ADAPTIVE EXTRACTION OF REPEATING NON-NEGATIVE TEMPORAL PATTERNS FOR SINGLE-CHANNEL SPEECH ENHANCEMENT

Yinan Li, Xiongwei Zhang, Meng Sun, Gang Min, Jibin Yang

Lab of Intelligent Information Processing, PLA University of Science and Technology, Nanjing, China

ABSTRACT

Estimating unknown background noise from single-channel noisy speech is a key yet challenging problem for speech enhancement. Given the fact that the background noises typically have the repeating property and the foreground speech is sparse and time-variant, many literatures decompose the noisy spectrogram directly in an unsupervised fashion when there is no isolated training example of the target speaker or particular noise types beforehand. However, recently proposed methods suffer from un-interpretable decomposed patterns, neglecting the temporal structure of the background noise or being constrained by the pre-fixed parameters. To settle these issues, we propose a novel method based on autocorrelation technique and convolutive non-negative matrix factorization. The proposed method can adaptively estimate the underlying non-negative repeating temporal patterns from noisy speech and identify the clean speech spectrogram simultaneously. Experiments on NOIZEUS dataset mixed with various real-world background noises showed that the proposed method performs better than some state-of-the-art methods.

Index Terms— Speech enhancement, non-negative repeating temporal patterns, autocorrelation, convolutive non-negative matrix factorization

1. INTRODUCTION

The goal of single-channel speech enhancement is to improve the intelligibility and fidelity of noisy speech by attenuating background noises in a single-channel recording. The difficulty arises from the unpredictable and high variability of real-world interferers. It is particular difficult to impose mathematical constraints on the encountered noise that are both discriminating enough to facilitate good separation, and sufficient flexible to handle unseen noises.

Numerous approaches have been proposed to address this challenging task. Among them, traditional algorithms such as spectral subtraction, Wiener filtering, and statistical-model-based method assume that the background noise is stationary and that the noise can be modeled by a single spectral profile or using a single speech-to-noise ratio (SNR) [1]. These algorithms are popular and widely used as they require neither the identity of the specific speaker nor knowing the noise type. An estimation of the noise ensures these algorithms work. The main problem of these approaches is that they are hard to handle non-stationary noises that are difficult to predict and estimate. Furthermore, complex noises are poorly modeled by a single spectral profile or SNR regardless of what kind of sophisticated models are used.

Another set of techniques that attracts much attention is grounded on compositional models such as non-negative matrix factorization (NMF) [2] and probabilistic component analysis (PLCA) [3]. They are able to figure out the component sources from the mixtures and evaluate their corresponding contributions in a highly interpretable fashion, whether the noise is stationary or not. One of the problems of these approaches is how to relate the obtained components to the ground truth knowledge, i.e. which component corresponds to speech and which one to noise. Although it can be addressed by using the supervised manner, isolated training example becomes essential [4]. This poses serious challenges due to the need of additional information is typically difficult to be satisfied. Moreover, when the actual source fails to cope with the training examples, the performance rapidly decline.

Recently, an emerging paradigm that explores the repetition property of spectrogram has been employed for speech separation and enhancement [5]. The basic premise underlying is that the background noise spectrogram has a repetitive low-rank structure while the foreground speech is time-varying and sparse. The premise is appealing as it requires neither the prior estimation of noise or speech model, nor the stationary assumption of background noise. The recently-developed robust principle component analysis (R-PCA), based on a well-behaved convex optimization, is a proper candidate to provide such a solution [6]. However, the decomposed negative components of RPCA are difficult to interpret. Meanwhile, the Euclidean distance used to define the objective function of RPCA is often criticized for over-emphasizing on the high-energy components [7].

In this paper, we propose a novel method in Section 2 that introduces the adaptive non-negative temporal model to the sparse and low-rank decomposition framework. The proposed method efficiently extracts the underlying unknown

This work is supported by the NSFC(61402519,61471394) and the Natural Science Foundation of Jiangsu Province (BK20140071, BK20140074).

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

background noise from mixture, enhancing the noisy speech simultaneously. We discuss the relationship between our method and the related prior work in Section 3. Section 4 reports the experimental results of the proposed algorithm and the competitive methods. We conclude the paper in Section 5.

2. THE PROPOSED METHOD

The proposed method treats background noise as a concatenation of a small subset of temporal repeated patterns and recasts the problem of noise modeling as estimating the underlying repeated patterns in the present of speech. Fig.1. gives an overview of the proposed method. As we can see in the flowchart, the proposed method can be divided into three steps: identification of the underlying repeating period T (Step1), extraction of the underlying non-negative repeating temporal patterns (Step2), and speech reconstruction (Step3).

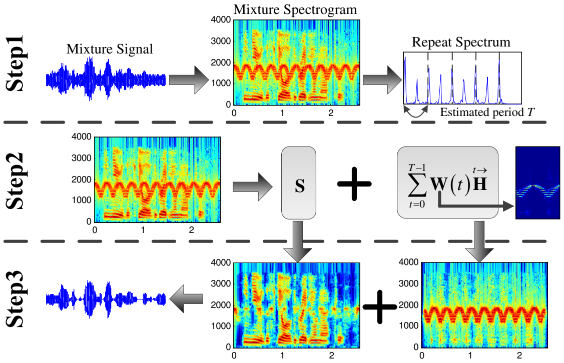


Fig. 1. The flowchart of the proposed framework.

Given the short-time Fourier transform (STFT) of noisy speech, we first derive its magnitude spectrogram $\mathbf{Y}$ which is a non-negative matrix $\mathbf{Y} \in \mathbb{R}_{\geq 0}^{m \times n}$ and retain the phase $\angle \mathbf{Y}$ for time-domain signal reconstruction later.

2.1. Repeating Period Estimation

Autocorrelation is one of the most widely used techniques to detect the underlying period. There are numerous methods dealing with this problem [5], [8]. By operating on the power spectrogram which emphasizes the appearance of peaks of periodicity, we can calculate the acoustic self-similarity vector whose j th element is as follows:

$$a(j) = \sum_{i=1}^{n} \sum_{k=1}^{m-j+1} \frac{Y(i,k)^2 Y(i,k+j-1)^2}{n(m-j+1)} \tag{1}$$

where i, j and k are used to denote the element position in a matrix.

Subsequently, after normalizing by its first term (i.e. $a(j) \leftarrow a(j)/a(1)$), we can predict the repeating period by following the recipe proposed in [5]. The procedure is quite straightforward, and its basic idea is to find which period in the self-similarity vector corresponding to the highest mean accumulated energy over its integer multiples.

2.2. Noise Pattern Extraction and Speech Enhancement

Once the repeating period T in terms of frames is estimated, we use it to reveal the underlying repeating background patterns and to enhance the degraded speech. The basic principle is in conformance with the sparse and low-rank framework [5], [6]. The differences lie in that, the non-negative constraint is imposed on the factorized components and that the generalized Kullback-Leibler (K-L) divergence, which is a better alternation of the Euclidean distance for audio processing, is used to define the objective function. Furthermore, the temporal structure of repeating background noise is also considered, utilizing the convolutive non-negative matrix factorization (CNMF) [9] for modeling the temporal information. The objective function is thus defined as follows:

$$\underset{\mathbf{W},\mathbf{H},\mathbf{S}}{\operatorname{argmin}} D_{KL}\left(\mathbf{Y}|| \sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\to}{\mathbf{H}} + \mathbf{S}\right) + \lambda \|\mathbf{S}\|_1 \tag{2}$$

where $D_{KL}(x||y) = x(\log x - logy) + (y - x)$ and time-sliced $\mathbf{W}(t) \in \mathbb{R}_{\geq 0}^{m \times r}$ is a set of bases (including r convolutive bases to represent the temporal structures of $\mathbf{Y}$) that share the same gain matrix $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times n}$. $\overset{t\to}{(\cdot)}$ is the operation that shifts t columns of the objective matrix to the right while $\overset{\leftarrow t}{(\cdot)}$ analogously shifts to the left. $\mathbf{S}$ is a non-negative sparse matrix. The regularization term that controls the sparsity of decomposition component is defined by ℓ_1 norm. We derive the update algorithms as follows,

$$\mathbf{W}(t) \leftarrow \mathbf{W}(t) \odot \frac{\left\{\mathbf{Y} \oslash \left(\sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\to}{\mathbf{H}} + \mathbf{S}\right)\right\} \overset{t\to}{\mathbf{H}}^{\mathrm{T}}}{\mathbf{1} \cdot \overset{t\to}{\mathbf{H}}^{\mathrm{T}}} \tag{3}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}(t)^{\mathrm{T}} \left\{\overset{\leftarrow t}{\mathbf{Y}} \oslash \left(\overset{\leftarrow t}{\left[\sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\to}{\mathbf{H}}\right]} + \overset{\leftarrow t}{\mathbf{S}}\right)\right\}}{\mathbf{W}(t)^{\mathrm{T}} \cdot \mathbf{1}} \tag{4}$$

$$\mathbf{S} \leftarrow \mathbf{S} \odot \frac{\mathbf{Y}}{(\lambda + 1) \cdot \left(\mathbf{S} + \sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\to}{\mathbf{H}}\right)} \tag{5}$$

where $\odot$ and $\oslash$ denote the element-wised matrix multiplication and division, respectively. Sparsity parameter λ can be viewed as a trade-off between speech distortion and noise reduction. We set r to be 1 by considering the repetition property. By iteratively updating these matrices, the objective function finally converges to a local minimum. The convergence of the algorithm can be proved in a similar way as in our previous work [10]. When the algorithm converges, the repeating temporal pattern of background noise is obtained (i.e. $\mathbf{W}(t)\,,\ t \in [0, T-1]$) and the clean speech spectrogram $\mathbf{S}$ is also estimated.

2.3. Speech Signal Reconstruction

Given the noise spectrogram reconstruction $\sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\rightarrow}{\mathrm{H}}$ and the speech spectrogram estimation $\mathbf{S}$, we use the soft time-frequency mask $\mathbf{M}$, whose element value is in the range of $0 \sim 1$, to further boost the performance of enhancement,

$$\hat{\mathbf{S}} = \mathbf{M} \odot \mathbf{Y} = \frac{\mathbf{S}}{\sum_{t=0}^{T-1} \mathbf{W}(t) \overset{t\rightarrow}{\mathbf{H}} + \mathbf{S}} \odot \mathbf{Y} \qquad (6)$$

Once the soft time-frequency mask is applied, we subsequently reconstruct the time-domain clean speech waveform using the noisy phase and inverse STFT.

3. RELATIONS TO PRIOR WORK

We compare our method with two previously established approaches based on the sparse and low-rank framework, REpeating Pattern Extraction Technique (REPET) [5] and Sparse and Low-rank Non-negative Matrix Factorization (SLNMF) [11]. REPET is a low computational method that uses the median operation to separate the repeating component from the noisy spectrogram. To accommodate to the median operation, the repeating period has to be estimated in the first 1/3 of the total length. Our method overcomes this potential weakness and is feasible once there is repetition. SLNMF is our preliminary attempt to solve the uninterpretable problem of the previous sparse and low-rank decomposition. The objective function of SLNMF is also defined by the generalized K-L divergence which is widely used for the task of speech enhancement. However, the parameter that controls the complexity of background noise (i.e. the rank of NMF used to approximate the low-rank background noise) is prefixed and can't be changed adaptively on the fly according to the practical noise. Meanwhile, estimating the rank of background noise on the presence of speech is a difficult task. The proposed method gets rid of the rank estimation problem and utilizes an adaptive method to estimate the temporal span of the noise pattern (by using the approach presented in section 2.1) as an alternation to model background noises of different complexity. Since repeating patterns of complicated noises tend to span over a relative long time, basis that contains more time-slices is preferred, and vice versa. In this way, the proposed method allows us to deal with the highly time-varying repeating noises that are difficult to handle by state-of-the-art methods.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setting

We evaluate the performance of the proposed approaches for single channel speech enhancement using the NOIZEUS dataset which contains 30 short English sentences spoken by three male and three female speakers. The noisy speech was synthesized by adding clean speech to a variety of noise signals with signal-to-noise ratios (SNRs) at -5dB, 0dB and 5dB. Noises were drawn from NOISEX-92 database [12], highly non-stationary noises used in [13] and other internet resources [1]. Four stationary noises (*pink*, *f16*, *factory 1* and *volvo*) and four non-stationary noises (*frogs*, *computer keyboard*, *helicopter*, *siren*) are included. All signals we used are resampled at 8 kHz and all the spectrograms are computed using a Hamming window of 512 samples (64ms) with a frame shift of 128 points.

4.2. Performance Measures

All the involved methods were evaluated by using two metrics, including the widely used signal-to-distortion ratio (SDR) in BSS-EVAL toolbox [14] and the perceptual evaluation of the speech quality (PESQ) score. For both metrics, higher values mean better performance. To overcome the effect of initialization, the results of proposed method are averaged across 10 different random initializations. Without loss of generality, we report the mean value for each metric on all types of noises.

4.3. Experimental Results

We compare the proposed method to related works such as REPET and SLNMF. It is worth to note that, to make the comparisons fair, the proposed method and REPET share the same parameters to segment the repeating period. The different lies that our method don't need to enforce the period to be shorter than 1/3 of the total length. Similarly, the rank of SLNMF was fixed to 2 in conformance with the parameter settings described in [11]. Besides, all STFT parameters were kept the same with the proposed method. We also compare our method with several traditional unsupervised methods, namely Multi-band Spectral Subtraction (MSS) [15], Minimum Mean Square Error (MMSE) [16] and the KLT subspace algorithm [17].

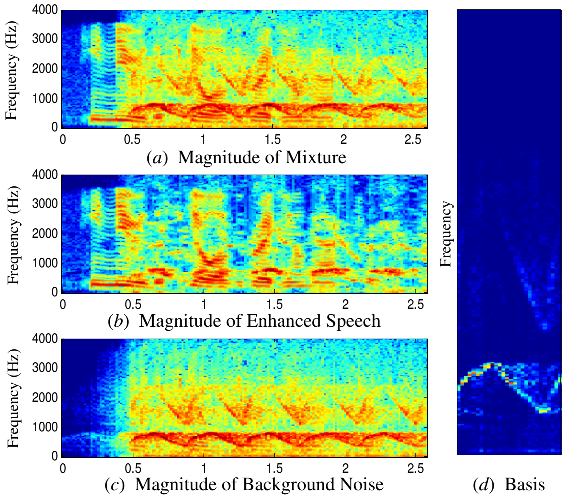
The results of various algorithms are given in Table 1. We can easily see that with respect to almost all metrics, methods based on the sparse and low-rank framework (i.e. REPET, SLNMF and the proposed method) outperforms traditional methods. The reason is partly because the sparse and low-rank decomposition scheme overcomes the stationary assumption of traditional methods to some extent.

[1] $http://www.soundsnap.com/$

Methods	Input SNRs					
	-5dB		0dB		5dB	
	SDR	PESQ	SDR	PESQ	SDR	PESQ
REPET	-0.29	1.02	4.57	1.52	5.68	1.96
SLNMF	0.16	1.52	5.04	2.03	6.27	**2.41**
MSS	-1.43	1.49	3.16	1.67	5.32	2.35
MMSE	-4.64	**1.62**	4.37	1.84	5.15	2.23
KLT	-3.41	1.23	3.64	1.48	5.14	2.01
Proposed	**0.71**	1.57	**5.17**	**2.16**	**6.84**	2.32

Table 1. Average evaluation from the test data results.

By further inspecting of our experimental results, we can see that the proposed method performs almost best among the algorithms using the sparse and low-rank framework. This can be explained as the adaptation strategy provides a more flexible model than SLNMF when dealing with unknown background noise. Besides, the non-negative constraint based model may serve as a more reasonable alternation compared with the median operation used by REPET, though more computational load is introduced.



(*a*) Magnitude of Mixture

(*b*) Magnitude of Enhanced Speech

(*c*) Magnitude of Background Noise

(*d*) Basis

Fig. 2. Example decomposition. The top left panel shows the spectrogram of speech mixed "*siren*" at 0dB. The middle left and bottom left panel shows the spectrogram of separated speech and background noise respectively. The right panel gives the non-negative convolutive basis estimated from the background noise in an adaptive way.

For better illustration, we present an intuitive experiment, as was presented in Figure 2, to show the efficiency of the proposed method.

From Figure 2 (a), we can see that the speech signal is corrupted by "*siren*". It is obvious to see that the spectrogram of "*siren*" is not stationary and we can also observe a large part of overlapping between the speech and "*siren*" spectrogram in both the low-frequency and the high-frequency components. Hence, it is a challenging task to separate speech directly from the noisy observation when we know nothing in advance about the speaker identity or the spectrogram structure of "*siren*".

Metrics	Methods		
	REPET	SLNMF	Proposed
SDR	3.49	3.69	**4.71**
PESQ	2.12	2.23	**2.34**

Table 2. Experimental results when dealing with "*siren*".

The proposed method treats "*siren*" as a kind of unknown repeating background noise whose pattern spans over a certain period of time and we use the autocorrelation technique to detect the period length in terms of frames. Subsequently, the estimated length is used as the number of time-slices of convolutive basis of CNMF to capture the temporal information of background noises. This can be viewed as alternation of the global low-rank approximation. We use the derived algorithms to iteratively estimate the underlying background noise and speech. Once the algorithm converges, the repeating pattern of background noise is estimated as is shown in the Figure. 2 (d). We can easily observe that both the length of pattern and the estimated convolutive basis match well with basic temporal spectrogram structure of "*siren*". Moreover, it is easy to see that the background noise and the foreground speech are largely separated as were presented in Figure. 2 (b) and (c).

We also present the performance on speech enhancement with "*siren*" noise and compare them with related works. These results demonstrated the advantage of the proposed method when dealing with background noise possessing the temporal repeating structures.

5. CONCLUSION

In this paper, an unsupervised single-channel speech enhancement method that can adaptively estimate the unknown repeating temporal patterns was proposed and evaluated. The proposed method introduces the non-negative temporal repeating patterns to the sparse and low-rank framework. The proposed method overcomes the problem of choosing parameters beforehand and can adaptively adjust to cope with various noise inferences without any prior training. Besides, by taking the advantage of non-negative model, potential repetitive temporal spectral regularities underlying in the noisy speech can easily be discovered. Experimental results showed that the proposed method performs better than traditional enhancement method as well as competitive methods such as REPET and SLNMF.

6. REFERENCES

[1] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] Zhiyao Duan, Gautham J Mysore, and Paris Smaragdis, "Online PLCA for real-time semi-supervised source separation," in *Latent Variable Analysis and Signal Separation*, pp. 34–41. Springer, 2012.

[4] Paris Smaragdis, Cédric Févotte, Gautham J Mysore, Nasser Mohammadiha, and Matthias Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[5] Zafar Rafii and Bryan Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73–84, 2013.

[6] Po-Sen. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Sing-voice separation from monaural recording using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 57–60.

[7] Tuomas Virtanen, Jort Florent Gemmeke, Bhiksha Raj, and Paris Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[8] Zafar Rafii and Bryan Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (I-CASSP)*. IEEE, 2011, pp. 221–224.

[9] Paris Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[10] Yinan Li, Xiongwei Zhang, Meng Sun, and Gang Min, "Unsupervised monaural speech enhancement using robust NMF with low-rank and sparse constraints," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. IEEE, 2015, pp. 1–4.

[11] Meng Sun, Yinan Li, Jort Gemmeke, and Xiongwei Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank N-MF with Kullback-Leibler divergence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1233–1242, 2015.

[12] Andrew Varga and Herman JM 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, 1993.

[13] Zhiyao Duan, Gautham J Mysore, and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments.," in *INTERSPEECH*, 2012.

[14] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] Sunil Kamath and Philipos Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE international conference on acoustics speech and signal processing*. IEEE, 2002, vol. 4, pp. 4164–4164.

[16] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[17] Yi Hu and Philipos C Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.