ONLINE REPET-SIM FOR REAL-TIME SPEECH ENHANCEMENT

Zafar Rafii

Northwestern University
EECS Department
Evanston, IL, USA

Bryan Pardo

Northwestern University
EECS Department
Evanston, IL, USA

ABSTRACT

REPET-SIM is a generalization of the REpeating Pattern Extraction Technique (REPET) that uses a similarity matrix to separate the repeating background from the non-repeating foreground in a mixture. The method assumes that the background (typically the music accompaniment) is dense and low-ranked, while the foreground (typically the singing voice) is sparse and varied. While this assumption is often true for background music and foreground voice in musical mixtures, it also often holds for background noise and foreground speech in noisy mixtures. We therefore propose here to extend REPET-SIM for noise/speech segregation. In particular, given the low computational complexity of the algorithm, we show that the method can be easily implemented online for real-time processing. Evaluation on a data set of 10 stereo two-channel mixtures of speech and real-world background noise showed that this online REPET-SIM can be successfully applied for real-time speech enhancement, performing as well as different competitive methods.

Index Terms— Blind source separation, real-time, repeating patterns, similarity matrix, speech enhancement

1. INTRODUCTION

Speech enhancement is the process of improving intelligibility and/or quality of a speech signal, generally when degraded by a noise signal [1]. Applications are numerous, and include speech amplification (e.g., in hearing aids), speech recognition (e.g., in speech-to-text softwares), and speech transmission (e.g., in mobile phones). Since they are generally intended for real-time applications, most of the algorithms for speech enhancement are online algorithms.

According to [1], traditional approaches for speech enhancement can be divided into four categories: spectral subtraction, Wiener filtering, minimum mean square error estimation, and subspace algorithms. Somewhat inspired by source separation techniques, recent methods have also been proposed based on Non-negative Matrix Factorization (NMF) [2] and Probabilistic Latent Component Analysis (PLCA) [3]. When multiple channels are available (e.g., in a two-channel mixture), spatial information can also be exploited in addition to temporal and spectral information, for example by using Independent Component Analysis (ICA) [4] or the Degenerate Unmixing Estimation Technique (DUET) [5]. Most of the methods for speech enhancement require a prior estimation of the noise model [3], and sometimes of the speech model as well [2].

Recently, the REpeating Pattern Extraction Technique (REPET) was proposed to separate the repeating background (typically the music accompaniment) from the non-repeating foreground (typically the singing voice) in musical mixtures [6, 7]. The basic idea is to identify the repeating elements in the audio, compare them to repeating models derived from them, and extract the repeating patterns via time-frequency masking. While the original REPET (and its extensions) assumes that repetitions happen periodically [6, 8, 7], REPET-SIM, a generalization of the method that uses a similarity matrix was further proposed to handle structures where repetitions can also happen intermittently [9]. The only assumption is that the repeating background is dense and low-ranked, while the non-repeating foreground is sparse and varied.

Repetitions happen in music, but in audio in general. In particular in noisy mixtures, the background noise can often exhibit a dense and low-ranked structure, while the signal of interest exhibits a sparse and varying structure. Under this assumption, REPET-SIM then appears as a justifiable candidate for noise/speech segregation. In particular, given the low computational complexity of the algorithm, the method can be easily implemented online for real-time speech enhancement. The advantages of this online REPET-SIM are that it can (obviously) work in real-time, it is very simple to implement, it does not require any pre-trained model (unlike [2] or [3]), it can deal with non-stationary noises (unlike spectral subtraction or Wiener filtering), and it can work with single-channel mixtures (unlike ICA or DUET).

The rest of this article is organized as follows. In Section 2, we first present an online implementation of the REPET-SIM method. In Section 3, we then evaluate the system for real-time speech enhancement, on a data set of 10 stereo two-channel mixtures of speech and real-world background noise, compared with different competitive methods. In Section 4, we conclude this article.

2. METHOD

2.1. REPET-SIM

REPET-SIM is a generalization of the REPET method for separating the repeating background from the non-repeating foreground in a mixture. The REPET approach is based on the idea that repetition is a fundamental element for generating and perceiving structure. In music for example, pieces are often composed of an underlying repeating structure (typically the music accompaniment) over which varying elements are superimposed (typically the singing voice). The basic idea is to identify the repeating elements in the audio, compare them to repeating models derived from them, and extract the repeating patterns via time-frequency masking [6, 8, 9, 7].

Specifically, REPET-SIM identifies the repeating elements in the audio by using a similarity matrix [9]. The similarity matrix is a two-dimensional representation where each bin (a, b) measures the (dis)similarity between any two elements a and b of a given sequence, given some metric. Since repetition/similarity is what makes the structure, a similarity matrix calculated from an audio signal can help to reveal the structure that underlies it [10]. Assuming that the repeating background is dense and low-ranked and the non-repeating foreground is sparse and varied, the repeating elements unveiled by the similarity matrix should then be those that basically make the repeating background.

Given the Short-Time Fourier Transform (STFT) X of a mixture, REPET-SIM first derives its magnitude spectrogram V. It then computes a similarity matrix S from V using the cosine similarity, and identifies for every time frame j in V, the frames j_k's that are the most similar to frame j using S. It then derives a repeating spectrogram model U by taking for every frame j in V, the element-wise median of the corresponding similar frames j_k's. It then refines the repeating spectrogram model U into W by taking the element-wise minimum between U and V, and derives a soft time-frequency mask M by normalizing W by V, element-wise. It finally derives the STFT of the estimated repeating background by symmetrizing M and applying it to the STFT of the mixture X [9].

While originally developed for separating a repeating background from a non-repeating foreground in musical mixtures, REPET-SIM appears as a justifiable candidate for noise/speech segregation. Indeed, in noisy mixtures, the background noise often exhibits a dense and low-ranked structure, while the signal of interest exhibits a sparse and varying structure.

2.2. Online Implementation

Given the low computational complexity of the algorithm, REPET-SIM can be easily implemented online for real-time processing. The online implementation simply implies processing the time frames of the mixture one by one, by using a sliding buffer that temporally stores past frames, given a maximal buffer size.

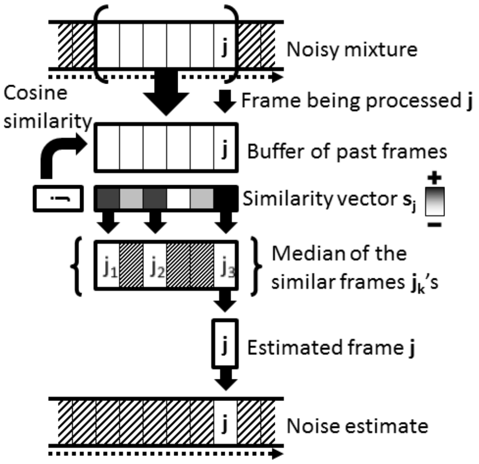

Fig. 1. Overview of the online REPET-SIM system.

Given a time frame of the STFT X of a mixture, we first derive its magnitude spectrum. We then calculate the cosine similarity between the frame being processed j and the B past frames, $j - B - 1, j - B - 2, \ldots$ and j, that were temporally stored in a buffer of maximal size b seconds (or B frames). We obtain a similarity vector s_j.

We then identify in the buffer, the frames j_k's ($\leq B$) that are the most similar to the frame being processed j using s_j, and we take their median for every frequency channel. We obtain an estimated frame for the noise. We then refine this estimated frame by taking the minimum between the estimated frame and the frame being processed j, for every frequency channel (see also [9]).

We finally synthesize the time frame for the STFT of the noise by mirroring the frequency channels and using the phase of the corresponding time frame of the STFT of the mixture. After inversion in the time domain, the speech signal is simply obtained by subtracting the background noise from the mixture signal. If the mixture is multichannel, the channels are processed independently.

3. EVALUATION

3.1. Data Set

The Signal Separation Evaluation Campaign (SiSEC) proposes a source separation task for two-channel mixtures of speech and real-world background noise[1]. We used the "development" data (*dev*), given that the original speech and noise signals were provided. We excluded the second part (domestic environment) because the recordings were too short ($\approx$ 1 second). Our data set then consists of 10 two-channel

[1] http://sisec.wiki.irisa.fr/tiki-index.php?page=Two-channel+mixtures+of+speech+and+real-world+background+noise

mixtures of one speech source and real-world background noise, of 10 second length and 16 kHz sampling frequency.

The background noise signals were recorded via a pair of microphones in different public environments (subway (*Su1*), cafeteria (*Ca1*), and square (*Sq1*)), and in different positions (center (*Ce*) and corner (*Co*)). Several recordings were made in each case (*A* and *B*), by adding a speech signal (male or female) to the background noise signal.

3.2. Competitive Methods

For the given data set, SiSEC featured the following systems:

- *Algorithm 5* is based on a first constrained ICA that estimates the mixing parameters of the target source, followed by a Wiener filtering to enhance the separation results [4].

- *Algorithm 8* is based on a first estimation of the noise from the unvoiced segments, followed by DUET [5] and spectral subtraction to refine the results, and a minimum-statistics-based adaptive procedure to refine the noise estimate [11].

- *Baseline* is based on a first estimation of the Time Differences Of Arrival (TDOA) of the sources, followed by a maximum likelihood target and noise variance estimation under a diffuse noise model, and a multichannel Wiener filtering [12]; this is the baseline algorithm proposed by SiSEC.

REPET-SIM is the proposed online method. The STFT was calculated using half-overlapping Hamming windows of 1024 samples, corresponding to 64 milliseconds at 16 kHz. The parameters of the algorithm were fixed as follows [9]: maximum number of repeating frames $k = 20$; minimum similarity between a repeating frame and the given frame $t = 0$; minimum distance between two consecutive repeating frames $d = 0.1$ second; and maximal buffer size $b = 2$ seconds ($B \approx 30$ frames). Pilot experiments showed that those parameters lead to overall good noise/speech segregation results.

SiSEC also featured *Algorithm 6* which is the same as *Algorithm 5* but with different settings, and *STFT Ideal Binary Mask* which represents the binary masks providing maximum SDR. We do not report their results, since *Algorithm 5* seems slightly better than *Algorithm 6*, and *STFT Ideal Binary Mask* is strictly better than all the methods. More details about the competitive methods and their results can be found online[2].

3.3. Performance Measures

The BSS_EVAL toolbox proposes a set of measures that intend to quantify the quality of the separation between a source and its estimate. The principle is to decompose the estimate of a source into contributions corresponding to the target source, the spatial distortion (if multichannel source), the interference from unwanted sources, and the artifacts related with additional noise. Based on this principle, the following measures were defined (in dB): source Image to Spatial distortion Ratio (ISR), Source to Interference Ratio (SIR), Sources to Artifacts Ratio (SAR), and finally Signal to Distortion Ratio (SDR) which measures the overall error [13].

Based on a similar principle, the PEASS toolkit proposes a set of new measures that were shown to be better correlated with human assessment of signal quality. The following measures were defined: Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), Artifacts-related Perceptual Score (APS), and finally Overall Perceptual Score (OPS) which measures the overall error [14].

3.4. Experimental Results

		dev_Su1_Ce_A		*dev_Su1_Ce_B*	
		sim	noi	sim	noi
REPET-SIM	SDR	-0.5	**15.4**	**5.2**	**14.1**
	OPS	15.9	**31.3**	30.7	**22.4**
Algorithm 5	SDR	**0.9**	5.7	-2.3	1.8
	OPS	**21.7**	10.0	**33.6**	9.7
Algorithm 8	SDR	-7.8	8.1	-0.7	8.2
	OPS	13.4	12.4	32.2	20.1
Baseline	SDR	-5.0	10.9	0.5	9.4
	OPS	20.5	29.9	28.9	18.3

Table 1. SDR (dB) and OPS results for the subway noises.

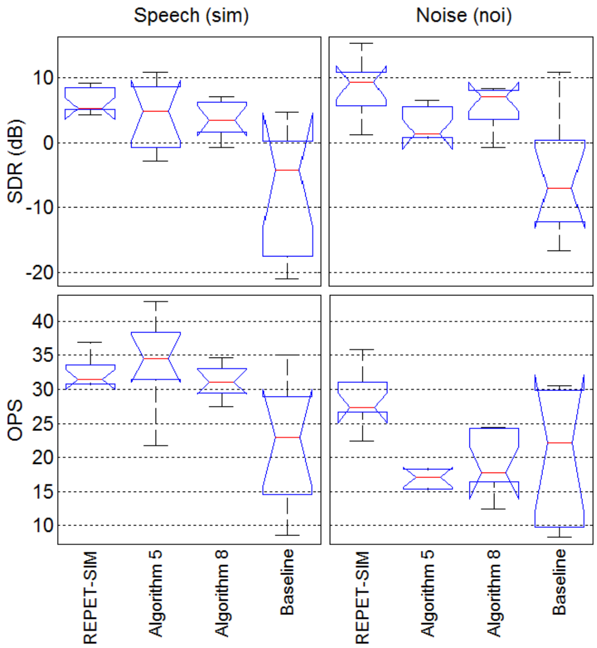

Fig. 2. SDR (dB) and OPS distributions for all the noises.

Tables 1, 2, and 3 show the results for the SDR (dB) and OPS, for the stereo speech estimates (sim) and stereo noise estimates (noi), for all the methods, respectively for the subway noises, the cafeteria noises, and the square noises. Figure 2 shows the distributions for all the noises. As we can

[2] http://www.irisa.fr/metiss/SiSEC11/noise/results_dev.html

		dev_Ca1_Ce_A		*dev_Ca1_Ce_B*		*dev_Ca1_Co_A*		*dev_Ca1_Co_B*	
		sim	noi	sim	noi	sim	noi	sim	noi
REPET-SIM	SDR	**5.4**	**1.3**	8.0	**3.7**	**9.2**	**5.6**	**9.2**	**5.6**
	OPS	33.6	23.6	23.7	**31.0**	30.7	**26.6**	30.7	**26.6**
Algorithm 5	SDR	4.7	0.8	**10.9**	2.8	5.1	0.8	5.1	0.8
	OPS	**42.9**	**24.0**	**35.4**	25.3	**31.4**	17.1	**31.4**	17.1
Algorithm 8	SDR	3.4	-0.8	6.3	2.1	7.1	3.6	7.1	3.6
	OPS	34.6	18.1	27.5	24.3	31.1	24.4	31.1	24.4
Baseline	SDR	0.3	-3.9	4.7	0.4	-3.5	-7.0	-3.5	-7.0
	OPS	8.9	9.7	33.1	27.8	22.9	8.3	22.9	8.3

Table 2. SDR (dB) and OPS results for the cafeteria noises.

		dev_Sq1_Ce_A		*dev_Sq1_Ce_B*		*dev_Sq1_Co_A*		*dev_Sq1_Co_B*	
		sim	noi	sim	noi	sim	noi	sim	noi
REPET-SIM	SDR	**4.4**	**9.1**	5.1	**9.5**	**5.1**	**10.7**	8.6	**10.8**
	OPS	32.9	**27.1**	32.1	**27.4**	34.1	**35.8**	36.9	**31.1**
Algorithm 5	SDR	-0.8	0.8	**8.7**	5.5	-2.8	0.8	**10.8**	6.5
	OPS	**38.4**	15.3	26.9	15.8	**36.5**	17.3	**42.6**	18.3
Algorithm 8	SDR	1.7	6.5	3.4	7.8	2.2	7.8	6.0	8.3
	OPS	30.3	17.4	**33.0**	16.4	29.4	14.0	34.4	17.0
Baseline	SDR	-21.1	-16.4	-21.1	-16.7	-17.5	-12.0	-14.4	-12.2
	OPS	23.6	25.9	8.6	17.9	35.0	30.5	14.5	29.9

Table 3. SDR (dB) and OPS results for the square noises.

see, *REPET-SIM* does almost always better than *Algorithm 8* and *Baseline*, and performs as well as *Algorithm 5*, sometimes getting better results, especially for the noise estimates. This makes sense, since REPET-SIM only models the noise.

Multiple comparison tests showed that, for the SDR, *REPET-SIM* is significantly better only when compared with *Baseline*, for both the speech and noise estimates. For the OPS, there is no significant difference between the different methods for the speech estimates; however *REPET-SIM* is significantly better than all the other methods for the noise estimates. We used a (parametric) analysis of variance (ANOVA) when the distributions were all normal, and a (non-parametric) Kruskal-Wallis test when at least one of the distributions was not normal. We used a Jarque-Bera normality test to determine if a distribution was normal or not. The online REPET-SIM was implemented in Matlab on a PC with Intel Core i7-2600 CPU of 3.40 GHz and 12.0 GB of RAM.

4. CONCLUSION

We have presented an online implementation of REPET-SIM, a generalization of the REPET method that uses a similarity matrix to separate the repeating background from the non-repeating foreground in a mixture. The method only assumes that the background noise is dense and low-ranked, while the speech signal is sparse and varied.

Evaluation on a data set of 10 stereo two-channel mixtures of speech and real-world background noise showed that this online REPET-SIM can be successfully applied for real-time speech enhancement, performing as well as different methods, while being computationally efficient.

Audio examples and source codes can be found online[3]. This work was supported by NSF grant number IIS-0812314.

5. RELATION TO PRIOR WORK

Traditional techniques for speech enhancement do not explicitly use the analysis of the repeating structure as a basis for noise/speech segregation [11, 1]. Most of the methods also require prior estimation of the noise model and/or speech model [2, 3]. Other methods require the availability of multiple channels [4, 12]. REPET-SIM is a method that was originally proposed for separating a music background from a voice foreground in musical mixtures, based on the assumption that the background is dense and low-ranked, and the foreground is sparse and varied. We proposed here to extend such assumption for background noise and foreground speech, and developed an online version of REPET-SIM that can be applied for real-time speech enhancement. The advantages of such a method are: it can (obviously) work in real-time, it is very simple to implement, it does not need any pre-trained model, it can deal with non-stationary noises, and it can work with single-channel mixtures.

[3] http://music.cs.northwestern.edu/research.php?project=repet

6. REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2007.

[2] Alexey Ozerov and Emmanuel Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, Florence, Italy, September 1 2011, pp. 86–87.

[3] Zhiyao Duan, Gautham J. Mysore, and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *13th Annual Conference of the International Speech Communication Association*, Portland, OR, USA, September 9-13 2012.

[4] Francesco Nesta and Marco Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, Florence, Italy, September 1 2011, pp. 18–23.

[5] Özgür Yilmaz and Scott Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] Zafar Rafii and Bryan Pardo, "A simple music/voice separation system based on the extraction of the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 22-27 2011.

[7] 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. 71–82, January 2013.

[8] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30 2012.

[9] Zafar Rafii and Bryan Pardo, "Music/voice separation using the similarity matrix," in *13th International Society for Music Information Retrieval*, Porto, Portugal, October 8-12 2012.

[10] Jonathan Foote, "Visualizing music and audio using self-similarity," in *ACM Multimedia*, Orlando, FL, USA, October 30-November 5 1999, pp. 77–80.

[11] Sundarrajan Rangachari and Philipos C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, February 2006.

[12] Charles Blandin, Alexey Ozerov, and Emmanuel Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, August 2012.

[13] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *7th International Conference on Independent Component Analysis and Signal Separation*, London, UK, September 9-12 2007.

[14] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, September 2011.