38th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)

May 26-31, 2013, Vancouver, BC, Canada

Online REPET-SIM for Real-time Speech Enhancement

Zafar Rafii & Bryan Pardo

Northwestern University, EECS Department, Evanston, IL, USA.
zafarrafii@u.northwestern.edu • pardo@northwestern.edu • http://music.cs.northwestern.edu/

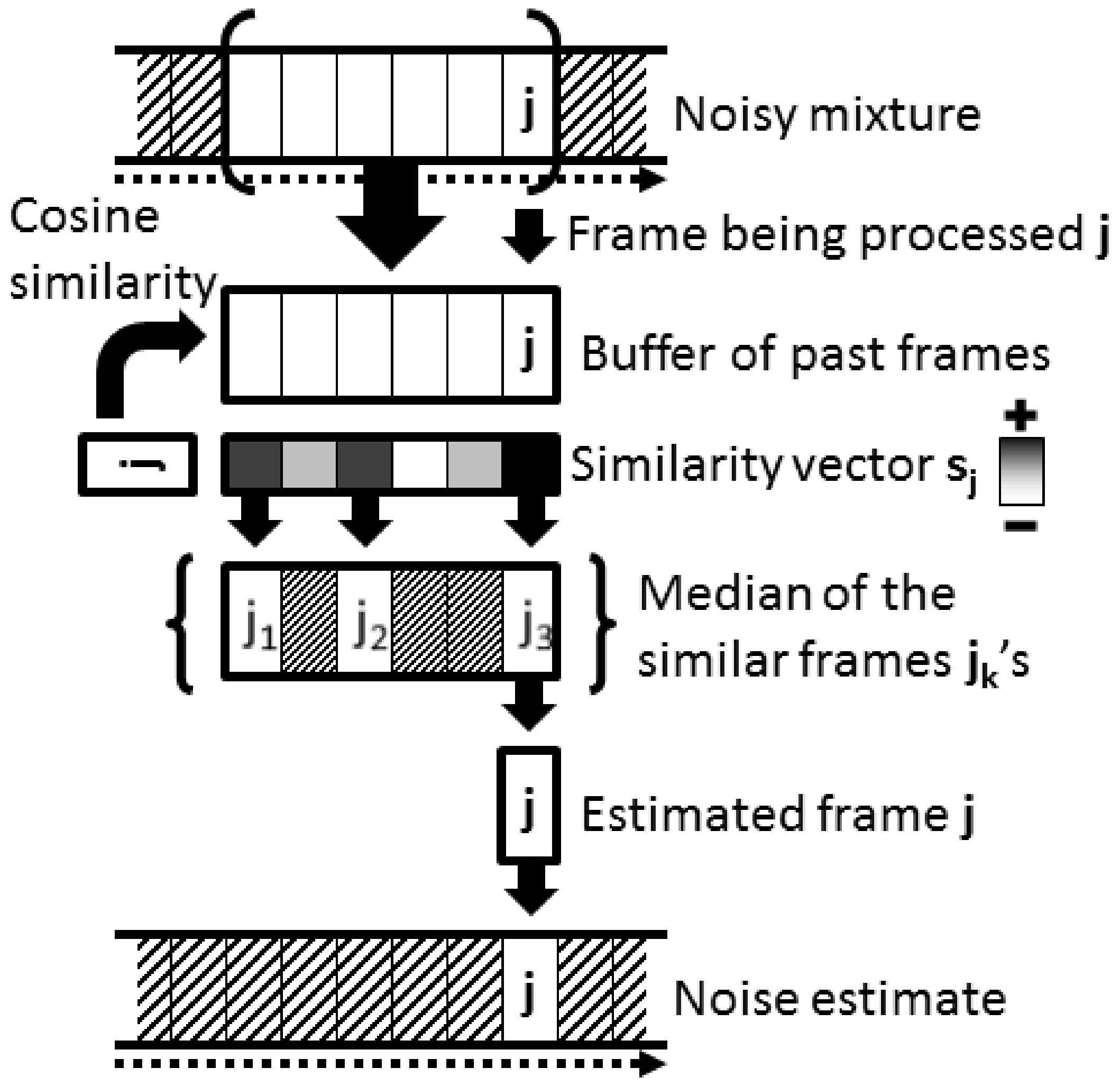
Introduction

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 is dense and low-ranked, while the foreground 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. Given the low computational complexity of the algorithm, we then show that the method can be easily adapted online for real-time speech enhancement.

Method

For every time frame being processed j in the magnitude spectrogram of a noisy mixture signal:

- **Step 1: Identify the repeating elements**
  - Compute the cosine similarity between frame j and the B past frames ($j-B-1$, $j-B-2$, $\ldots$ and j) that have been stored in a buffer, and get the similarity vector s_j
  - Identify the k ($\leq B$) past frames j_k's that are the most similar to frame j using s_j
- **Step 2: Derive a repeating model**
  - Take the median of the k identified past frames j_k's (for every frequency channel), and get an estimated frame for the noise
  - Take the minimum between the estimated frame and frame j (for every frequency channel), and get a refined estimated frame for the noise
- **Step 3: Extract the repeating structure**
  - Synthesize the estimated noise signal from the estimated frames using the phase of the noisy mixture signal
  - Subtract the estimated noise signal from the noisy mixture signal, and get an enhanced speech signal



Noisy mixture

Cosine similarity

Frame being processed j

Buffer of past frames

Similarity vector s_j

Median of the similar frames j_k's

Estimated frame j

Noise estimate

Figure : Overview of the online REPET-SIM.

Evaluation

Signal Separation Evaluation Campaign (SiSEC)[1]:

- **Data set**
  - 10 two-channel mixtures of speech and real-world background noise (10 second length)
  - different noise signals recorded via a pair of microphones in different environments (subway, cafeteria, and square), and different positions (center and corner)
  - different speech signals added to the noise signals (male and female)
- **Competitive (online) methods**[2]
  - *Algorithm 5*: Independent Component Analysis (ICA) + Wiener filtering
  - *Algorithm 8*: Degenerate Unmixing Estimation Technique (DUET) + spectral subtraction
  - *Baseline*: Time Differences Of Arrival (TDOA) of the sources + Wiener filtering
- **Performance measures** (higher = better)
  - *BSS Eval*: Signal to Distortion Ratio (SDR)
  - *PEASS*: Overall Perceptual Score (OPS)

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

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

Results

Speech/noise separation performance on 10 two-channel mixtures of speech and real-world background noise, compared with 3 different competitive online methods, using SDR (dB) and OPS:

- **Table 1: 2 noisy mixtures simulated in a subway**
(1 noise signal × 1 mic position × 2 speech signals)

		mix 11		mix 12	
		speech	noise	speech	noise
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 2: 4 noisy mixtures simulated in a cafeteria**
(1 noise signal × 2 mic positions × 2 speech signals)

		mix 11		mix 12		mix 21		mix 22	
		spe	noi	spe	noi	spe	noi	spe	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 3: 4 noisy mixtures simulated in a square**
(1 noise signal × 2 mic positions × 2 speech signals)

		mix 11		mix 12		mix 21		mix 22	
		spe	noi	spe	noi	spe	noi	spe	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

Analysis

- **Comparative results**
  - Number of "wins" for each of the methods:

	SDR		OPS	
	speech	noise	speech	noise
REPET-SIM	**6**	**10**	0	**9**
Algorithm 5	4	0	**9**	1
Algorithm 8	0	0	1	0
Baseline	0	0	0	0

  - *REPET-SIM* got higher SDRs for both speech and noise, and higher OPS's for noise only

- **Statistical analysis**
  - Difference in medians given a sign test between *REPET-SIM* and each of the other methods:

	SDR		OPS	
	speech	noise	speech	noise
Algorithm 5	no	yes	(yes)	yes
Algorithm 8	yes	yes	no	yes
Baseline	yes	yes	no	yes

  - *REPET-SIM* got statistically higher SDRs, except for speech compared with *Algorithm 5*, and statistically higher OPS's for noise only

- **Relation to prior work**

Unlike traditional techniques in real-time speech enhancement, the online REPET-SIM:

- does not need any pre-trained model
- can deal with non-stationary noises
- works with single-channel mixtures

Conclusion

Evaluation on 10 two-channel mixtures of speech and real-world background noise showed that the online REPET-SIM can be successfully applied for real-time speech enhancement, performing as well as different competitive methods. Audio examples and source codes can be found online[3]. This work was supported by NSF grant numbers IIS-0812314 and IIS-1116384.

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