CHiME Data Separation Based on Target Signal Cancellation and Noise Masking

Zbyněk Koldovský, Jiří Málek and Jan Nouza
Technical University of Liberec, Studentská 2,
461 17 Liberec, Czech Republic
zbynek.koldovsky@tul.cz
jiri.malek@tul.cz, jan.nouza@tul.cz

Miroslav Balík
Brno University of Technology,
Purkyňova 464/118, 612 00 Brno,
Czech Republic
balik@feec.vutbr.cz

***Abstract*—The task of the CHiME challenge is to separate distant speech from noise and recognize the commands being spoken. We propose a separation approach suitable for the CHiME data that relies on the existence of a speech-only recording and assumes fixed positions of the speaker and hearer and stationary reverberant conditions. Using the clean speech segment, a filter that suppresses the target speech is designed. Then, its output provides an estimate of the noise, when the noise is present and comes from different positions. The estimated noise is suppressed from original recordings by an adaptive filter, which outputs the enhanced target signal. The experiments with CHiME data show that this approach improves the keyword recognition accuracy by 7–27%.**

I. Introduction

PASCAL CHiME speech separation and recognition challenge (CHiME) considers the problem of distant speech recognition in a noisy environment. The task is to process a dataset that consists of short utterances of a speaker who is standing in a fixed position that is about 2 meters distant from a dummy head. The utterances are short commands of a fixed structure and are mixed with natural background noises at various signal-to-noise levels. The task to recognize the utterances posses a challenge for several signal processing disciplines such as audio source separation, feature extraction, or robust speech recognition [1].

In this paper, we focus on the separation of the target speech signal from noise, which can be considered as a part of a more complex solution. We apply a simple separation approach that is based on the fact that the positions of the speaker and of the dummy head are fixed. Specifically, using a short noise-free recording of the speaker, we design a time-invariant MISO filter that, applied to signals from microphones, suppresses the source coming from the position of the speaker. Since the recordings are obtained by two microphones (the binaural setup), the non-target sources remain in the filtered signal, which provides an estimate of the noise. Then, the estimated noise is suppressed from original recordings by a time-variant filter (masking), which outputs the enhanced target signal.

Many two-microphone noise reduction techniques have already been proposed [2]. The basic idea considered here was already used, e.g., in [3] or [4] or in much earlier works cited therein. Compared to [3], we estimate the noise signals using time-invariant filter, which is effective on CHiME data thanks to the fixed position of the speaker. Next, we apply the adaptive (masking) filter to the sum of channels, that is, not to each channel separately as in [3], which was used to preserve the binaural hearing allowing source localization. Since in CHiME datasets the speaker is standing directly in front of the dummy head, the sum of both channels already slightly improves the signal-to-noise ratio.

We evaluate the performance of the separation in terms of two standard measures, i.e., by the Signal-to-Noise Ratio (SNR) and Signal-to-Distortion Ratio (SDR). The performance is also measured in terms of keyword recognition accuracy achieved by a baseline recognizer provided within the CHiME challenge. The recognizer is originally trained on a database of noise-free data. On a testing set of noise-free data it achieves the keyword recognition accuracy about 93%, while on the noisy data it achieves 31–83% depending on the noise level. We show that the recognition of enhanced noisy signals by the approach proposed here is improved by up to 27%.

The following section describes the problem and the proposed separation approach for the task. The experimental evaluations and comparisons conducted on CHiME data are provided by Section III. Section IV concludes the paper and outlines further perspectives.

II. The Model and Solution

A CHiME recording can be described by

$$\begin{aligned} x_{\mathrm{L}}(n) &= \{h_{\mathrm{L}} * s\}(n) + y_{\mathrm{L}}(n), \\ x_{\mathrm{R}}(n) &= \{h_{\mathrm{R}} * s\}(n) + y_{\mathrm{R}}(n) \end{aligned} \tag{1}$$

where n is the time index, $*$ denotes the convolution, $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are, respectively, the signals from the left and right microphone, $s(n)$ is the speaker's signal, and $y_{\mathrm{L}}(n)$ and $y_{\mathrm{R}}(n)$ are noise signals that are usually strongly correlated or even correspond to one noise signal only that is differently filtered on each channel. The microphone-source impulse responses denoted by $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ are the same in all recordings since the positions of the microphones and of the speaker are fixed.

A. Noise Estimation by Target Signal Cancellation

To cancel the target signal $s(n)$ by filtering and suppressing $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$, the need is, generally, to find filters $g_{\mathrm{L}}(n)$

and $g_{\mathrm{R}}(n)$ such that

$$\{g_{\mathrm{L}} * h_{\mathrm{L}}\}(n) = \{g_{\mathrm{R}} * h_{\mathrm{R}}\}(n), \tag{2}$$

because then the difference

$$\{g_{\mathrm{L}} * x_{\mathrm{L}}\}(n) - \{g_{\mathrm{R}} * x_{\mathrm{R}}\}(n) = \{g_{\mathrm{L}} * y_{\mathrm{L}}\}(n) - \{g_{\mathrm{R}} * y_{\mathrm{R}}\}(n) \tag{3}$$

does not contain the contribution of $s(n)$ and provides information about the noise signals.

To find the filters $g_{\mathrm{L}}(n)$ and $g_{\mathrm{R}}(n)$, a priori information is needed. Here, we consider the situation when a noise-free recording of the target is available, i.e., when $y_{\mathrm{L}}(n) = y_{\mathrm{R}}(n) = 0$. Then, $g_{\mathrm{L}}(n)$ and $g_{\mathrm{R}}(n)$ can be searched by solving

$$\{g_{\mathrm{L}} * x_{\mathrm{L}}\}(n) - \{g_{\mathrm{R}} * x_{\mathrm{R}}\}(n) = 0. \tag{4}$$

The solution of (4) is not unique. For example, $g_{\mathrm{L}}(n)$ and $g_{\mathrm{R}}(n)$ could be, respectively, chosen as the inverse filters of $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$, which dereverberate the target signal $s(n)$. However, such filters are usually not causal, and they seriously modify the spectra of the noise signals in (3), thereby degrading the quality of the noise estimate.

The need is that the noise estimates provided by (3) are as close to $y_{\mathrm{L}}(n)$ or $y_{\mathrm{R}}(n)$ as possible (in order to suppress them from (1) in the further processing stage). This could be formulated so that $g_{\mathrm{L}}(n)$ and $g_{\mathrm{R}}(n)$ are close to the unit impulse function $\delta(n)$ as much as possible. To this end, the fact that $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ do not differ much, because the microphones (ears) are close to each other, can be used.

For example, in [3], $g_{\mathrm{L}}(n)$ is put equal to $\delta(n)$, and (4) is optimized subject to $g_{\mathrm{R}}(n)$ only, and vice versa[1]. It means that the target signal in one microphone is equalized to have the same response as in the other microphone.

In case of the CHiME scenario, we propose to find an equalizing filter $g(n)$ such that it satisfies the condition

$$\{g * x_{\mathrm{L}}\}(n) - (x_{\mathrm{L}}(n) + x_{\mathrm{R}}(n))/2 = 0 \tag{5}$$

(assuming noise-free recordings $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$). The reason is that $(x_{\mathrm{L}}(n) + x_{\mathrm{R}}(n))/2$ might be even closer to $x_{\mathrm{L}}(n)$ (or symmetrically to $x_{\mathrm{R}}(n)$), therefore, $g(n)$ might be closer to $\delta(n)$.

In [3], the equalizing filters $g_{\mathrm{L}}(n)$ and $g_{\mathrm{R}}(n)$ are prelearned by using the normalized least mean square algorithm on noise-free data. In case of CHiME, we found that $g(n)$ can be computed non-adaptively using any noise-free utterance available in the development CHiME dataset. The filter of a given length $g(n)$ is found as the solution of a least squares problem

$$\min_{g} \sum_{n=1}^{N} \Big(\{g * x_{\mathrm{L}}\}(n) - (x_{\mathrm{L}}(n) + x_{\mathrm{R}}(n))/2\Big)^2, \tag{6}$$

where N denotes the number of samples.

B. Noise Subtraction

Once $g(n)$ is given,

$$v(n) = \{g * x_{\mathrm{L}}\}(n) - (x_{\mathrm{L}}(n) + x_{\mathrm{R}}(n))/2 \tag{7}$$

provides the estimate of noise in a noisy recording[2]. Since the speaker is standing directly in front of the dummy head, it is efficient to subtract the estimated noise from

$$u(n) = (x_{\mathrm{L}}(n) + x_{\mathrm{R}}(n))/2 \tag{8}$$

by an adaptive filter applied to the signal in the time-frequency domain [5].

In [3], a Wiener-like filter from [6] using a priori knowledge of SNR is applied. Here, we apply a similar filter proposed in [7] that is driven by a single parameter that allows a trade-off between the achieved SNR and SDR.

Let $U(k,\ell)$ and $V(k,\ell)$ be the short-time Fourier transform of $u(n)$ and $v(n)$, respectively, where k is the frequency index and ℓ is the time-frame index. The filter is defined in the time-frequency domain by

$$W(k,\ell) = \frac{|U(k,\ell)|^2}{|U(k,\ell)|^2 + \tau|V(k,\ell)|^2} \tag{9}$$

where τ is a free non-negative parameter. The time-frequency representation of the final output signal is defined through

$$\widehat{S}(k,\ell) = W(k,\ell)U(k,\ell). \tag{10}$$

Note that for $\tau = 0$, $W(k,\ell) = 1$, so the signal u remains unchanged. On the other hand, if both τ and $|V(k,\ell)|$ have large values compared to $|U(k,\ell)|$, $W(k,\ell)$ is close to zero, which suppresses the (k,ℓ)th time-frequency bin in the output signal. It follows that τ controls the achieved SNR and SDR, and (9) could be seen as a “fuzzy” variant of the binary mask proposed in [5].

III. Results

This section summarizes the proposed separation approach and describes the results when it was applied to the CHiME development and final dataset of isolated utterances sampled at the rate of 16 kHz. Each dataset contains 600 utterances at 6 different noise levels.

Initially, the filter $g(n)$ of length 2000 taps was computed by finding the minimum in (6) using a randomly chosen noise-free utterance (from file `s1_pgak4p.wav`), whose length is less than two seconds.

Next, we found it effective to remove frequencies below 70 Hz prior to further processing of the recordings. These frequencies usually correspond to noise only and could have large energy. To this end, we designed the high-pass Butherworth filter $f(n)$ of the fourth order with the cut-off frequency at 70 Hz.

Each noisy recording of the development and final dataset was processed in the following steps.

[1] Thanks to the symmetry, two different estimates of noise given by (2) can be obtained, which is partly beneficial for the retrieval of stereo signal preserving the binaural hearing.

[2] In [3], the noise estimates are yet equalized by minimizing the mean square distance between them and original recordings x_{L} and x_{R} in order to restore their spectra. In case of CHiME, we found that this equalization deteriorates the final performance, so we did not apply it.

TABLE I
AVERAGE RESULTS IN TERMS OF SNR IMPROVEMENT AND SDR ACHIEVED ON THE DEVELOPMENT DATASET

original SNR level	-6dB	-3dB	0dB	3dB	6dB	9dB
			$\tau = 1$			
SNR impr. [dB]	1.66	1.47	1.36	1.16	0.99	0.74
SDR [dB]	17.53	17.92	18.96	20.03	20.99	22.06
			$\tau = 10$			
SNR impr. [dB]	4.34	3.81	3.38	2.82	2.26	1.65
SDR [dB]	7.89	8.09	8.87	9.57	10.21	10.82
			$\tau = 50$			
SNR impr. [dB]	6.45	5.71	4.93	4.06	3.18	2.39
SDR [dB]	1.90	1.98	2.65	3.22	3.74	4.20
			$\tau = 100$			
SNR impr. [dB]	7.33	6.50	5.58	4.61	3.60	2.78
SDR [dB]	-0.59	-0.57	0.08	0.61	1.10	1.53

1) Apply the high-pass filter $f(n)$ to $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ to remove the low-frequency noise below 70 Hz.
2) Compute $u(n)$ according to (8).
3) Using $g(n)$, compute $v(n)$ according to (7).
4) Apply the adaptive filter (9) to $u(n)$ using the noise estimate $v(n)$; the length of frame in the short-time Fourier transform is 1024 samples, and the shift of frame is 32 samples.
5) Store the output signal into a file.

A. Resulting SNR and SDR on the Development Dataset

In case of the development dataset, the clear utterances are provided, which allows the evaluation of SNR and SDR. Let $\widehat{s}(n)$ denote an enhanced signal. It can be written as a sum of two terms

$$\widehat{s}(n) = \widetilde{s}(n) + \widetilde{y}(n), \tag{11}$$

where $\widetilde{s}(n)$ is the contribution of the target speech and $\widetilde{y}(n)$ is the residual noise. We define the SNR of the enhanced signal $\widehat{s}(n)$ as

$$\text{SNR} = \frac{\sum_{n=1}^{N} |\widetilde{s}(n)|^2}{\sum_{n=1}^{N} |\widetilde{y}(n)|^2}, \tag{12}$$

and the SDR of the enhanced signal as

$$\text{SDR} = \frac{\sum_{n=1}^{N} |\widetilde{s}(n)|^2}{\sum_{n=1}^{N} |\{(h_{\mathrm{L}} + h_{\mathrm{R}})/2 * s\}(n) - \widetilde{s}(n)|^2}, \tag{13}$$

where N denotes the length of the recording.

The resulting SNR and SDR averaged over all 600 utterances are summarized in Table I for each of the 6 noise levels. The results are shown also for different choices of the parameter τ in (9). As expected, the results demonstrate that τ controls the resulting SDR and SNR so that for higher values of τ SNR increases while SDR decreases, and vice versa.

B. Recognition Results

The processed datasets were sent to the baseline recognizer provided by the CHiME organizers[3].

TABLE II
KEYWORD RECOGNITION ACCURACY ACHIEVED ON THE DEVELOPMENT DATASET

SNR level	-6dB	-3dB	0dB	3dB	6dB	9dB
untreated [%]	31.08	36.75	49.08	64.00	73.83	83.08
			$\tau = 1$			
enhanced [%]	44.92	52.50	65.50	76.42	83.50	89.08
			$\tau = 10$			
enhanced [%]	**54.58**	**63.50**	**73.00**	**82.08**	**87.67**	**90.00**
			$\tau = 50$			
enhanced [%]	51.83	60.42	69.17	78.92	82.58	86.08
			$\tau = 100$			
enhanced [%]	45.50	54.50	63.42	72.08	78.00	81.67

TABLE III
KEYWORD RECOGNITION ACCURACY ACHIEVED ON THE FINAL DATASET

SNR level	-6dB	-3dB	0dB	3dB	6dB	9dB
untreated [%]	30.33	35.42	49.50	62.92	75.00	82.42
	$g(n)$ derived for the development dataset, $\tau = 10$					
enhanced [%]	44.75	54.00	65.33	74.83	83.92	87.50
	$g(n)$ derived for the test dataset, $\tau = 10$					
enhanced [%]	**52.08**	**62.00**	**74.75**	**81.83**	**88.00**	**91.25**

The results are evaluated in terms of recognition accuracy of keywords (i.e. percents of the letter and digit tokens recognized correctly) and are shown in Tables II and III, respectively, for the development and testing datasets. To compare, both tables show the recognition score achieved on untreated noise data.

1) Development dataset: The results achieved on the development dataset are shown for various choices of τ. It is seen that τ can be tuned to make a trade-off between SNR and SDR and that the optimum value could be different for each recording. The optimum may also depend on the recognizer. For instance, the results in Table I and II are indicative of a different sensitivity of the used recognizer to SDR than to SNR.

To conclude, the best results for the development dataset are achieved when $\tau = 10$, where the recognition score is improved by up to 27%.

2) Test dataset: The test dataset contains different utterances than the development dataset but was constructed in the same way. The only difference is that the source-microphone impulse responses $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ in (1) were measured under different conditions (e.g. doors open/closed, curtains drawn/undrawn). This means that the filter $g(n)$ derived using a noise-free signal from the development dataset may not be so efficient on testing data. The results in Table III confirm this claim, because the maximum improvement of the recognition score is about 19% only.

We therefore manually selected a noise-free segment of a test data (the first second of `s10_sgwg1s.wav` from the 9dB SNR dataset) and recomputed $g(n)$ by minimizing (6). Table III shows that the results of the repeated experiment using the novel $g(n)$ are comparable with those achieved on the development dataset.

IV. CONCLUSION

We have proposed a separation approach that is suitable for the CHiME data. It improves the recognition score by 7–27% depending on the original SNR. The approach is not blind as it relies on the existence of a speech-only recording, fixed

[3] `http://www.dcs.shef.ac.uk/spandh/chime/PCC/data/pcchome.tar.gz`

positions of the speaker and hearer, and stationary reverberant conditions. The advantage of the approach consists in its simplicity and speed.

The experiment with the test dataset has demonstrated the sensitivity of the method to small changes of reverberation. When the reverberation is slightly changed, the signal is still enhanced, but the recognition score is not so good. Although a short segment of clean speech (one second) suffices to update the filter $g(n)$, the changes may be relatively faster in more realistic conditions (e.g., small movements of speaker's and hearer's head). It is therefore a challenge for the further research to put together non-blind approaches such as the one proposed here and blind methods that might be used for fine-tuning of the separation system against changes about which no a priori information is provided.

ACKNOWLEDGEMENT

This work was supported by Grant Agency of the Czech Republic through the project P103/11/1947 and by Technology Agency of the Czech Republic through the project TA01011142.

REFERENCES

[1] J. Benesty, M. M. Sondhi, Y. Huang (Eds), *Springer Handbook of Speech Processing*, Springer, 2007.

[2] J. Benesty, S. Makino, and J. Chen (Eds.), *Speech Enhancement*, 1st edition, Springer-Verlag, Heidelberg, 2005.

[3] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model", *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2009)*, pp. 133 - 136, New Paltz, New York, Oct. 2009.

[4] B. Albouy and Y. Deville, "Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures", *Proc. of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 361-366, Nara, Japan, April 1-4, 2003.

[5] Ö. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking", *IEEE Trans. on Signal Processing*, vol. 52, no. 7, July 2004.

[6] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation", *Proc. of ICASSP 1996*, vol. 2, pp. 629–632, 1996.

[7] Z. Koldovský, J. Nouza, and J. Kolorenč, "Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA", *Proc. of Interspeech 2006*, Pittsburgh PA, USA, 17.-21. September, pp. 2578–2581, 2006.