SPARSE POWER SPECTRUM BASED ROBUST VOICE ACTIVITY DETECTOR

Datao You, Jiqing Han, Guibin Zheng, Tieran Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

This paper presents a robust approach to improve the performance of voice activity detector (VAD) in low signal-to-noise ratio (SNR) noisy environments. To this end, we first generate sparse representations by Bregman Iteration based sparse decomposition with a learned over-complete dictionary, and derive a kind of audio feature called sparse power spectrum from the sparse representations. we then propose a method to calculate the short segment average spectrum and long segment average spectrum from sparse power spectrum. Finally, we design a criterion to detect speech region and non-speech region based on the above average spectrum. Experiments show that the proposed approach further improves the performance of VAD in low SNR noisy environments.

Index Terms— Sparse decomposition, sparse spectrum, average energy, voice activity detection.

1. INTRODUCTION

Voice activity detector is a mean to distinguish speech segments from non-speech segments in an audio stream. VAD plays a critical role on increasing the capacity of transmission and speech storage by reducing average bit-rate, therefore it is widely applied in many speech applications [1,9], including mobile communication services, real-time internet telephony, automatic speech recognition, speech enhancement, and variable rate speech coding.

A variety of VAD algorithms were proposed over the past decades [1-5]. Most of these methods use features that depend on frequency-domain energy, zero-crossing rate, correlation coefficients, and periodicity estimation. Since their shortcomings, such as sensitive to noises and inaccuracy in SNR estimating, these methods do not work well in low SNR noisy environments. Recently, statistical model based algorithms have been proposed [6-9]. This kind of VADs are posed as a hypothesis testing problem with statistical models of speech and noise, although assumptions made about the statistics of noise do not always hold in practice. The statistical model based algorithms are not good in low SNR noisy environments, since the inherent high-fluctuation of a-posteriori SNR and the abruption of transient speech segments are more than the noise segments[16].

It is known that speech signal is mainly superposed by a group of underlying components [10-12], which provide important cues for improving the detection performance. Obtaining these cues is a difficult task, the important reason is that the elements of decomposition set do not well reflect the speech underlying components in most signal processing approaches [10]. Currently, Bregman Iteration based sparse decomposition not only work well in reducing noise, but also shows good performance at capturing the underlying components from noisy speech [13]. Based on Bregman Iteration based sparse decomposition, we propose a VAD approach to improve the detection performance under low SNR noisy condition.

The proposed approach contains the following steps. First, an over-complete dictionary is learned from speech corpus by online dictionary learning algorithm [14]. It has been shown that learned dictionary is well adapted to natural speech underlying components [14]. Next, the Bregman Iteration based sparse decomposition is used to extract the speech signals from noisy speech and generate sparse representations [13]. Then, squaring the sparse representations and add them to get a kind of sparse power spectrum. After that, the short segment average spectrum (SSAS) and long segment average spectrum (LSAS) are calculated from sparse power spectrum. Finally, a detection criterion is designed to detect speech and non-speech segments. The criterion is based on short segment and long segment average energy. Experimental results show that the proposed approach further improves the detection performance of VAD in low SNR noise environments.

2. SPARSE DECOMPOSITION

2.1. Dictionary learning

In order to well adapt to the speech signals, an over-complete dictionary is learned from speech corpus. The dictionary learning is an optimization problem [10], which can be estimated by solving the following regression problem

$$\underset{\Psi,C}{\operatorname{argmin}}\, \lambda\|C\|_0 + \frac{1}{2\delta}\|X - \mathbf{\Psi}C\|_2^2 \qquad (1)$$

where $\|.\|_0$ represents l_0-norm operator; λ and δ^2 denote the weight of the l_0-norm function and the variance of residu-

This research is partly supported by the National Natural Science Foundation of China under grant No. 91120303 and 61071181.

al energy respectively; $X = [x_1, \cdots, x_T]$ is a speech frame set, in which each column x_t is a D-dimensional vector; $C = [c_1, \cdots, c_T]$ is the sparse coefficient set, in which each column c_t is a L-dimensional vector; $\mathbf{\Psi} = [\psi_1, \cdots, \psi_L] \in R^{D \times L}$ represents an over-complete dictionary which is initialized by cosine vectors, and each column ψ_l of $\mathbf{\Psi}$ is a unitary vector.

Since there are two problems in (1): both of l_0-norm and joint optimization of $\mathbf{\Psi}$ and C are non-convex functions, so it is difficult to conduct. Fortunately, for l_0-norm problem, Candés et.al [15] proved that it can be replaced by l_1-norm; and for the joint optimization problem, it can be simplified by alternating optimization between $\mathbf{\Psi}$ and C. Then the modified optimization formula can be rewritten as follow

$$\left\{ \begin{array}{l} \underset{C}{\operatorname{argmin}} \lambda \|C\|_1 + \frac{1}{2\delta} \left\|X - \mathbf{\Psi} C\right\|_2^2 \\ \underset{\Psi}{\operatorname{argmin}} \lambda \|C\|_1 + \frac{1}{2\delta} \left\|X - \mathbf{\Psi} C\right\|_2^2 \end{array} \right. \tag{2}$$

2.2. Sparse decomposition

For a given noisy speech segment $s_t = x_t + n_t \, (t = 1, \cdots, T)$ and the learned dictionary $\mathbf{\Psi}$, the sparse decomposition is

$$\underset{c_t}{\operatorname{argmin}} \lambda \|c_t\|_1 + \frac{1}{2\delta} \left\|s_t - \mathbf{\Psi} c_t\right\|_2^2 \tag{3}$$

where c_t denotes the sparse representation vector of s_t, and satisfies

$$\|c_t\|_0 \ll L \tag{4}$$

Sparse decomposition is effective to encode the incoming signals, but it is a difficult problem to reduce noises from noisy speech for the minimization equation. Currently, J-F. Cai et.al [13] have argued that Bregman Iteration algorithm is an efficient and robust-to-noise algorithm for solving the minimization equation (4). The procedures of Bregman Iteration are

$$\left\{ \begin{array}{l} v^{k+1} = v^k - \mathbf{\Psi}^{'} (\mathbf{\Psi} c_t^k - s_t) \\ c_t^{k+1} = \Upsilon_\delta (\delta v^{k+1}) \end{array} \right. \tag{5}$$

where $c^0 = v^0 = 0$, and

$$\Upsilon_\zeta(\omega) := [\gamma_\zeta(\omega(1)), \cdots, \gamma_\zeta(\omega(L))]^{'} \tag{6}$$

(7) is the soft threshold operator with

$$\gamma_\zeta(\xi) = \left\{ \begin{array}{ll} 0, & if\, |\xi| \leq \zeta \\ sgn(\xi)(|\xi| - \zeta), & if\, |\xi| > \zeta \end{array} \right. \tag{7}$$

and the stopping criteria is

$$std(s_t - \mathbf{\Psi} c_t^{k+1}) < \delta \, or \, iter. < 1000 \tag{8}$$

where $std(.)$ denotes the standard deviation operator of $s_t - \mathbf{\Psi} c_t^{k+1}$; the iteration of (6) will stop whenever the std of residual $s_t - \mathbf{\Psi} c_t^{k+1}$ is less than δ or the number of iterations exceeds 1000, and the final result c_t^{k+1} is assigned to c_t.

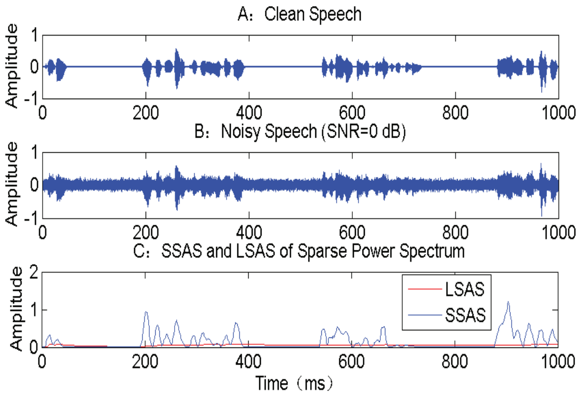

Fig. 1. SSAS and LSAS curves of a noisy speech stream (SNR=0 dB and the added noise is white noise).

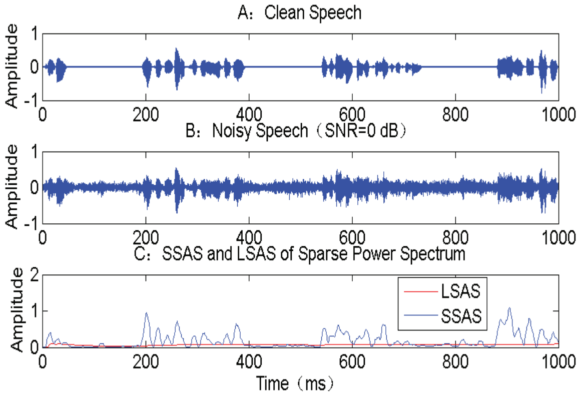

Fig. 2. SSAS and LSAS curves of a noisy speech stream (SNR=0 dB and the added noise is babble noise).

3. THE PROPOSED APPROACH

3.1. Sparse power spectrum

On the basis that the additive noises are reduced from noisy speech by Bregman Iteration algorithm, we square the elements of the above sparse representation vector c_t and add them to generate a kind of energy feature, which we call sparse power spectrum. The formula is

$$e_t = \frac{1}{L} \sum_{l=1}^{L} c_{l,t}^2 \tag{9}$$

where $c_{l,t}$ is the lth element of sparse representation vector c_t; and e_t denotes the sparse power spectrum of c_t.

3.2. SSAS and LSAS

We found that the sparse power spectrum of speech segments generally greater than the non-speech segments, even in low

SNR noisy environments. Due to this finding, it is reasonable to assume that the discrimination of the SSAS, which is the average of a small number of adjacent sparse power spectrums, is more robust and discriminating (see Fig. 1, 2). Then, it is also rational to refer that the LSAS, which is the average of a great number of adjacent sparse power spectrums, is more stationary and can be used as a threshold of the SSAS between speech segment and the non-speech segment (see Fig. 1, 2). Based on the above assumption and reference, we proposed a simple and efficient algorithm to detect the speech segments and non-speech segments in low SNR noisy environments.

Given sparse power spectrum $e_t \, (t = 1, \cdots, T)$, the SSAS can be computed by the following formula

$$y_t = \frac{1}{2I+1} \sum_{i=t-I}^{t+I} e_i \qquad (10)$$

where I represents the displacement length from the sparse power spectrums e_t, which is assigned to 3 in this paper; and y_t denotes the SSAS of e_t.

Given sparse power spectrum $e_t \, (t = 1, \cdots, T)$, the LSAS calculating representation of LSAS is

$$\beta_t = \frac{1}{t - t_t} \sum_{i=t_t}^{t} e_i \qquad (11)$$

where t_t is the start time for calculating the LSAS of e_t; in this paper, t_t is either assigned to $t - 6000$ when $t > 6000$, or assigned to 1; and β_t denotes the LSAS of e_t.

3.3. Detection criterion

Fig. 1 and 2 show the SSAS and LSAS curves of a speech stream in noise environments (SNR=0 dB). The speech comes from a man speaker of TIMIT corpus. The noises are white and babble noise respectively, and both of them are taken from NOISEX-92 database. The blue curves of part C represent the SSAS and the red curves represent the LSAS in the two figures, and the comparisons of SSAS and LSAS in the both features indicate that the above assumption and reference accord with the factual cases. So we can design a detection criterion based on the differences between SSAS and LSAS.

When the SSAS and LSAS of noisy frame $s_t \, (t = 1, \cdots, T)$ is given, the formula of the designed detection criterion is

$$\begin{cases} H_0 : & y_t < \beta_t \\ H_1 : & y_t \geq \beta_t \end{cases} \qquad (12)$$

where H_0 and H_1 denote speech absence and presence, respectively; y_t is the SSAS of noisy audio frame s_t; and β_t is the LSAS of noisy audio frame s_t.

4. EXPERIMENTS AND ANALYSIS

The duration of test speech data was 240s which concatenated by recordings of four males and four females. The total recording duration of each speaker was approximately 30s. These speech recordings were selected from TIMIT corpus, and were sampled at 8000 Hz. The speech material was marked manually at each 10 ms frame. In the test data, 55.84% of the whole frames were labeled as active speech frames, in which 38.16% were voiced sound and 17.66% were unvoiced sound frames. In order to simulate adverse environments, the white, babble, and factory noises were mixed with the clean test data of 0 dB SNR. These noises were taken from NOISEX-92 database. For evaluating the performance of the proposed approach, the detection probability (PD) and false alarm probability (PF) are exploited. In addition, Gaussian-LRT [7] was exploited as the baseline, and the proposed approach is named SSAS-LSAS for short.

Fig. 3, 4, and 5 present the ROC curves of the proposed approach under noises condition (SNR=0 dB). Fig. 3 shows that, in white noise environment, the performance of SSAS-LSAS is a little less than the baseline when PF < 0.058, but the performance of SSAS-LSAS obviously outperforms the baseline when the PF > 0.058. Fig. 4 displays that, under factory noise condition, the performance of SSAS-LSAS is better than the baseline when $0.016 <$ PF < 0.16, and the performance of the SSAS-LSAS is similar with the baseline when PF < 0.016 and PF > 0.16. Fig. 5 shows that SSAS-LSAS obviously outperforms the baseline.

From these experimental results, a conclusion can be derived that the proposed approach further improves the detection hit rates of VAD in low SNR noisy environments.

5. CONCLUSION

In this paper, we proposed a robust VAD approach to improve the detection performance under low SNR noisy conditions. We first learn a over-complete dictionary from speech corpus to well adapt the natural speech underlying components. Then, we generate sparse representations by Bregman Iteration based sparse decomposition with the learned over-complete dictionary. After that, we derive a kind of audio feature named sparse power spectrum, and go further to generate the SSAS and LSAS for detection. Finaly, we present detection criterion based on the above SSAS and LSAS. Experiments show that the proposed approach can further improve the performance of VAD under low SNR and noise conditions.

6. REFERENCES

[1] L.R. Rabiner, and M.R. Sambur, ”Voiced-unvoiced-silence detection using Itakura LPC distance measure, ” *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 323-326, 1977.

[2] J.D. Hoyt, and H. Wechsler, ”Detection of human speech

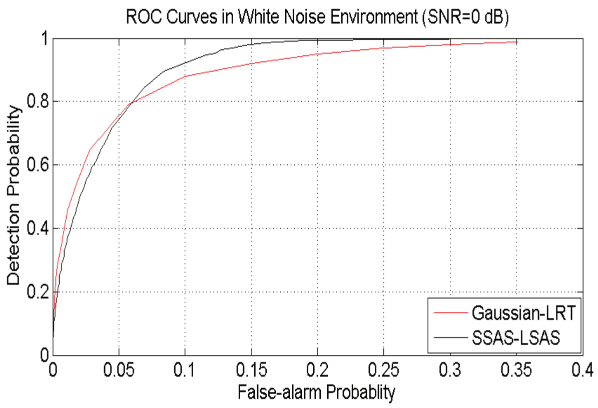

Fig. 3. Evaluation of SSAS-LSAS and baseline under white noise condition (SNR=0 dB).

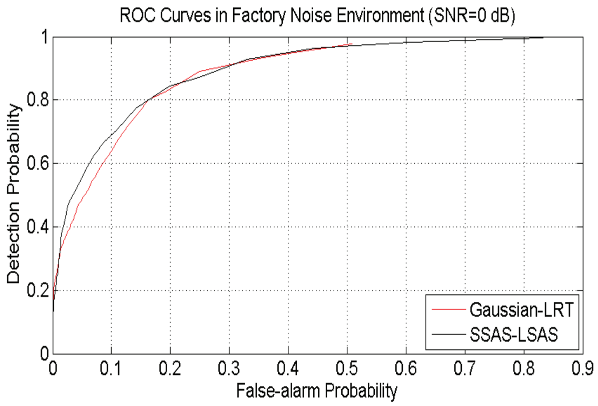

Fig. 4. Evaluation of SSAS-LSAS and baseline under factory noise condition (SNR=0 dB)

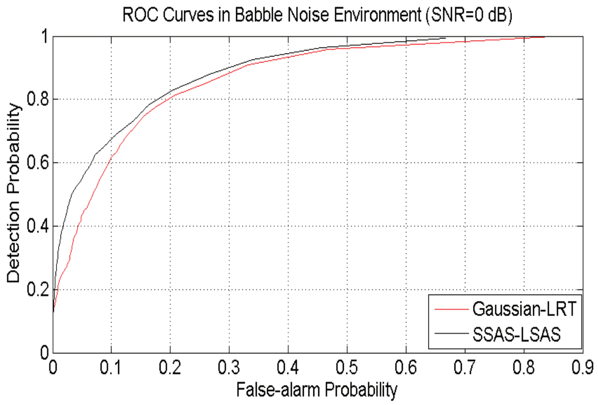

Fig. 5. Evaluation of SSAS-LSAS and baseline under babble noise condition (SNR=0 dB).

in structured noise," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 237-240, 1994.

[3] J.C. Junqua, B. Reaves, and B. Mark, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," *Euronspeech*, pp. 1371-1374, 1991.

[4] J.A. Haigh, and J.S. Mason, "Robust voice activity detection using cepstral feature," *IEEE TELCON*, pp. 321-324, 1993.

[5] R. Tuchker, "Voice activity detection using a periodicity measure," *Inst. Elect. Eng.*, Vol. 139, pp. 377-380, 1992.

[6] J. Sohn, N.S. Kim an W.A. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, Vol. 6, No. 1, pp. 1-3, Jan. 1999.

[7] Y.D. Cho, K.A. Naimi and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 7-11, May, 2001.

[8] J. Ramirez, J.C. Segura, C. Benitez, L. Garcia, and A. Rubio,"Statistical voice detection using a multiple observation likelihood ratio test", *IEEE Signal Processing Letters*, vol. 12, no. 10, Oct. 2005.

[9] J.W. Shin, H.J. Kwon, S.H. Jin and N.S. Kim, "Voice activity detection based on conditional MAP criterion," *IEEE Signal Processing Letters*,Vol. 15, 2008.

[10] S. Mallat, "A Wavelet Tour of Signal Rrocessing, the Sparse way," Academic Press, 2009.

[11] E. Smith, M.S. Lewicki, "Efficient Coding of Time-relative Structure Using Spikes," *Neural Computation*, Vol. 17, pp. 19-45 2005.

[12] G. Peyre, "Best basis compressed sensing," *IEEE Transaction on Signal Processing*, Vol. 58, No. 5, pp. 2613-2622, May. 2010.

[13] J-F. Cai, S. Osher, and Z. Shen, "Linearized bregman iteration for compressed sensing," *Math. Comp.*, pp. 1515-1536, 2009.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *in Proc. 26th ICML*, 2009.

[15] E. Candès, Z. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, Vol. 35, No. 6, pp. 2313-2351, 2007.

[16] P.K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 19, No. 3, pp. 600-613, Mar. 2011.