Close speaker cancellation for suppression of non-stationary background noise for hands-free speech interface

Jani Even [1], *Carlos Ishi* [1], *Hiroshi Saruwatari* [2], *Norihiro Hagita* [1]

[1] ATR-IRC, Japan
[2] Graduate school of information science, Nara institute of science and technology, Japan
even@atr.jp

10.21437/Interspeech.2010-326

Abstract

This paper presents a noise cancellation method based on the ability to efficiently cancel a close target speaker contribution from the signals observed at a microphone array. The proposed method exploits this specificity in the case of the hands-free speech interface when the target user is close to the microphone array and the noise is a diffuse background noise. This method is in particular able to deal with non-stationary noise. The method can be divided in three steps. First, the steering vector pointing at the target user is estimated from the covariance of the observed signals. Then the noise estimate is obtained by canceling the user's contribution. During this step the speech pauses are also estimated. Finally a post-filter is used to suppress this estimated noise from the observed signals. The post-filter strength is controlled by using the estimated noise during the speech pauses as reference. A 20k-words dictation task in presence of non-stationary diffuse background noise at different SNR levels illustrates the effectiveness of the proposed method.
Index Terms: speech enhancement, hands-free speech interface, noise cancellation

1. Introduction

Speech is the most natural communication medium for human consequently it is of great interest to develop speech interfaces for controlling machines [1]. Among the different choices for designing a speech interface, the hands-free speech interface in which the voice of the user is picked up at a distance is one of the most natural approach because it suppresses the need to use a hand-held microphone or a headset. But this convenience comes at the expense of a higher level of ambient noise in the captured speech signal. Consequently, for a hands-free speech interface operating in realistic situation, it is necessary to deal with the noise. This can be done in the recognizer by training on noisy data but this solution is only suitable if we have a good knowledge of the noise present during operation. Another approach is to apply a denoising algorithm before feeding the received speech utterance to the recognizer.

For single microphone hands-free speech interfaces, several denoising methods use the speech pauses to get an estimate of the noise spectral density and then suppress the estimated noise from the observed signal using a nonlinear post-filter [2, 3]. These approaches are very effective for suppressing relatively stationary noise for which the noise spectral density is well estimated during speech pauses.

In the case of speech interfaces equipped with a microphone array, beamforming [4] can be applied to enhance the quality of the captured speech. However in presence of diffuse background noise, the performance of beamforming is limited by the number of microphones. To overcome this limitation, we can use an architecture similar to the generalized side lobe canceler (GSC) [5]. In the GSC approach, first a spatial filter is applied to the signal observed by the microphone array. This spatial filter is composed of two paths: a fixed beamformer gives a slightly enhanced speech signal and a blocking matrix gives an estimate of the noise. Then these two estimates are combined with a nonlinear post-filter in order to obtained the final enhanced speech signal. The GSC approach can be modified by changing the post-filter [6] or modifying the spatial filter [7, 8]. Contrary to the original GSC, the approaches proposed in [7, 8] use an adaptive beamformer that does not constrain the user to be at a fixed position.

In this paper, we consider a speech interface where a single user is talking close to a microphone array and can be considered as the dominant signal. Moreover, we assume that the noise is a diffuse background noise created by sources far from the array. This diffuse background noise is not assumed to be stationary. This scenario describes the use of a personal device in a noisy environment, thus it is of great interest to find an effective solution. In this paper, we will show that by exploiting the specificity of this problem, it is possible to perform effective denoising using an approach combining an adaptive beamformer, a noise estimator (adaptive blocking matrix) and a nonlinear post-filter like the ones in [7, 8] but with a reduced computation cost. In the following, we first present the model of the studied speech interface and describe a noise cancellation method similar to the ones in [8]. Then we introduce the simplified method and detail how to get the adaptive beamformer and the blocking matrix from the dominant eigenvector of the observed signals covariance matrix and how to suppress the estimated noise from the observed signals. Finally, the performance of the proposed method is assessed by a 20k-word dictation task in presence of a real-world non-stationary diffuse background noise.

2. Hands-free speech interface model

In this paper, we consider the situation where the target user is close to the microphone array whereas the diffuse background noise is created by sources far from the microphone array. Consequently, only the target user is assumed to be a point source and have a clear direction of arrival (DOA). We define the model of this hands-free speech interface in the frequency domain. The frequency domain signals are obtained using a short time Fourier transform (STFT) of size F. In the remainder f denotes the frequency bin and k denotes the frame index. The mixing model in the fth frequency bin is

This work was supported in part by the Ministry of Internal Affairs and Communications.

$$\mathbf{X}(f,k) = \left[\begin{array}{c|c} \mathbf{H}_\theta(f) & \mathcal{I}_n \end{array} \right] \left[\begin{array}{c} S(f,k) \\ \mathbf{N}(f,k) \end{array} \right], \quad (1)$$

where $S(f,k)$ is the speech component, $\mathbf{N}(f,k)$ is a vector containing the n components of the diffuse background noise, $\mathcal{I}_n$ is the identity matrix of size n and

$$\mathbf{H}_\theta(f) = \{\exp(j2\pi(f/F)f_s \frac{id}{c} \sin\theta(f))\}_{i\in[0,n-1]}$$

is a $n \times 1$ steering vector depending of the target speech DOA $\theta(f)$ (also of the sampling frequency f_s, the microphone spacing d, and the sound velocity c). Note that the vector $\mathbf{H}_\theta(f)$ is function of the frequency. The reason is that the *apparent* DOA at a given frequency, that accounts for the effect of the reflection and the reverberation, differs from the *physical* DOA of the speech, which is an angle defined by the user position relatively to the microphone array. It is a realistic assumption that, in a given frequency bin, the target speech component is statistically independent of the diffuse background components. But the statistical independence of the diffuse background noise components is not assumed.

3. FD-BSS with nonlinear post-filter

In this section, we present the general idea behind the denoising method in [8, 9]. This method use frequency domain blind signal separation (FD-BSS) to blindly estimate the target speech and the noise.

In the fth frequency bin, the estimate $Y(f,t)$ is obtained by applying an unmixing matrices $\mathbf{W}(f)$ to the observed signals

$$\mathbf{Y}(f,k) = \mathbf{W}(f)\mathbf{X}(f,k)$$

$\mathbf{W}(f)$ is updated to minimize the mutual information of $\mathbf{Y}(f,k)$ and converges to

$$\mathbf{W}(f) = \left[\begin{array}{c} \frac{1}{n}\mathbf{H}_\theta^H(f) \\ \mathbf{W}_\perp(f) \end{array} \right] \quad (2)$$

where $\mathbf{W}_\perp(f)$ is a $n1 \times n$ matrix of rank $n-1$ such that $\mathbf{W}_\perp(f)\mathbf{H}_\theta(f) = O_{n-1\times 1}$. Namely the first component is a delay and sum (DS) beamformer in the target speech direction whereas the other components have spatial nulls in the speech direction [8]. Then the noise estimate $\widehat{\mathbf{N}(f,t)}$ is obtained by projecting back the noise components [10] and suppressed from the observed signal applying a nonlinear post-filter channel-wise. Finally the speech estimate $\widehat{S}(f,t)$ is obtained by applying a delay and sum (DS) beamformer in the direction θ of the target speech to merge the different channels. Fig. 1 illustrates this in the case of a Wiener post-filter. These approaches are able to deal with non stationary noises and do not constrain the user's DOA.

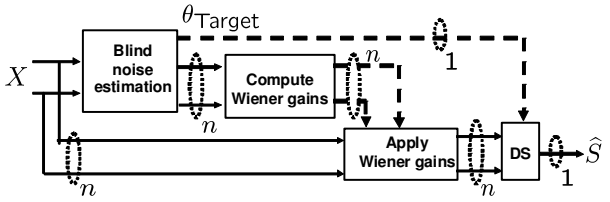

Figure 1: Blind noise estimation with channel-wise Wiener post-filter.

4. Proposed approach

4.1. Steering vector estimation

In eigenspace beamforming it is common to estimate the steering vector pointing to each user by using the periods where only that user is active [11]. During these periods, the eigenvector corresponding to the largest eigenvalue of the estimated covariance matrix is collinear to the steering vector pointing to the active user.

For the hands-free speech interface with a single dominant user, it is reasonable to use the same method to estimate the steering vector pointing to the user. With the model defined in Eq.(1), we can write the covariance of the observed signal as

$$\Gamma_{\mathbf{X}}(f) = \sigma_S^2 \mathbf{H}_\theta(f)\mathbf{H}_\theta^H(f) + \Gamma_{\mathbf{N}}(f),$$

where $\Gamma_{\mathbf{N}}(f)$ is the covariance of the noise and $\sigma_S^2 = \mathcal{E}\{|S(f,k)|^2\}$. Because we assume that the target user speech is the dominant signal, the eigenvector $\mathbf{V}(f)$ corresponding to the largest eigenvalue of the estimated covariance $\widehat{\Gamma_{\mathbf{X}}}(f)$ is taken as steering vector estimate $\widehat{\mathbf{H}_\theta(f)} = \mathbf{V}(f)$.

The quality of this estimation depends of how dominant is the target speech and also affect the noise estimation (see next section).

4.2. Noise estimation

In the propose method, the noise estimate is obtained by subtracting the speech contribution from the observation

$$\widehat{\mathbf{N}(f,k)} = \left(\mathcal{I}_n - \frac{1}{n^2}\Gamma_{\mathbf{X}}(f)\lambda^H \mathbf{V}(f)\lambda \mathbf{V}^H(f) \right) \mathbf{X}(f,k)$$

where λ is a scalar such that $\mathcal{E}\{|z(f,k)|^2\} = 1$ where $Z(f,k) = \frac{1}{n}\lambda \mathbf{V}^H(f)\mathbf{X}(f,k)$.

Assuming perfect estimation of $\mathbf{H}(f)$ we have

$$Z(f,k) = \lambda \left[\begin{array}{cc} 1 & \frac{1}{n}\mathbf{H}_\theta(f) \end{array} \right] \left[\begin{array}{c} S(f,k) \\ \mathbf{N}(f,k) \end{array} \right],$$

then the constraint on $Z(f,k)$ gives $|\lambda|^2 = \frac{1}{\sigma_S^2 + \frac{\sigma_\theta^2}{n^2}}$, where $\sigma_\theta^2 = \mathbf{H}_\theta^H(f)\Gamma_{\mathbf{N}}(f)\mathbf{H}_\theta(f)$.

We can rewrite the noise estimate (dropping frequency index)

$$\widehat{\mathbf{N}(k)} = \left(\mathcal{I}_n - \frac{1}{1 + \frac{\sigma_\theta^2}{n^2\sigma_S^2}} \left(\mathcal{I}_n + \frac{\Gamma_{\mathbf{N}}}{n^2\sigma_S^2} \right) \right) \mathbf{H}_\theta S(k)$$
$$+ \left(\mathcal{I}_n - \frac{1}{1 + \frac{\sigma_\theta^2}{n^2\sigma_S^2}} \left(\frac{1}{n}\mathcal{I}_n + \frac{\Gamma_{\mathbf{N}}}{n^2\sigma_S^2} \right) \mathbf{H}_\theta \mathbf{H}_\theta^H \right) \mathbf{N}(k) \quad (3)$$

The quantity σ_θ^2 can be seen as the noise power in the direction θ, thus the quality of the noise estimation is function of the ratio of power in the direction θ $\frac{\sigma_\theta^2}{n^2\sigma_S^2}$ and of the ratio of power in all directions $\frac{\Gamma_{\mathbf{N}}}{n^2\sigma_S^2}$.

In our case, the two ratios are small as we consider a dominant speech signal. Thus the noise estimate is

$$\widehat{\mathbf{N}(k)} \approx \left(\mathcal{I}_n - \frac{1}{n}\mathbf{H}_\theta \mathbf{H}_\theta^H \right) \mathbf{N}(k)$$

meaning that the noise is underestimated in the direction θ where a spatial null suppresses the close target speech. This noise estimate is equivalent to the one obtained using the projection back in [8] as showed in [12].

When the target user is not clearly dominant, in addition to the estimation error on the steering vector, estimation errors appear in the noise estimate because the two ratios are no longer small. The target speech leaks in the noise estimate ,first term of Eq.(3), and the noise is distorted, second term of Eq.(3).

4.3. Detection of speech pauses

Because the close target speech is efficiently canceled by a spatial null (like in FD-BSS based methods [8]), the ratio

$$\frac{<|\widehat{\mathbf{N}(f,k)}|^2>_{F_0}}{<|\mathbf{X}(f,k)|^2>_{F_0}} \quad (4)$$

where $<\cdot>_{F_0}$ denotes an averaging on the frequency band F_0 that contains the speech is a good indicator of speech presence. This ratio is closer to one when there is speech pauses and decreases when the speech is active. This is illustrated in fig.2 where the ratio (scaled) is superimposed on the spectrogram of the observed signal (the data are the same as in Sect.5).

We use the K frames where the ratio Eq.(4) is the closest to one to estimate $<|\widehat{\mathbf{N}(f,k)}|^2>_{F_0,K}$ where $<\cdot>_{F_0,K}$ denotes an averaging on the frequency band F_0 and on the K selected frames. These K frames appear as circled values in fig.2. This average quantity is regarded as the average power of the diffuse background noise.

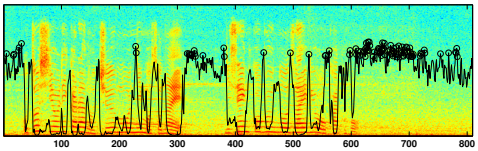

Figure 2: Spectrogram and ratio Eq.(4) with the K selected frames marked (circles).

4.4. Post-filter

We use a post-filter similar to the one used in [9, 12] The Wiener gain for the ith signal is

$$G^{(i)}(f,k) = \frac{|X^{(i)}(f,k)|^2}{|X^{(i)}(f,k)|^2 + \alpha|\widehat{N}^{(i)}(f,k)|^2}$$

where the subscript (i) denotes the ith component and α is a parameter controlling the noise reduction. The ith component of the filtered target speech is

$$\widehat{S}^{(i)}(f,k) = \sqrt{G^{(i)}(f,k)|X^{(i)}(f,k)|^2}\frac{X^{(i)}(f,k)}{|X^{(i)}(f,k)|}.$$

Finally the speech estimate $\widehat{S}(f,k)$ is obtained by applying a delay and sum (DS) beamformer in the estimated direction $\widehat{\theta}$ of the target speech (the estimated DOA θ is obtained from the steering vectors $\mathbf{V}(f)$ and averaged on the frequency band where the speech is present).

In this paper, we set the parameter α such that

$$\frac{<|\mathbf{X}(f,k)|^2>_K}{<|\mathbf{X}(f,k)|^2>_K + \alpha <|\widehat{\mathbf{N}(f,k)}|^2>_K} = G_{\min}$$

where $<\cdot>_K$ denotes the averaging on the K frames with smallest ratio Eq.(4) and $G_{\min}$ is the desired attenuation of the average power of the diffuse background noise $<|\widehat{\mathbf{N}(f,k)}|^2>_K$.

5. Experimental results

To evaluate the effectiveness of the proposed denoising method for hands-free speech interface, we use a 20K-words dictation task in presence of a real-world noise. We used the JNAS database [13] and the recognizer JULIUS (version 4.1.1 fast)

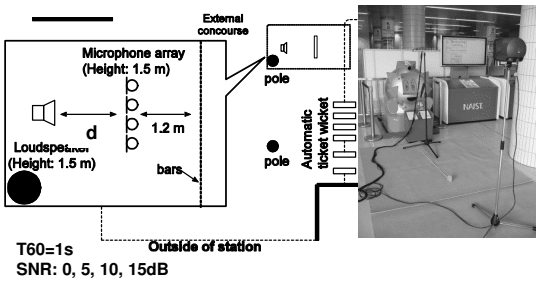

Figure 3: Experimental setting.

Table 2: *System specifications.*

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior (JNAS)
Test data	Adult and Senior female (JNAS)

using Phonetically Tied Mixture (PTM) model [14]. The open test set is composed of 100 utterances (female speakers). The conditions used in recognition are given in Table 2. The experimental data were generated using recording from a train station hall with a eight microphone linear array (spacing 2.15cm), see experimental setting in Fig. 3. The test sentences are convoluted with the measured impulse response and mixed at different SNR levels with the recorded non-stationary diffuse background noise.

Three acoustic models are used a clean model, a clean model with super-imposed office noise at 30dB SNR and a clean model with super-imposed office noise at 25dB SNR. For the two models with super-imposed office noise, the same office noise is added with an SNR of 30dB to the utterances before performing recognition (masking noise).

Three different steering vector estimation methods are compared: the use of the dominant eigenvector, the use of FD-BSS as in [8] and the use of frequency domain blind signal extraction (FD-BSE) as in [12]. These three methods are respectively designated by 'Proposed', 'BSS+WF' and 'BSE+WF' in Table 1. 'Unprocessed' refers to the unprocessed signals and 'DS' to the output of the DS beamformer (eigenvector) without the nonlinear post-filter. The post-filter is the same for all steering vector estimation methods. The STFT uses a 512 point Hanning window with 128 points shift and a 1024 points FFT. The number of frames used for estimating the average noise power is $K = 100$ and the attenuation $G_{\min}$ is taken in the set $\{-33, -30, -23, 20, -13, -10, -3\}$ dB. Table 1 shows the result for the best $G_{\min}$.

We can see that for the simple case of a close target user in diffuse background noise, using the covariance to estimate the steering vector gives results that are as good as the one obtained from more complex estimation methods. Moreover the computational cost is greatly reduced compared to FD-BSS and FD-BSE (in this simulation 0.3s for the proposed method, 126s for FD-BSE and 360s for FD-BSS but both the adaptive methods were applied with a larger than necessary number of iterations).

The influence of the attenuation $G_{\min}$ (in dB) of the av-

Table 1: Word accuracy for the different acoustic models and SNR levels.

AM	clean							clean + 30dB office noise							clean + 25 dB office noise						
SNR (dB)	0	5	10	15	20	25	30	0	5	10	15	20	25	30	0	5	10	15	20	25	30
unprocessed	0.70	6.16	25.42	57.11	75.46	82.10	82.86	3.36	17.65	57.46	77.21	84.50	86.03	86.75	6.15	27.66	62.57	78.17	84.96	87.31	86.43
DS	2.41	12.33	40.07	61.05	73.76	77.73	79.29	9.37	38.20	67.40	81.33	85.08	87.32	86.75	16.93	48.94	72.23	80.43	87.57	86.36	87.50
BSE + WF	**11.93**	32.97	**60.18**	**72.33**	**78.00**	81.93	**83.44**	**26.90**	55.33	**76.99**	83.17	87.62	88.27	**90.22**	**30.59**	58.16	**75.79**	84.06	87.28	88.20	88.65
BSS + WF	9.00	29.00	57.77	70.87	77.17	79.00	80.98	19.32	51.87	72.34	81.21	85.40	86.09	88.63	22.24	55.74	73.73	82.67	84.70	86.12	87.17
Proposed	9.37	**35.71**	57.99	70.32	77.43	**82.63**	83.37	24.03	**56.39**	76.62	**85.02**	**88.19**	**89.52**	89.78	30.07	**58.85**	75.52	84.36	**88.27**	**89.54**	**89.46**

erage diffuse background noise power $< |\widehat{\mathbf{N}(f,k)}|^2 >_K$ on the word accuracy is presented in Fig. 4 for 0dB (□), 5dB (▽), 10dB (○) and 15dB (×) SNR (for the clean model with 30dB office noise super-imposition). We can see that the choice of the attenuation is especially critical at lower SNR levels. An attenuation of -20dB is a good compromise for an unknown input SNR level (a possible development is to estimate the SNR using the speech pauses).

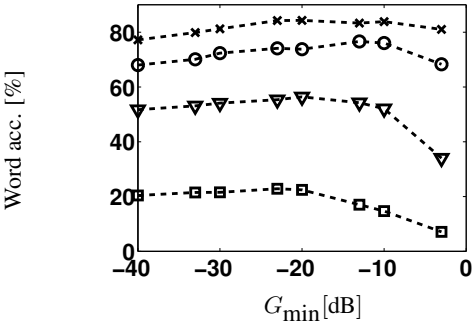

Figure 4: Word acc. (for the clean model with 30dB office noise super-imposition) vs attenuation $G_{\min}$ (in dB) for 0dB (□), 5dB (▽), 10dB (○) and 15dB (×) SNR levels.

Fig. 5 shows an example of spectrograms for 15dB SNR with $G_{\min} = -20$dB and $K = 100$.

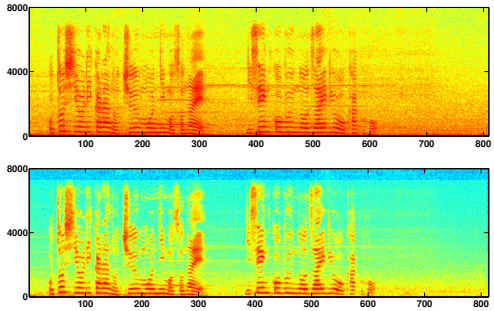

Figure 5: Spectrograms of observation (top) and processed signal (bottom).

6. Conclusions

In this paper, we showed that in the case of a single dominant target speaker in presence of diffuse background noise, using the dominant eigenvector of the covariance matrix is an affordable alternative to more complex blind processing method. We also proposed a practical way to set the strength of the post-filter using the estimated speech pauses. When the target user is not the only dominant signal this approach fails thus the future work is to detect such cases and shift to FD-BSS or FD-BSE estimation when necessary.

7. References

[1] R. Rosenfeld, D. Olsen, and A. Rudnicky, "Universal speech interfaces," *Interactions*, vol. 8, no. 6, pp. 34–44, 2001.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[4] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Prapagation*, vol. AP-30, pp. 27–34, 1982.

[5] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[6] S. Doclo et al., "Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction," *EUSIPCO'04*, pp. 2007–2010, 2004.

[7] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, pp. 29–37, 2008.

[8] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009.

[9] Y. Takahashi, Y. Uemura, H. Saruwatari, K. Shikano, and K. Kondo, "Structure selection algorithm for less musical-noise generation in integration systems of beamforming and spectral subtraction," *2009 IEEE Workshop on Statistical Signal Processing SSP2009, Cardiff, Wales, UK*, pp. 701–704, 2009.

[10] N. Murata, S. Ikeda, and A. Zieh, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[11] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[12] J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Speech enhancement in presence of diffuse background noise: Why using blind signal extraction?" *International Conference on Acoustics, Speech, and Signal Processing ICASSP 2010, Dallas, USA*, pp. 4770–4773, 2010.

[13] K. Ito et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoust. Soc. of Japan*, vol. 20, pp. 196–206, 1999.

[14] "Julius, an open-source large vocabulary csr engine - http://julius.sourceforge.jp."