic root to DFT coefficients the VAD method in [4] and [5] could also be improved. The accuracy at EER of [4] and [5] with and without cubic root operation on DFT coefficients are shown in table 2 together with the performance of our proposed method. "Tan_r3" and "Ramírez_r3" respectively represent Tan's and Ramírez's method with cubic root operation on DFT coefficients. Note that bold values in neighbor rows indicate the performance is elevated by employing cubic root operation when comparing "Tan" and "Tan_r3", "Ramírez" and "Ramírez_r3". It could be found from table 2 that cubic root compression indeed improves statistical model based VAD using DFT coefficients, and it is believed that the method in [2] could also be improved by this way because based on [2] the work of [4] just replaces the HMM based hang-over scheme

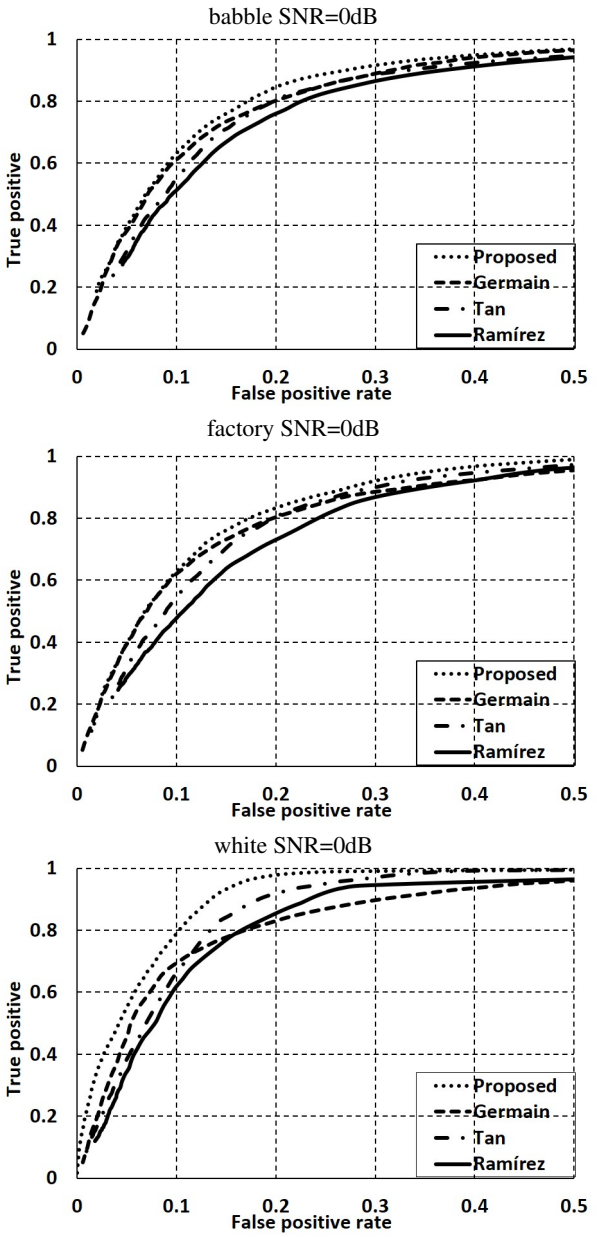

Figure 2: ROC curves of four methods with TIMIT mixed with babble, factory and white noise under 0dB SNR. Dotted lines represent the performance of our proposed method. Dashed lines represent the performance of [8]. Dashed lines with dot represent the performance of [5]. Solid lines represent the performance of [4].

[1] In 1-0 classifier system, TP means the prediction is 1 and the sample is really 1. FN means the prediction is 0 and the sample is actually 1. FP means the prediction 1 and the sample is actually 0. TN means the prediction is 0 and the sample is really 0.

Table 1: Accuracy at EER of four different VAD methods under three types of noise and three levels of SNR.

	babble			factory			white		
	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Proposed	**81.9%**	**84.4%**	**86.9%**	**81.5%**	**85.5%**	**87.2%**	**87.4%**	**88.2%**	**88.5%**
Germain	80.1%	83.7%	86.6%	80.3%	83.3%	86.4%	81.5%	84.8%	87.3%
Tan	80.1%	83.0%	84.5%	80.1%	82.5%	84.5%	84.6%	86.2%	86.3%
Ramírez	78.5%	78.4%	83.5%	77.4%	81.1%	83.6%	82.1%	85.1%	85.7%

Table 2: Accuracy at EER of five different VAD methods under three types of noise and three levels of SNR.

	babble			factory			white		
	0dB	5dB	10dB	0dB	5dB	10dB	0dB	5dB	10dB
Proposed	81.9%	84.4%	86.9%	81.5%	85.5%	87.2%	87.4%	88.2%	88.5%
Tan_r3	79.4%	**84.1%**	**86.3%**	79.6%	**84.4%**	**85.9%**	**84.7%**	**87.9%**	**88.3%**
Tan	80.1%	**83.0%**	**84.5%**	80.1%	**82.5%**	**84.5%**	**84.6%**	**86.2%**	**86.3%**
Ramírez_r3	**79.2%**	**83.1%**	83.5%	**77.9%**	**82.6%**	**84.8%**	**86.2%**	**87.4%**	**88.1%**
Ramírez	**78.5%**	**78.4%**	83.5%	**77.4%**	**81.1%**	**83.6%**	**82.1%**	**85.1%**	**85.7%**

with MOLRT but the feature used is not changed.

6. Conclusion

In this paper, we propose an efficient way to improve the performance of statistical model based VAD. A novel feature based on Mel-frequency subband coefficients is employed and MOLRT is used for decision-making for the task of statistical model based VAD. The highlight of the proposed Mel-frequency subband coefficients is that instead of logarithm operation cubic root nonlinear compression is used. Through experiments using TIMIT and Noisex-92 database, it is firmly proved that by using the proposed feature statistical model based VAD algorithms could be improved a lot. Also the proposed method is superior to the recently proposed method in [8], meanwhile maintaining speaker and noise independence and easily being extended to online algorithm. In another experiment we find that the performance of traditional statistical model based VAD methods could also be elevated by simply applying cubic root on DFT coefficients.

In future work, we would like to compare different subband coefficients obtained from different filter banks such as gammatone filter bank and Bark-scale filter Bank in the task of statistical model based VAD.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 91120015, No. 11161140319).

The authors also would like to thank F. Germain for his code.

8. References

[1] T. Virtanen, R Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Janaury 1999.

[3] Y. Ephraim and D. Malah, "Spech enhancement using a minimum mean-squre error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109-1121, December 1984.

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

[5] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, March 2010, pp. 4466-4469.

[6] J. Wu and X. Zhang, "An efficient voice activity detection algorithm by combining statistical model and energy detection", *Eurasip Journal on Advances in Signal Processing*, vol. 2011, pp. 1-10, 2011.

[7] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "TIMIT acoustic-phonetic continous speech corpus CD-ROM", in *NIST*, 1993.

[8] F. G. Germain, D. L. Sun and G. J. Mysore, "Speaker and noise independent voice activity detection", in *Proceedings of Interspeech*, August 2013, pp. 732-736.

[9] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners", *Journal of Acoustical Society of America*, 126(3), pp. 1486-1494, 2009.

[10] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Hoboken, NJ: Wiley-IEEE Press, 2006.

[11] H. Hermansky, "Perceptual linear prediction analysis of speech", *Journal of Acoustical Society of America*, vol. 87, pp. 1738-1752, April 1990.

[12] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, March 2012, pp. 4101-4104.

[13] X. Zhao and D. L. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, May 2013, pp. 7204-7208.