BLIND SIGNAL EXTRACTION BASED JOINT SUPPRESSION OF DIFFUSE BACKGROUND NOISE AND LATE REVERBERATION

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano, and ‡*Tomoya Takatani*

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
‡ TOYOTA MOTOR CORPORATION, Aichi, Japan
email: even@is.naist.jp

ABSTRACT

The performance of automatic speech recognition for signals acquired through a hands-free speech interface is limited by the adverse effect of the noise and the reverberation. Frequency domain blind signal processing techniques, like blind signal separation, have been used with success for suppressing the noise in real situation but they usually do not take into account the reverberation. In this paper, we present a method based on frequency domain blind signal extraction that is aimed at suppressing both the adverse effect of the noise and the reverberation.

1. INTRODUCTION

The hands-free speech interface not only allows the user to interact with the machine in a natural way by using speech but it also frees the user from carrying a microphone or a headset as the speech is picked up at a distance by means of a microphone array. But this ease of use comes with a non negligible cost: the performance of automatic speech recognition system is deteriorated by the effect of the noise and the room reverberation.

Several microphone array techniques can be used to improve the captured speech by reducing these adverse effects [1, 2]. Among these techniques, frequency domain blind signal separation (FD-BSS), see review paper [3], has been used with success for estimating the diffuse background noise present in the hands-free speech interface [4]. In particular, as FD-BSS gives a better estimate of the diffuse background noise than of the target speech, it has to be combined with some nonlinear post-filtering techniques in order to improve the quality of the captured speech. However the approach proposed in [4] does not suppress the adverse effect of the reverberation.

As showed by the authors of [5, 6, 7], the late reverberation is the most harmful to the automatic speech recognition system. Consequently, they proposed approaches that suppress the later part of the reverberation by means of nonlinear filters (for example spectral subtraction of an estimated late reverberant speech). But these approaches were proposed in the noise free case.

In this paper, we present a method that combines frequency domain blind signal extraction (FD-BSE) [8] and nonlinear filter to suppress the noise as in [4] (we do not use single channel spectral subtraction but channel-wise Wiener filters as nonlinear post-filters as in [8]). But the proposed architecture also suppresses the late reverberation effect in a channel-wise manner by using another set of Wiener filters. The late reverberation effect is estimated by using the output after noise suppression, some *a priori* knowledge of the room reverberation and the information given by FD-BSE on the user's position (see [9] for late reverberation suppression using statistical room impulse response model). The effectiveness of the proposed method is illustrated by a dictation task performed with a hands-free speech interface in presence of both diffuse background noise and reverberation.

Notations: throughout the paper, vectors and matrices are in bold face, for signals $X(f,k)$ is the frequency domain representation of $x(t)$ and for filters $H(f)$ is the frequency domain representation of $h(\tau)$.

2. HANDS-FREE SPEECH INTERFACE

Let us first define the model of the hands-free speech interface, when a single user is talking in a noisy and reverberant room. We assume that the noise is a diffuse background noise created by noise sources far from the microphone array and that the user, closer to the array, is a point source.

The multi-dimensional signal (n components) received at the microphone array $\mathbf{x}(t)$ is the sum of the speech contribution $\mathbf{x}_S(t)$ and the diffuse background noise contribution $\mathbf{x}_N(t)$.

$$\mathbf{x}(t) \quad = \quad \mathbf{x}_S(t)+\mathbf{x}_N(t).$$

The multi-dimensional speech contribution reflects the effect of the room impulse response $\mathbf{h}(\tau)$ on the clean speech $s(t)$ and is composed of an early part $\mathbf{h}_E(\tau)$ and a late part $\mathbf{h}_L(\tau)$

$$\begin{aligned}\mathbf{x}_S(t) &= (\mathbf{h}_E(\tau)+\mathbf{h}_L(\tau))*s(t)\\ &= \mathbf{x}_E(t)+\mathbf{x}_L(t)\end{aligned}$$

where $\mathbf{x}_E(t)$ and $\mathbf{x}_L(t)$ are the early reverberant speech and the late reverberant speech.

Modern hidden Markov model (HMM) based speech recognizers are able to cope with the filtering effect of the room impulse response up to a certain delay τ_d (for example by applying cepstrum mean normalization). Thus $\mathbf{h}_E(\tau)$ and $\mathbf{h}_L(\tau)$ are defined as

$$\mathbf{h}_E(\tau) = \begin{cases}\mathbf{h}(\tau) & \text{for } \tau \le \tau_d\\ 0 & \text{for } \tau > \tau_d\end{cases}$$

$$\mathbf{h}_L(\tau) = \begin{cases}\mathbf{h}(\tau) & \text{for } \tau > \tau_d\\ 0 & \text{for } \tau \le \tau_d\end{cases}$$

meaning that the effect of $\mathbf{h}_E(\tau)$ is handled by the recognizer whereas the effect of $\mathbf{h}_L(\tau)$ must be handled by the signal processing front end (for the recognizer we use [10] the early/late reverberation threshold is $\tau_d = 75$ ms as shown in [7]).

To present the technique used to suppress the diffuse background noise, we use a simplified frequency domain model of the hands-free speech interface (not taking explicitly into account the late reverberation). The frequency domain signals are obtained using a short time Fourier transform of size F. In the remainder f denotes the frequency bin and k denotes the frame index. Considering that the user is a point source, the mixing model in the fth frequency bin is approximated by

$$\mathbf{X}(f,k) \approx \mathbf{H}_\theta(f)S_1(f,k)+\mathbf{N}(f,k), \qquad (1)$$

where $S_1(f,k)$ is the anechoic speech component, $\mathbf{N}(f,k)$ is a vector containing the n components of the diffuse background noise 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 $n\times 1$ vector depending of the speech direction of arrival (DOA) $\theta(f)$ (also of the sampling frequency f_s, microphone inter spacing d, and 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 the angle defined by the user's position relatively to the microphone

 ISSN 2076-1465

array. With this model, some amount of the late reverberation effect is included in the noise.

We can reformulate (1) as a noiseless instantaneous mixture

$$\mathbf{X}(f,k) = [\ \mathbf{H}_\theta(f)\ |\ \mathscr{I}_n\ ] \begin{bmatrix} S_1(f,k) \\ \mathbf{N}(f,k) \end{bmatrix},$$

where $\mathscr{I}_n$ is the identity matrix of size n.

For convenience we define

$$\mathbf{S}(f,k) = [S_1(f,k), S_2(f,k), \cdots, S_{n+1}(f,k)]^T$$
$$\text{with}\quad S_2(f,k), \cdots, S_{n+1}(f,k) = \mathbf{N}(f,k).$$

Then the noiseless instantaneous mixture is re-written as

$$\mathbf{X}(f,k) = \mathbf{A}(f)\mathbf{S}(f,k). \tag{2}$$

It is a realistic assumption that, in a given frequency bin, the target speech component is statistically independent of the diffuse background noise components. But the statistical independence of the diffuse background noise components is not assumed.

3. DIFFUSE BACKGROUND NOISE SUPPRESSION

In the fth frequency bin, the estimates $Y(f,k)$ of the separated components estimate are obtained by applying demixing matrices $W(f)$ to the observed signals

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

this matrix is updated in order to minimize the mutual information of the components of $Y(f,k)$ (see [3, 4] for FD-BSS method details).

In [4], Takahashi et al. showed that in this situation the square matrix $\mathbf{W}(f)$ estimated by BSS is such that the row corresponding to the speech component estimate is a delay and sum (DS) beamformer in the direction of the speech's apparent DOA at that frequency. The other rows corresponding to the estimates of the noise components are null beamformers at the speech's apparent DOA at that frequency.

After separation, assuming that the speech component is the first component of $\mathbf{Y}(f,k)$, the noise estimate $\widehat{\mathbf{X}_N}(f,k)$ is obtained by projecting back the noise components

$$\widehat{\mathbf{X}_N}(f,k) = \mathbf{W}(f)^{-1}\mathbf{D}\mathbf{W}(f)\mathbf{X}(f,k)$$

where $\mathbf{D}$ is a diagonal matrix with entries $[0,1,\cdots,1]$ along the diagonal. We have

$$\widehat{\mathbf{X}_N}(f,k) \approx [\, O_{n\times 1} \,|\, \mathscr{I}_n\,]\,\mathbf{S}(f,k).$$

Consequently the quality of the noise estimate is highly superior to that of the speech estimate as the null beamformers efficiently suppress the speech (a point source) from the estimated noise components whereas the DS beamformer does not suppress the noise from the estimated speech component. For this reason the authors in [4] propose to use FD-BSS for estimating the diffuse background noise and then apply a nonlinear post-filter to suppress the noise. This architecture, called blind spatial subtraction array (BSSA), is composed of two paths (see Fig. 1). The primary path (bottom) is a DS beamformer in the user's direction

$$\widehat{X_S}(f,k) = \mathbf{B}(\theta(f))\mathbf{X}(f,k)$$

and the second path (top) is the FD-BSS based noise estimation. The same DS beamformer is applied to the noise estimate

$$\widehat{X_N}(f,k) = \mathbf{B}(\theta(f))\widehat{\mathbf{X}_N}(f,k)$$

then spectral subtraction is used to suppress the diffuse background noise from the primary path

$$|\widehat{S}(f,k)| = \begin{cases} |\widehat{X_S}(f,k)|^2 - \alpha|\widehat{X_N}(f,k)|^2 \\ \text{if } |\widehat{X_S}(f,k)|^2 - \alpha|\widehat{X_N}(f,k)|^2 > 0 \\ \beta|\widehat{X_S}(f,k)|^2 \ \text{ else} \end{cases} \tag{3}$$

where the over subtraction parameter α and the flooring parameter β control the processing.

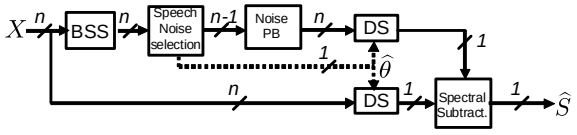

Figure 1: BSSA architecture.

4. SUPPRESSION OF THE LATE REVERBERATION EFFECT

In presence of heavy reverberation, the performance drop observed for the automatic speech recognition based on HMM is mainly caused by the later part of the reverberation $\mathbf{h}_L(\tau)$ that cannot be handled by the recognizer (see [7] for example). For this reason it is necessary to suppress this effect by pre-processing the speech with a dereverberation algorithm.

The method proposed in this paper uses the framework presented in [6] and in [7]. The speech signal has a strong correlation within each local time frame due to articulatory constraints but early and late reflections are uncorrelated. Consequently the authors of [6] proposed to estimate the early reverberant component by subtracting in the power spectrum domain an estimate of the late reverberant component to the observed signal.

A blind estimation of the late reverberant component with multi-step forward linear prediction was proposed in [11] where an effective suppression of the late reverberation was achieved by spectral subtraction.

The method proposed in [7] uses prior knowledge to avoid the costly blind estimation of the late reverberant component. It assumes that the late part of the impulse response $\mathbf{h}_L(\tau)$ that creates the late reverberation is not varying significantly within the room contrary to the early part of the impulse response $\mathbf{h}_E(\tau)$ that is strongly affected by the position of the speaker and the microphone array within the room. Consequently it is possible to obtain an acceptable estimate of the late reverberation for the room by measuring one impulse response before hand (thus this method is designed for systems that operate in a given room). The method in [7] uses the received speech to estimate the late reverberant speech and requires a modification of the spectral subtraction in order to compensate the estimation error on the late reverberant part.

5. PROPOSED JOINT SUPPRESSION OF DIFFUSE BACKGROUND NOISE AND LATE REVERBERATION EFFECT

5.1 Suppression of the diffuse background noise

In FD-BSE, at the fth frequency bin, the estimate $y(f,k)$ is obtained by applying an extracting vector $\mathbf{W}(f)$ to the observed signals

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

The vector $\mathbf{W}(f)$ that extract the speech component can be obtained by the method presented in [8] that minimize the cost function

$$J(\mathbf{W}(f)) = \frac{1}{2}\mathscr{E}\{|Y(f,k)|\}^2 \quad \text{under the constraint}$$
$$\mathscr{E}\left\{|Y(f,k)|^2\right\} = 1 \quad \text{with an iterative gradient descent.}$$

Then the diffuse background noise is estimated by subtracting the projection of the speech component from the observation

$$\widehat{\mathbf{X}_N}(f,k) = \left(\mathscr{I}_n - \mathbf{\Gamma}_\mathbf{X}(f)\lambda^H\mathbf{W}^H(f)\lambda\mathbf{W}(f)\right)\mathbf{X}(f,k)$$

where $\mathbf{\Gamma}_\mathbf{X}(f)$ is the covariance of $\mathbf{X}(f,k)$ and λ is a scalar such that $Z(f,k) = \lambda\mathbf{W}(f)\mathbf{X}(f,k)$ verifies $\mathscr{E}\{|Z(f,k)|^2\} = 1$. With a few assumption on the diffuse background noise we have [12]

$$\widehat{\mathbf{X}_N}(f,k) = \left[\ O_{n\times 1}\ \middle|\ \mathscr{I}_n - \tfrac{1}{n}\mathbf{H}_\theta\mathbf{H}_\theta^H\ \right]\mathbf{S}(f,k).$$

To suppress the diffuse background noise effect, a Wiener filter is applied on each component of the observed signal

$$\widehat{X_{Si}}(f,k) = \mathscr{W}\left(X_i(f,k), \alpha_N \widehat{X_{Ni}}(f,k)\right).$$

where α_N is a parameter controlling the noise reduction and the Wiener filter $\mathscr{W}(\cdot,\cdot,\cdot)$ is defined as

$$\begin{aligned} S(f,k) &= \mathscr{W}(X(f,k),N(f,k)) \\ &= \sqrt{G(f,k)|X(f,k)|^2}\frac{X(f,k)}{|X(f,k)|} \\ \text{with } G(f,k) &= \frac{|X(f,k)|^2}{|X(f,k)|^2+|N(f,k)|^2}. \end{aligned}$$

5.2 Suppression of the late reverberation effect

Assuming the noise suppression was efficient, the n components of $\widehat{\mathbf{X}_S}(f,k)$ contains the early reverberant speech components $\mathbf{X}_E(f,k)$ and the late reverberant speech components $\mathbf{X}_L(f,k)$.

As proposed in [11, 7] we use nonlinear filtering to suppress the late reverberation effect: another channel-wise Wiener filter is applied to the signal after noise suppression

$$\widehat{X_{Ei}}(f,k) = \mathscr{W}\left(\widehat{X_{Si}}(f,k), \alpha_R \widehat{X_{Li}}(f,k)\right)$$

where the $\widehat{X_{Li}}(f,k)$ are the components of the late reverberant speech estimate and α_R controls the filter strength.

Thus the focus is on the determination of the late reverberant speech estimate $\widehat{\mathbf{X}_L}(f,k)$. Using the relation

$$\mathbf{x}_L(t) = \mathbf{h}_L(\tau) * s(t)$$

this estimation is separated in two tasks: Obtaining an estimate of the filter $\mathbf{h}_L(\tau)$ and obtaining an estimate of the signal $s(t)$.

The estimate of the late reverberation filter exploits the fact that the late reverberation is rather room dependent and can be approximated by using a synthetically generated tail. Here we use a simple random tail with exponential decay

$$h_i(\tau) = a u(\tau) e^{-d(\tau-\tau_0)}$$

where a is a scalar, $u(\tau)$ a Gaussian random random variable with zero mean and unit variance, τ_0 is the limit between early and late reverberation ($\tau_0 = 75$ ms for our recognizer) and d is a decay factor. The decay factor is set to have an impulse response with a given T_{60} (the time after which the power of the tail decreased by 60dB). We use the approximation

$$d = \frac{\ln 10^6}{2(T_{60}-\tau_0)}$$

obtained by neglecting $u(\tau)$ while computing the integral in the power ratio

$$\frac{\int_{T_{60}}^{\infty} a^2 u(t)^2 e^{-2d(t-\tau_0)}dt}{\int_{\tau_0}^{\infty} a^2 u(t)^2 e^{-2d(t-\tau_0)}dt} \approx \frac{\int_{T_{60}}^{\infty} a^2 e^{-2d(t-\tau_0)}dt}{\int_{\tau_0}^{\infty} a^2 e^{-2d(t-\tau_0)}dt}.$$

Consequently the method requires an estimate of the reverberation time T_{60} and setting a value to a.

In the second task, the estimate of $s(t)$ is just instrumental in obtaining $\widehat{\mathbf{X}_L}(f,k)$. We propose an approach that uses the output of the noise suppression stage $\widehat{\mathbf{X}_S}(f,k)$ to get an intermediary signal for suppressing the late reverberation effect.

The FD-BSE method estimates a vector $\mathbf{W}(f)$ from which we can estimates a projection back filter for the speech signal

$$\begin{aligned} \widehat{\mathbf{X}_S}(f,k) &= \Gamma_{\mathbf{X}}(f)\lambda^H \mathbf{W}^H(f)\lambda \mathbf{W}(f)\mathbf{X}(f,k) \\ &= \mathbf{K}(f)\mathbf{X}(f,k). \end{aligned}$$

The energy for each microphone of the filter $\mathbf{K}(f)$ is used to get the scale parameters a of the synthetic tail.

Assuming that $\mathbf{W}(f)$ converged to $\frac{\lambda}{n}\mathbf{H}_\theta^H(f)$ we have

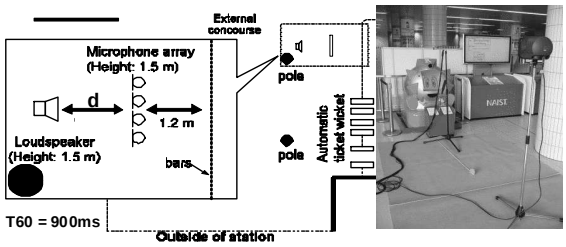

Figure 3: Experimental setting.

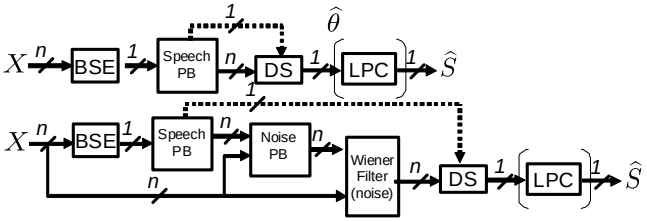

Figure 4: Comparison methods.

$$\frac{\{\mathbf{K}(f)\}_{p+1,q}}{\{\mathbf{K}(f)\}_{p,q}} = \exp(j2\pi(f/F)f_s\frac{d}{c}\sin\theta(f))$$

from which we can estimate $\theta(f)$ by taking

$$\widehat{\theta(f)} = \text{asin}\left(\frac{cF}{2\pi f f_s d}\text{angle}\left(\frac{\{\mathbf{K}(f)\}_{p+1,q}}{\{\mathbf{K}(f)\}_{p,q}}\right)\right)$$

this method is quite similar to the one in [13] but does not require a matrix inversion (estimating $\theta(f)$ directly from $\mathbf{W}(f)$ is possible but it is less robust in practice when the relation $\mathbf{W}(f) = \frac{\lambda}{n}\mathbf{H}_\theta^H(f)$ is approximate). Then a mean DOA $\widehat{\theta}$ is obtained from the $\widehat{\theta(f)}$ and used to apply a DS beamformer in the direction $\widehat{\theta}$ to $\widehat{\mathbf{X}_S}(f,k)$. The output of this DS beamformer is used as speech estimate to get the late reverberant speech. This signal is slightly closer to the true $S(f,k)$ than the components of $\widehat{\mathbf{X}_S}(f,k)$ because of the DS beamformer but it is a coarse estimate as the room impulse response effect is still present. However, the method is quite robust to this mismatch as can be seen in Sect.6.

5.3 Architecture

Fig. 2 shows the proposed architecture. The BSE algorithm is used to obtain both the DOA estimate $\widehat{\theta}$ and the diffuse background noise (a n component signal) then the first set of n Wiener filters suppress the noise. After noise suppression, the upper path estimates the late reverberant speech. First the DS beamformer in the direction $\widehat{\theta}$ gives the intermediary speech estimate, then the synthetic tail is applied (in the time domain) to this signal to get the n component estimate of the late reverberant speech that is suppressed by the second set of Wiener filters. Finally the speech estimate $\widehat{S}(f,k)$ is obtained by applying the DS beamformer in the direction $\widehat{\theta}$ of the estimated target speech to merge the output components of the second set of Wiener filters.

6. EXPERIMENTAL RESULTS

The simulation uses data recorded in a train station, see Fig. 3. A four ($n = 4$) microphone array (inter microphone spacing of 2.15 cm) was used to record the diffuse background noise, and estimate the impulse responses from two locations, 50 cm and 150 cm, in front of the array (DOA of 0^o). Since our goal is speech recognition, a 20K-word Japanese dictation task from the database JNAS is used as performance measure [14]. The test set (100 signals, female

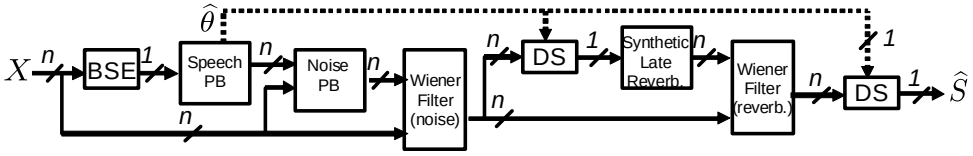

Figure 2: Proposed architecture.

Table 1: *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)

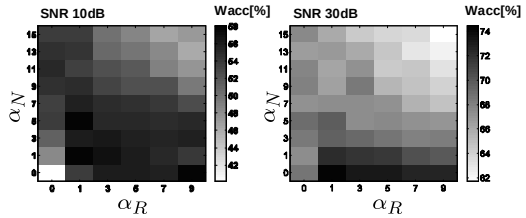

Figure 5: Effect of the set of parameters (α_N, α_R) on the word accuracy.

speakers only) is convoluted with the impulse responses and mixed with the recorded noise at different SNRs.

The quality of the speech estimate given by the proposed method (**prop**) is compared to the quality of the speech estimate obtained by: unprocessed signal (**obs**), FD-BSE alone (**bse**), FD-BSE with noise suppression by channel-wise Wiener filter (**bse-w**) and each of the three previous approaches cascaded with the multi-LPC dereverberation (**obs-lpc**, **bse-lpc** and **bse-w-lpc**). Fig. 4 shows some of these methods. The LPC block refers to the dereverberation method in [11] where the delay is $d = 400$ and the prediction filter is 3000 taps (these parameters correspond to the ones in [11]).

For the frequency domain processing, the short time Fourier transform uses a 512 point hamming window with 50% overlap. The separation is performed by 600 iterations of a BSE method with adaptation step of 0.3 divided by two every 200 iterations (the method is presented in [8]). For the proposed method, the parameter τ_0 is set to 75 ms and we use $T_{60} = 450$ ms (this is a mismatched value, simulations with $T_{60} = 900$ ms were also performed but the under estimation of T_{60} gave better results; maybe because the latest part of the reverberation is masked by the noise).

The recognizer is JULIUS [15] using Phonetically Tied Mixture (PTM) model. The conditions used in recognition are given in Table 1. The acoustic model is a clean model with super-imposed noise (office noise 30dB SNR). The recognition was performed with and without a masking noise; the same office noise as the acoustic model is mixed with the processed signal before recognition is performed (the mixing SNR is 30dB).

The word accuracies achieved with the different methods are given in Table. 2. The word accuracies displayed for **bse-w**, **bse-w-lpc** and **prop** are the higher one obtained from the parameter sets $\alpha_N = \{0,1,3,5,7,9,11,13,15\}$ and $\alpha_R = \{0,1,3,5,7,9\}$.

The proposed method gives the best performance except for 10 dB SNR when the user is close to the microphone array and there is no masking noise in which case the suppression of the diffuse noise alone (**bse-w**) is the most efficient method. Using a masking noise especially improves the performance at the higher SNR. But the proposed method is less affected than the **lpc** methods by the presence or not of the masking noise.

Fig. 5 shows the effect of the parameters α_N and α_R on the word accuracy for a distance of 150 cm at both 10 dB and 30 dB SNR. The case $(\alpha_N = 0, \alpha_R = 0)$ corresponds to **bse** and the cases $(\alpha_N \neq 0, \alpha_R = 0)$ correspond to **bse-w** whereas all the other cases correspond to **prop**. At 10 dB of SNR, good performance is achieved by taking a set of parameters with similar average sizes or by taking a set with a big and a small coefficient. Whereas at 30 dB of SNR it is preferable to have a small α_N as one can expect.

7. CONCLUSION

In this paper, we proposed a method that both suppresses the diffuse background noise and the late reverberant speech in order to improve automatic speech recognition performance while using a hands-free speech interface. The method that relies on blind signal processing for the noise and some *a priori* knowledge for the reverberation proved to be efficient in a realistic simulation. The next development is to include a blind estimation of T_{60} in the method and to propose better strategy for the choice of the set of parameters (α_N, α_R) (done in a room/SNR dependent manner now). Preliminary results showed that the method is robust to error estimation on T_{60} but also that the selection of appropriate values for the set of parameters (α_N, α_R) is closely related to the estimate of T_{60}.

REFERENCES

[1] L.J. Griffiths and C.W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Prapagation*, AP-30:27–34, 1982.

[2] S. Doclo, A. Spriet, and M. Moonen. Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction. *in Proc. EUSIPCO, Vienna, Austria*, pages 2007–2010, 2004.

[3] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra. *A Survey of Convolutive Blind Source Separation Methods*. Springer, 2007.

[4] 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*, 17(4):650–664, 2009.

[5] K. Lebart and J.M. Boucher. A new method based on spectral subtraction for speech dereverberation. *Acta Acoustica*, 87:359–366, 2001.

[6] K. Kinoshita, T. Nakatani, and M. Miyoshi. Efficient dereverberation framework for automatic speech recognition. *In Proceedings of ICSLP*, pages 92–95, 2005.

[7] R. Gomez, J. Even, H. Saruwatari, and K. Shikano. Distant-talking robust speech recognition using late reflection components of room impulse response. *International Conference on Acoustics, Speech, and Signal Processing ICASSP, Las Vegas, USA*, pages 4581–4584, 2008.

Table 2: Word accuracy for the different methods.

SNR (dB)	10 dB				30 dB			
distance (cm)	50 cm		150 cm		50 cm		150 cm	
masking	no	yes	no	yes	no	yes	no	yes
obs	57.58	56.56	39.99	39.80	86.17	85.71	67.98	67.22
obs-lpc	62.41	63.05	42.52	43.73	82.75	89.02	65.80	74.38
bse	57.77	57.40	40.11	43.09	86.24	86.92	68.55	69.16
bse-lpc	50.03	55.01	36.48	41.03	79.42	89.66	59.14	74.57
bse-w	**70.79**	71.35	55.13	55.58	85.65	88.21	68.99	72.54
bse-w-lpc	61.55	64.72	45.91	50.54	80.34	89.66	60.38	75.30
prop	70.11	**71.83**	**58.25**	**59.48**	**87.00**	**90.72**	**74.45**	**79.39**

[8] J. Even, H. Saruwatari, and K. Shikano. Blind signal extraction based speech enhancement in presence of diffuse background noise. *2009 IEEE Workshop on Statistical Signal Processing SSP2009, Cardiff, Wales, UK*, pages 513–516, 2009.

[9] E.A.P. Habets, S. Gannot, I. Cohen, and P.C.W. Sommen. Joint dereverberation and residual echo suppression of speech signals in noisy environments. *IEEE Transactions on Audio, Speech, and Languages Processing*, 16(8):1433–1451, 2008.

[10] A. Lee et al. Julius - an open source real-time large vocabulary recognition engine. *EUROSPEECH,*, pages 1691–1694, 2001.

[11] K. Kinoshita, T. Nakatani, and M. Miyoshi. Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation. *In Proceedings of ICASSP*, 2006.

[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*, pages 4770–4773, 2010.

[13] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. Speech and Audio Processing*, 12:530–538, 2004.

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

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