VOICE ACTIVITY DETECTION USING CONVOLUTIVE NON-NEGATIVE SPARSE CODING

Peng Teng and Yunde Jia

School of Computer, Beijing Institute of Technology, China
Beijing Laboratory of Intelligent Information Technology, China
{tengpeng, jiayunde}@bit.edu.cn

ABSTRACT

This paper presents a voice activity detection (VAD) approach using convolutive non-negative sparse coding (CNSC) to improve the detection performance in low signal-to-noise (SNR) conditions. Our idea is to use noise-robust feature for speech signal detection *while* noise is reduced away. We first use magnitude spectrum as the non-negative and additive low-level representation of audio signals, and learn a speech dictionary from clean speech as well as a noise dictionary from noise samples. Then, the two dictionaries are concatenated to form a global dictionary, and an audio signal is decomposed into coefficient vectors using CNSC on the global dictionary. Only coefficients corresponding to the bases from the speech dictionary are taken as the features for the signal. At last, the activity labels is given by decoding a conditional random field (CRF) which is constructed to model the context of an audio signal for VAD. Experiments demonstrate that our VAD approach has an excellent performance in low SNR conditions.

Index Terms— voice activity detection, convolutive non-negative sparse coding, conditional random fields

1. INTRODUCTION

Voice activity detection (VAD), to detect the presence of speech in an audio signal degraded by noise, is widely applied in numerous modern speech communication systems. In the last decade, since a statistical model based VAD approach with impressive performance was proposed by Sohn *et al.* [1], many VAD algorithms focus on statistical model based approaches which have the decision rules derived from the likelihood rate (LR) to discriminate speech/nonspeech [2–5]. Treating VAD as a binary classification problem, some classifiers based on statistical learning theory were also employed [6–9]. For example, Jo *et al.* [8] introduced support vector machines based VAD with feature vectors consisting of LRs in each frequency bin. Concerning the context in an audio signal, Saito *et al.* [10] developed a VAD system based on conditional random fields (CRF) [11] using multiple popular features. Most of these methods adopt noise-sensitive features, e.g., those depending on time-domain statistics, frequency-domain energy, and correlation coefficients. To address these issues, You *et al.* [12] proposed a VAD algorithm based on the sparse coding of audio signals, showing good robustness to noise in low SNR conditions.

Sparse coding is namely the linear decomposition of a signal with a few weighted bases (so-called the sparseness constraint) from an over-complete dictionary, so that the signal can be represented as sparse vectors (coefficient vectors) consisting of these weights. Especially, sparse coding with a learned dictionary instead of a predefined one (e.g., based on wavelets) has recently led to state-of-the-art results in numerous low-level signal processing tasks, such as image denoising [13], audio processing [14] due to the noise-robust representation provided by the coefficient vector. Besides the sparseness, the non-negativity is another popular constraint to the linear decomposition of signals. The non-negativity constraint demands the dictionary and the coefficient vectors are both non-negative. A linear decomposition with this constraint, called nonnegative matrix factorization (NMF) [15], gives "*parts*" based representation as only additive combinations of bases are allowed in representation. Constrained by both the sparseness and the non-negativity, the linear decomposition is namely non-negative sparse coding (NSC) [16, 17]. Considering the temporal dependency in audio signals, Wang [18] extended NSC to a more general framework, called convolutive non-negative sparse coding (CNSC), using a convolutive decomposition model instead of the linear decomposition model. Recently, an online CNSC algorithm [19] is developed. It has been employed in the tasks of speech separation [19] and speech overlap detection [20] with good performances.

In this paper we propose a VAD approach using CNSC to improve the detection performance in low SNR conditions. The improvement is achieved by adopting noise-robust features *while* reducing the noise away from audio signals, which was not considered in the earlier studies [1–10, 12] . First, we use magnitude spectrum as the low-level representation of audio signals. Magnitude spectrum is non-negative in its elements, and supposed to be approximatively additive between two simultaneous audio signals, i.e., speech signals and noise signals in the VAD task. Next, an over-complete convolutive dictionary of speech signals (speech dictionary) is learned from clean speech signals using CNSC, as well as a dictio-

nary of noise signals (noise dictionary) is learned from noise sample signals using convolutive NMF (CNMF). Then, we concatenate the two dictionaries to form a global dictionary. According to the global dictionary we decompose a given audio signal using CNSC into a sequence of coefficient vectors. Discarding the coefficients corresponding to the noise dictionary, only coefficients corresponding to the speech dictionary are selected as the noise-robust features of the signal. At last, we train a CRF with a linear chain structure to model the correlation between feature sequences and voice activity labels along an audio signal. For a given audio signal, its voice activity labels are obtained by decoding the CRF with the input of its feature sequences. Experimental results show that our VAD approach has an excellent performance in low SNR conditions.

2. AUDIO SIGNAL ANALYSIS USING CNSC

Let $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ be the magnitude spectrum of an audio signal with T time frames, where $\boldsymbol{x}_t = [x_t^{(1)}, \cdots, x_t^{(l)}, \cdots, x_t^{(L)}]^\top$ denotes the magnitude of the t-th time frame; l is the frequency-bin index, and $x_t^{(l)} \geq 0$. $\boldsymbol{x}_t$ can be approximated by a linear combination of an over-complete set of bases $\boldsymbol{d}_m$ with weights $\alpha_t^{(m)}$:

$$\boldsymbol{x}_t = \mathbf{D}\boldsymbol{\alpha}_t, \tag{1}$$

where $\mathbf{D} = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_M], M > L$, is called a dictionary, $\boldsymbol{\alpha}_t = [\alpha_t^{(1)}, \cdots, \alpha_t^{(M)}]^\top$ is called a coefficient vector. Denoting $\mathbf{A} = [\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_T]$, with the non-negativity constraint $\mathbf{D} \in \mathbb{R}_{\geq 0}^{L \times M}$ and $\mathbf{A} \in \mathbb{R}_{\geq 0}^{M \times T}$, the magnitude spectrum $\mathbf{X}$ can be decomposed using non-negative matrix factorization (NMF) as

$$\mathbf{X} \approx \mathbf{D}\mathbf{A}. \tag{2}$$

With a typical sparseness constraint on $\mathbf{A}$, this decomposition can be achieved by minimizing the distance between the original matrix and its approximation:

$$(\hat{\mathbf{D}}, \hat{\mathbf{A}}) = \arg\min_{\mathbf{D},\mathbf{A}} \| \mathbf{X} - \mathbf{D}\mathbf{A} \|_F^2 + \lambda \sum_{ij} \mathbf{A}^{(ij)} \tag{3}$$

where $\hat{\mathbf{D}}$ and $\hat{\mathbf{A}}$ are the estimated optimal values of $\mathbf{D}$ and $\mathbf{A}$, respectively, $\| \cdot \|_F$ denotes the Frobenius norm, $\mathbf{A}^{(ij)}$ is the ij-th elements of $\mathbf{A}$, and the non-negative constant λ controls the sparsity of $\mathbf{A}$. This is so-called the non-negative sparse coding (NSC) of audio signals, and it is equal to NMF when $\lambda = 0$. However, the decomposition shown in Eq.(2) does not consider the connections between the neighboring columns of $\mathbf{X}$ which are essential to audio signals. To address this issue, a convolutive variant of Eq.(2) is employed. The convolutive NMF (CNMF) takes the form:

$$\mathbf{X} \approx \sum_{r=0}^{R-1} \mathbf{D}_r \overset{r\rightarrow}{\mathbf{A}}, \tag{4}$$

where R is the convolutive range, $\mathbf{D}_r \in \mathbb{R}_{\geq 0}^{L \times M}$. The operator $\overset{r\rightarrow}{\cdot}$ is column shift operator that shifts r columns of the matrix to the right, and vacated columns are filled with zeros. For simplicity, the R-cardinality dictionary set $\{\mathbf{D}_r\}$ $(r = 0, \cdots, R-1)$ for the R-range convolutive operator is also called a *dictionary*, and the convolutive operator in Eq.(4) is expressed as

$$\mathbf{X} \approx \{\mathbf{D}_r\} \otimes \mathbf{A}. \tag{5}$$

Convolutive NSC (CNSC) is namely CNMF with the sparseness constraint. The dictionary $\{\mathbf{D}_r\}$ for CNSC can be learned according to Eq.(6), and the decomposition of $\mathbf{X}$ with a given $\{\mathbf{D}_r\}$ can be achieved according to Eq.(7), both using the CNSC algorithm proposed in [18] or [19]. CNSC is equal to CNMF when $\lambda = 0$.

$$\{\hat{\mathbf{D}}_r\} = \arg\min_{\{\mathbf{D}_r\}} \| \mathbf{X} - \{\mathbf{D}_r\} \otimes \mathbf{A} \|_F^2 + \lambda \sum_{ij} \mathbf{A}^{(ij)} \tag{6}$$

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \| \mathbf{X} - \{\mathbf{D}_r\} \otimes \mathbf{A} \|_F^2 + \lambda \sum_{ij} \mathbf{A}^{(ij)} \tag{7}$$

3. AUDIO SIGNALS SPARSE DECOMPOSITION FOR VAD

We propose a novel scheme of audio signals sparse decomposition for VAD. The idea is to represent speech with noise-robust features while noise is reduced away from the signals. Since the magnitude spectrum of a noisy speech signals $\mathbf{X}$ can be deemed as the sum of speech magnitude spectrum $\mathbf{X}^{\mathrm{s}}$ and noise magnitude spectrum $\mathbf{X}^{\mathrm{n}}$, we decompose $\mathbf{X}$ using CNSC on a global dictionary formed by concatenating a speech dictionary $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ and a noise dictionary $\{\mathbf{D}^{\mathrm{n}}{}_r\}$. This decomposition can be formulated by

$$\begin{aligned} \mathbf{X}^{\mathrm{s}} + \mathbf{X}^{\mathrm{n}} = \mathbf{X} &\approx \{[\mathbf{D}^{\mathrm{s}}{}_r, \mathbf{D}^{\mathrm{n}}{}_r]\} \otimes \begin{bmatrix} \mathbf{A}^{\mathrm{s}} \\ \mathbf{A}^{\mathrm{n}} \end{bmatrix} \\ &= \{\mathbf{D}^{\mathrm{s}}{}_r\} \otimes \mathbf{A}^{\mathrm{s}} + \{\mathbf{D}^{\mathrm{n}}{}_r\} \otimes \mathbf{A}^{\mathrm{n}} \end{aligned} \tag{8}$$

where $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ is over-complete and learned from clean speech signals using CNSC to obtain noise-robust bases for representing speech; $\{\mathbf{D}^{\mathrm{n}}{}_r\}$ is low-rank and learned from noise signal samples using CNMF so that it can fit noise well with its few bases. $\mathbf{X}$ is decomposed into sparse coefficient vectors consisting of two parts: the coefficients in $\mathbf{A}^{\mathrm{s}}$ corresponding to $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ and the coefficients in $\mathbf{A}^{\mathrm{n}}$ corresponding to $\{\mathbf{D}^{\mathrm{n}}{}_r\}$. In the ideal case (e.g., bases from the two dictionaries are distinctly dissimilar), these two parts of coefficients can represent the true contributions of bases from the two dictionaries in constructing $\mathbf{X}$, respectively,

$$\begin{cases} \mathbf{X}^{\mathrm{s}} \approx \{\mathbf{D}^{\mathrm{s}}{}_r\} \otimes \mathbf{A}^{\mathrm{s}} \\ \mathbf{X}^{\mathrm{n}} \approx \{\mathbf{D}^{\mathrm{n}}{}_r\} \otimes \mathbf{A}^{\mathrm{n}} \end{cases}. \tag{9}$$

If $\mathbf{A}^{\mathrm{n}}$ is discarded, the noise is supposed to be reduced away, leaving only the residual of fitting the noise with bases from $\{\mathbf{D}^{\mathrm{n}}{}_r\}$. The residual can be deemed as white noise. Then, $\mathbf{A}^{\mathrm{s}}$ seems like to be obtained by decomposing a white noise degraded speech signal on $\{\mathbf{D}^{\mathrm{s}}{}_r\}$. Therefore, under the assumption described in Eq.(9), only $\mathbf{A}^{\mathrm{s}}$ is considered in our VAD approach and used as noise-robust features for $\mathbf{X}$, independently of the original noise.

4. VAD CONTEXT MODELING BASED ON CONDITIONAL RANDOM FIELDS

The goal of the VAD task is to give a sequence of voice activity labels $\mathbf{H} = [H_1, \cdots, H_t, \cdots, H_T]$ along a given audio signal $\mathbf{X}$ where $H_t \in \{0, 1\}$ indicates speech absence or presence at the t-th time frame $\boldsymbol{x}_t$. Let $\boldsymbol{y}_t$ be an observed feature vector derived from $\boldsymbol{x}_t$, and correspondingly $\mathbf{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T]$ be an observed feature vector sequence along $\mathbf{X}$. We model the correlations between $\mathbf{H}$ and $\mathbf{Y}$ using a graphical model with a linear chain structure, as shown in Fig.1. Let $G = (V, E)$ be a graph and $\mathbf{H}$ being indexed by the vertices of G, say H_t. A pair $(\mathbf{H}, \mathbf{Y})$ is called a CRF, if when conditioning on $\mathbf{Y}$, the variables H_t obey Markov property with respect to the graph: $p(H_t|\boldsymbol{y}_t, H_{V-\{t\}}) = p(H_t|\boldsymbol{y}_t, H_{\mathcal{N}_t})$, where $\mathcal{N}_t$ is the set of neighbors of node t and $H_{\mathcal{N}_t}$ is the joint vector of variables in the subscript set. Let $\mathcal{C}(\mathbf{H}, \mathbf{Y})$ be the set of maximal cliques of G. Given an observation $\mathbf{Y}$ and parameters $\boldsymbol{\theta}$, the distribution over a label sequence $\mathbf{H}$ can be defined as:

$$p_{\boldsymbol{\theta}}(\mathbf{H}|\mathbf{Y}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{Y})} \prod_{c \in \mathcal{C}(\mathbf{H}, \mathbf{Y})} \phi^c_{\boldsymbol{\theta}}(\mathbf{H}_c, \boldsymbol{y}_c) \tag{10}$$

$$Z_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{\mathbf{H}} \prod_{c \in \mathcal{C}(\mathbf{H}, \mathbf{Y})} \phi^c_{\boldsymbol{\theta}}(\mathbf{H}_c, \boldsymbol{y}_c) \tag{11}$$

where $\phi^c_{\boldsymbol{\theta}}$ is the positive-valued potential function of clique c and $Z_{\boldsymbol{\theta}}(\mathbf{Y})$ is the observation dependent normalization. For a linear chain, i.e., first-order state dependency depicted in Fig.1, the cliques include pairs of neighboring labels (H_{t-1}, H_t) and feature-label pairs $(H_t, \boldsymbol{y}_t)$. Therefore, for a model with T time frames, the CRF in Eq.(10) can be rewritten in terms of exponentiated feature functions F_θ as:

$$p_{\boldsymbol{\theta}}(\mathbf{H}|\mathbf{Y}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{Y})} \exp\left(\sum_{t=1}^{T} F_{\boldsymbol{\theta}}(H_{t-1}, H_t, \boldsymbol{y}_t) \right) \tag{12}$$

$$Z_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{\mathbf{H}} \exp\left(\sum_{t=1}^{T} F_{\boldsymbol{\theta}}(H_{t-1}, H_t, \boldsymbol{y}_t) \right). \tag{13}$$

In our approach, F_θ is computed in terms of weighted sums over the features of the cliques:

$$\begin{aligned} F_{\boldsymbol{\theta}}(H_{t-1}, H_t, \boldsymbol{y}_t) = & \sum_{b \in \{0,1\}^2} \gamma_b f_b(H_{t-1}, H_t) \\ & + \sum_{h \in \{0,1\}} \sum_{l=1}^{L} \beta_{h,l} g_{h,l}(H_t, y_t^{(l)}) \end{aligned} \tag{14}$$

where the two kind of feature functions are transition feature functions $f_b(H_{t-1}, H_t)$ defined as

$$f_b(H_{t-1}, H_t) = \begin{cases} 1 & \text{if : } b = [H_{t-1}, H_t] \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

and observation feature functions $g_{h,l}(H_t, y_t^{(l)})$ defined by

$$g_{h,l}(H_t, y_t^{(l)}) = \begin{cases} y_t^{(l)} & \text{if : } H_t = h \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

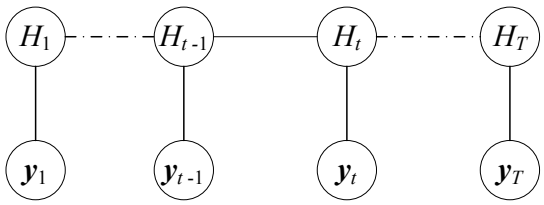


Fig. 1. Graphical model representation of CRF for VAD.

with parameters $\boldsymbol{\theta} = \{\gamma_b, \beta_{h,l}\}$. Notice that the model with tied parameters $\boldsymbol{\theta}$ is used across all cliques, in order to seamlessly handle models of arbitrary size, i.e., sequences of arbitrary length. Assuming a fully labeled training set $\{\mathbf{H}^n, \mathbf{Y}^n\}_{n=1,\cdots,N}$, the CRF parameters $\boldsymbol{\theta}$ can be obtained by maximizing the conditional log-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{H}^n|\mathbf{Y}^n) \tag{17}$$

Likelihood maximization can be performed using a gradient ascent method, e.g., BFGS [21]. With a trained CRF, letting $p(\cdot)$ denote $p_{\hat{\boldsymbol{\theta}}}(\cdot)$ for simplicity, the best activity label sequence $\hat{\mathbf{H}}$ to a feature vector sequence $\mathbf{Y}$ can be decided by decoding the CRF, i.e., solving

$$\hat{\mathbf{H}} = \arg\max_{\mathbf{H}} p(\mathbf{H}|\mathbf{Y}) \tag{18}$$

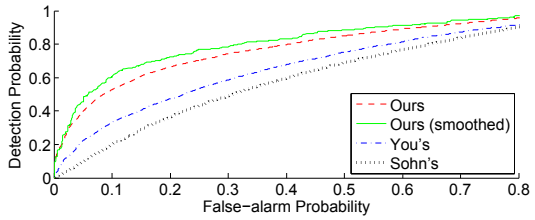
using the well-known Viterbi algorithm similarly to that in a hidden Markov Model. However, we use a soft decision scheme instead of Viterbi algorithm. We employ the Forward-Backward algorithm to calculate $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$ where $q_t = p(H_t = 1|\boldsymbol{y}_{t-C}, \cdots, \boldsymbol{y}_{t+C})$ served as the *a posteriori* of activity of the t-th time frame where C controls the range of the context that is concerned. Then, we actually obtain an activity label sequence $\hat{\mathbf{H}}$ determined by a decision threshold η:

$$\hat{H}_t = \begin{cases} 1: & q_t \geq \eta \\ 0: & q_t < \eta \end{cases}, \tag{19}$$

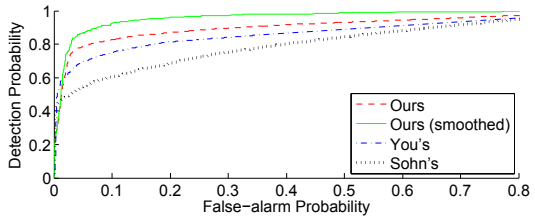
so that a trade-off between detection probability and false alarm probability of VAD can be easily made with a tuned η. In addition, $\hat{\mathbf{H}}$ can be further smoothed for more reasonable global decision.

5. EXPERIMENTS

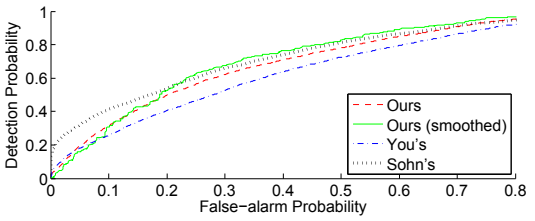
TIMIT [22] corpus is used for the experiments with its word transcription for the VAD evaluation. Three typical noise sources from NOISEX-92 [23] corpus: the factory, white and babble noise, are selected for the simulations of the practical noisy environments. All recorded audio signals are sampled at 16 kHz. We randomly select 128 sentences, 8 sentences (excluding the two dialects) spoken by each of 16 speakers from TIMIT TEST set. 64 sentences from half of the speakers are concatenated as a long utterance with silence of random length added between each two sentences, and the other 64



(a) in factory noise



(b) in white noise



(c) in babble noise

Fig. 2. ROC curves to evaluation of CNSC based VAD under factory (a), white (b) and babble (c) noises, respectively, at SNR $= 0$ dB.

sentences are concatenated in the same manner. These two long utterances are about 338 seconds and 331 seconds long and with 51.6% and 50.3% of speech signals, respectively. We add white noise to the first long utterance at SNR $= -5$ dB served as the only training utterance of the CRF. We add factory, white and babble noise to the second long utterance at SNR $= 0$ dB , respectively, obtaining three noisy utterances to simulate speech signals recorded in real noise environments. The noise adding is implemented by using FaNT [24].

For an audio signal, Short Time Fourier Transform (STFT) is performed with the analysis window of length 32 ms and the window shift of 16 ms. Magnitude spectrum is the magnitude square values of the STFT out. For computational simplicity, the magnitude spectrum for each time frame is reduced to a 22-dimension vector where each element is the average of magnitude spectrum placed on each critical band of Bark frequency scale from 20 Hz to 8000 Hz. The speech dictionary $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ is learned using an online CNSC algorithm [19] from 3696 sentences, 8 sentences spoken by each of 462 speakers from the TIMIT TRAIN set with parameters of $M = 80$, $R = 4$ and $\lambda = 0.01$. The noise dictionary $\{\mathbf{D}^{\mathrm{n}}{}_r\}$ for a certain type of noise is learned using the same algorithm from the noise samples in NOISEX-92 with parameters of $M = 5$, $R = 4$ and $\lambda = 0$.

The CRF is trained independently of noise, using the white noise degraded -5 dB long utterance mentioned above. The reason is twofold. First, once specific noise is fitted by the noise bases (which are learned from the noise itself) then reduced, there is still residual left in the signals. We deem the residual as white-like noise, so that $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ is similar to extracted from signals degraded by white noise. Second, a proper low-SNR training utterance is also benefit of good noise robust. Therefore, we decompose the training utterance only on $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ using CNSC, and use the resulting coefficient vector sequence to train the CRF.

For a noisy utterance, $\{\mathbf{D}^{\mathrm{n}}{}_r\}$ is learned first, and then concatenated with $\{\mathbf{D}^{\mathrm{s}}{}_r\}$ to construct the global dictionary. Then, the decomposition of the utterance is performed according to Eq.(8), and $\mathbf{A}^{\mathrm{s}}$ is used as the input $\mathbf{Y}$ of the CRF. $\boldsymbol{q}$ is calculated with $C = 8$, and $\hat{\mathbf{H}}$ is determined by a given η. We use an optional smoothing post-process where we constrain the detected durations of speech presence or absence are longer than 0.5 secs. The detection probability and false alarm probability of our approach are exploited to evaluate the performance. In addition, Sohn's VAD approach [1] is exploited as the baseline, and You's VAD approach [12] using the same 22-dimension reduced magnitude spectrum as ours is implemented by ourselves for comparison. The ROC curves of our VAD approach (with and without the smoothing post-process) compared with Sohn's and You's under different noises at SNR = 0 dB are illustrated in Fig. 2. It is demonstrated that our approach outperforms the baseline and further improves the VAD performance in factory and white noises conditions. However, our approach degrades in babble noise. This is because that in such conditions the learned noise bases are extremely similar to combinations of some speech bases, which causes much confusion in the decomposition with a sparseness constraint. This issue will be considered in our future work.

6. CONCLUSION

We have proposed a VAD approach using CNSC to improve the detection performance under low SNR noisy conditions by adopting noise-robust feature for speech signal detection *while* reducing noise contribution away. We have learned an over-complete speech dictionary from clean speech using CNSC, as well as a low-rank noise dictionary from noise samples using CNMF. Then, we concatenate the two dictionaries as a global dictionary to decompose a given audio signals using CNSC. We only consider the coefficients corresponding to the speech dictionary as the features of single time frames. A CRF with a linear chain structure is constructed to model the correlation between the features and speech activities along an audio signal, and trained independently of the noise conditions. The experiments demonstrate our approach further improves the performance of VAD in low SNR conditions.

References

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

[2] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 737–740.

[3] J. Ramírez, J.C. Segura, J.M. Górriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2177–2189, 2007.

[4] J. Ramírez, J.C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *Signal Processing Letters, IEEE*, vol. 12, no. 10, pp. 689–692, 2005.

[5] J.W. Shin, H.J. Kwon, S.H. Jin, and N.S. Kim, "Voice activity detection based on conditional map criterion," *Signal Processing Letters, IEEE*, vol. 15, pp. 257–260, 2008.

[6] J. Wu and X.L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *Signal Processing Letters, IEEE*, vol. 18, no. 8, pp. 466–469, 2011.

[7] J. Wu and X.L. Zhang, "Maximum margin clustering based statistical vad with multiple observation compound feature," *Signal Processing Letters, IEEE*, vol. 18, no. 5, pp. 283–286, 2011.

[8] Q.H. Jo, J.H. Chang, JW Shin, and NS Kim, "Statistical model-based voice activity detection using support vector machine," *Signal Processing, IET*, vol. 3, no. 3, pp. 205–210, 2009.

[9] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *Signal Processing Letters, IEEE*, vol. 19, no. 8, pp. 507–510, 2012.

[10] A. Saito, Y. Nankaku, A. Lee, and K. Tokuda, "Voice activity detection based on conditional random fields using multiple features," in *Proc. Interspeech*, 2010, pp. 2086–2089.

[11] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[12] D. You, J. Han, G. Zheng, and T. Zheng, "Sparse power spectrum based robust voice activity detector," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*. IEEE, 2012, pp. 289–292.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 2272–2279.

[14] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.

[15] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[16] P.O. Hoyer, "Non-negative sparse coding," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 557–565.

[17] P.O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[18] W. Wang, "Convolutive non-negative sparse coding," in *Proceedings of the IEEE International Joint Conference on Neural Networks*. IEEE, 2008, pp. 3681–3684.

[19] D. Wang, R. Vipperla, and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," in *Proceedings of the twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] R. Vipperla, J.T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012, pp. 4181–4184.

[21] R. Malouf et al., "A comparison of algorithms for maximum entropy parameter estimation," in *Proceedings of the sixth conference on natural language learning*, 2002, pp. 49–55.

[22] J.S. Garofolo, *TIMIT: Acoustic-phonetic Continuous Speech Corpus*, Linguistic Data Consortium, 1993.

[23] Rice University, *NOISEX-92 Database*, [Online.] Available: http://spib.rice.edu/spib/select_noise.html.

[24] H.-Guenter Hirsch, *FaNT: Filtering and Noise Adding Tool*, Software available at http://aurora.hsnr.de /download.html.