BLIND SOURCE SEPARATION IN MOBILE ENVIRONMENTS USING A PRIORI KNOWLEDGE

Erik Visser, Te-Won Lee

SoftMax Inc
4180 La Jolla Village Dr., Suite 455
La Jolla CA 92037
evisser@softmax.com, tlee@softmax.com

ABSTRACT

A speech enhancement scheme including blind source separation and background denoising based on minimum statistics is studied in mobile environments. To accommodate the dependence of the separated output signals on the spatial properties of the recorded source signals, these blind signal processing steps are complemented by an adaptive separated output channel selection stage using prior knowledge about the desired speaker speech content. The resulting scheme performance is illustrated by speech recognition experiments on real recordings corrupted by various noise sources and shown to outperform conventional beamforming and single channel denoising techniques as well as an equivalent scheme with fixed output channel selection.

1. INTRODUCTION

A number of single and multiple microphone based signal processing algorithms are nowadays available to address speech enhancement tasks in real environments. They often use a combination of probabilistic frameworks with statistical models of desired speech signals [1] and spatial information about signal mixtures by using an array of microphones with a known geometry to *suppress* interfering signals, also known as beamforming [2]. The drawback of these methods is the extensive use of a priori information about the acoustical environment and sources involved to achieve good performance limiting its robustness and flexibility in unknown environments.

Blind Source Separation (BSS) algorithms are an interesting alternative to these methods since, by design, they do not require a priori information about the signals involved to achieve good signal separation. Numerous contributions to the field of blind source deconvolution have focused on the design of algorithms with attractive theoretical properties such as good convergence, numerical stability and efficient tessellation of time-frequency bands. Although these basic "mechanics" of the ICA algorithm are largely "blind", i.e. operate without a priori source information, a certain number of implicit and explicit constraints are included in the ICA problem formulation to achieve meaningful results. As previous work in the field has shown [3, 4], this leads to dependence of the achievable separation performance on the spatial configuration of mixed signal sources. Indeed specific mixing scenarios may actually result in a singular problem case for which no satisfactory separating solution can be computed [3]. Another consequence of this spatial dependency in blind source deconvolution is the switching of the output order of separated sources in mobile microphone or source configurations. In order to benefit from the significant observed SNR improvements in the correct desired separated channel over the recorded signals in such situations, an additional stage needs to process the content of the separated audio channels if blind source separation is to be successfully integrated in speech recognition or communication systems. This in turn requires the availability of prior knowledge about the desired source signal to be identified. In this paper we will address some of these issues for a two microphone setup.

2. BLIND SOURCE SEPARATION AND BACKGROUND DENOISING

We consider an analytical framework with m different microphone mixture signals $x(t)$ composed of m point source signals $s(t)$ and additive background noise $n(t)$ (as previously discussed in [5])

$$\mathbf{x}(t) \;=\; \sum_{\tau=0}^{P} \mathbf{A}(\tau)\, \mathbf{s}(t-\tau) \;+\; \mathbf{n}(t)$$

where P is the convolution order, $\mathbf{A}(\tau)$ is a $m \times m$ mixing matrix.

While a multitude of frequency domain blind source separation algorithms for convolved mixtures have been developed, we focus again in this study on the example of a second order decorrelation approach presented in [6] to illustrate the generic channel selection problems typically encountered.

The Multiple Adaptive Decorrelation (MAD) algorithm [6] is designed for separating m recorded mixtures $\mathbf{x}(t) = \sum_{\tau=0}^{P} \mathbf{A}(\tau)\, \mathbf{s}(t-\tau)$ into m original sources $\mathbf{s}(t)$ by finding a sequence of m $\mathbf{x}$ m unmixing filter matrices $\mathbf{W}(\tau)$ such that $\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{Q} \mathbf{W}(\tau)\, \mathbf{x}(t-\tau)$, Q being the filter length. Q being the filter length. The unmixing filter computation is executed in the frequency domain by minimizing the cost function

$$\hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}_s = \arg \min_{\hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}_s} \sum_{t} \sum_{\omega=1}^{T} ||\mathbf{W} \hat{R}_x(\omega, t) \mathbf{W}^H - \Lambda_s(\omega, t)||^2 \quad (1)$$
$$s.t.\ \mathbf{W}(\tau) = 0, \forall \tau > Q,\ Q << T,$$
$$\mathbf{W}_{ii}(\omega) = 1$$

where the cross correlation of the measurements is denoted by $\hat{R}_x(\omega, t) = \mathsf{E}\left[\mathbf{X}(\omega, t)\, \mathbf{X}^H(\omega, t)\right]$ and that of the sources by $\hat{\Lambda}_s(\omega, t) = \mathsf{E}\left[\mathbf{S}(\omega, t)\, \mathbf{S}^H(\omega, t)\right]$. Scaling issues are solved by the second constraint fixing the diagonal elements of the filter matrices to unity. Background denoising is achieved through a minimum statistics type denoising algorithm. Minimum statistics based denoising algorithms seek to determine minimum noise power in each spectral subband over a finite time horizon. These noise power estimates are then used to compute the coefficients of a time-varying Wiener filter [7].

3. CHANNEL SELECTION

Although the objective function in (1) is designed to achieve separation of any type of mixed signals, the imposed constraints on filter length and amplitude introduce dependencies on the spatial properties of separable source signals. On one hand a constrained filter length prevents the modeling of a reverberation scenario with limited room transfer function complexity. Also the scaling constraint forces the output order of separated solutions to be determined by the relative amplitude strength of source signals at a given microphone location. In other words, for a two microphone setup, if source A is closer to microphone 1 than source B, separated source A will be output in the same order as the mixture recorded at microphone 1 is fed into the BSS algorithm. The fixing of direct feedforward filter coefficients to unity in problem (1) is necessary to solve independence of the separated solutions from scaling. In other formulations, whitening of the source separations is avoided in this manner as well [8]. The scaling constraint in (1) therefore leads to switching of a moving source from one channel to another as illustrated by Figure 1. In this case a speaker was moving from the right to the left side of a two microphone setup with the switching occurring in the middle of the recording.

As illustrated by Figure 2, a channel selection stage therefore has to included before feeding the processed signal to a subsequent application. The channel switching decision is made iteratively by employing known desired speech signal properties, for example a suitably designed command word tracking procedure. By comparing the relative log likelihoods of recognized words between each separated channel in speaker independent recognition systems against a command list or relative distance scores from templates in speaker dependent systems, a switching decision is made on a moving time horizon basis as soon as the relative score factors exceed a certain threshold. In this way a desired speaker's motion in space can be suitably tracked in the separated channels.

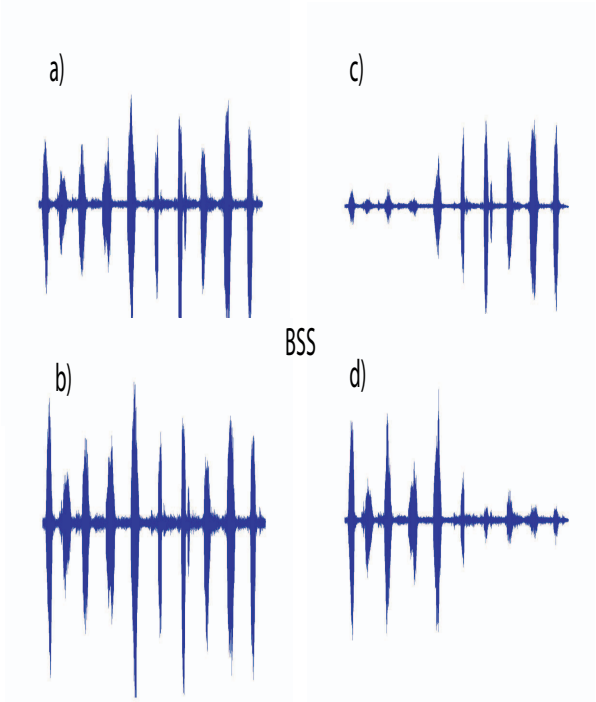

Fig. 1. Channel switching of separated solution: recorded left microphone signal a), recorded right microphone signal b), separated left signal c), separated right signal d)

4. EXPERIMENTS

The scheme depicted in Figure 2 is evaluated in the following experiments. Commands were uttered in an office environment towards a two microphone setup. Different scenarios were investigated as illustrated by Figure 3. In a first static scenario, a human speaker is uttering a sequence of command words at a distance of about 1 m on the left side of the center divide line between two microphones arranged 15 cm apart. A loudspeaker located at the same distance from the microphones is playing different noise sound files from a computer. Three different noise cases are considered: 1) military tank noise containing gunshot salves, 2) music and 3) babble noise. A total of about 120 words from a list of 10 commands are recorded in several sentences for each noise case and processed by a voice recognition system to quantify the obtained speech enhancement. In a

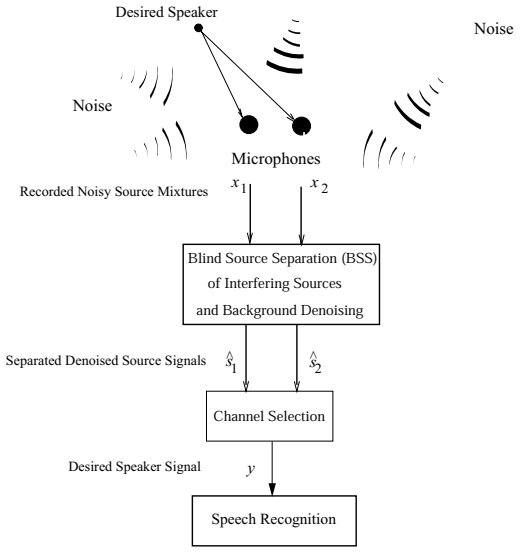

Fig. 2. Proposed Speech Enhancement Scheme

second scenario the adaptability of the BSS channel selection scheme is tested in a mobile scenario. While the loudspeaker remains immobilized at its previous position, the human speaker is uttering command sentences where some of them are spoken in its initial position. He then moves towards his end position illustrated by Figure 3 while issuing commands and finally remains positioned close the right microphone. The correct separated channel was selected through the use of a previously trained speaker dependent voice recognition system. The approach uses energy based as well as zero crossing rate measurements to discriminate noise from speech intervals.

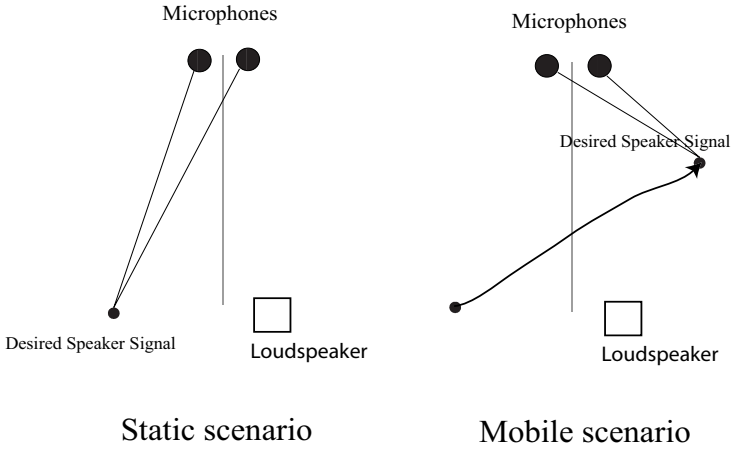

Fig. 3. Experimental setup: static scenario configuration on the left and mobile scenario on the right

The proposed scheme was compared to spectral subtraction type algorithms [1] with suitable adaptation of noise over- subtraction factors [9]. Beamforming where one mixture is delayed and summed to the other to emphasize the desired signal amplitude by in-phase summation is also applied assuming the desired speaker's location to be known. The recorded sound file SNR i.e. before speech enhancement was found to be between 5 and 20 dB. The word recognition accuracy results are displayed in Figure 4 for each noise case in static and mobile scenarios. The result obtained with the adaptive channel selection module is displayed to the far right of each noise case scenario figure. The performance obtained by applying beamforming followed by spectral subtraction to the left and right recorded channel are marked by DNL and DNR respectively; the accuracies found when evaluating the left and right BSS denoised files with no channel selection adaptation are denoted by BSSL and BSSR, respectively. ascBSS refers to the performance measured when adapting the channel selection in the mobile scenario only. Indeed no channel selection was needed for the static scenario. The numerical values for the static case are listed in Table 1 while those for the mobile case are listed in Table 2.

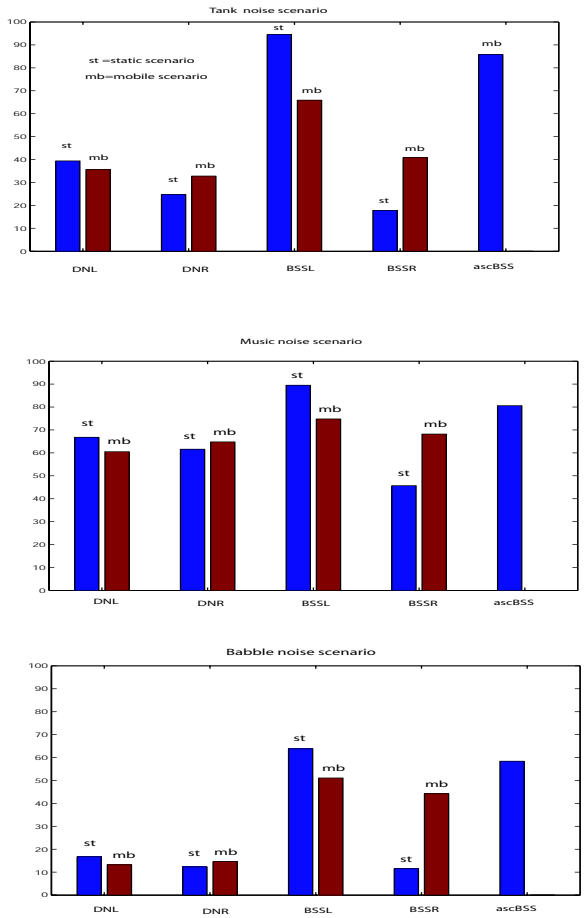

Fig. 4. Word recognition accuracy results

Static Scenario	tank noise	music	babble
DNL	39.4	66.8	16.7
DNR	24.7	61.6	12.3
BSSL	94.5	89.5	63.8
BSSR	17.8	45.7	11.5

Table 1. Word recognition accuracy (%) for various enhancement methods in the static scenario (see text for discussion)

Mobile Scenario	tank noise	music	babble
DNL	35.6	60.5	13.2
DNR	32.8	64.8	14.6
BSSL	65.9	74.8	51.1
BSSR	40.8	68.2	44.2
ascBSS	85.8	80.6	58.4

Table 2. Word recognition accuracy (%) for various enhancement methods in the mobile scenario (see text for discussion)

From Figure 4 and Table 1, we can see that the accuracy is overall better for the left channels in the static scenario since that is the channel recorded closest to the human speaker's position. All three different noise cases exhibit similar qualitative results. The babble noise case exhibits the worst performance overall due to its high incidence of highly interfering speech components. Also, still for the static scenario, one can see that BSS handles the non stationary noise interference much better than spectral subtraction. This is however only true if the left BSS channel is selected. The right BSS channel actually yields worse performance than the spectral subtracted channels since most of the desired speaker's speech content has been removed from this channel. This stresses the importance of a priori knowledge of the desired speaker's position or suitable channel selection procedures which is illustrated for the mobile scenario. As can be seen from Figure 4 and Table 2, the difference between left and right BSS channel recognition accuracy in the mobile scenario is less marked than in the static case. The left channel BSS performance is still better than the right one as most of the command have been uttered on the side closest to the left microphone. Also both BSS channels surpass the accuracy obtained with spectral subtraction. However better results are obtained when the channel selection is included (ascBSS) as indicated by Figure 4. One can see that although the channel selected for speech recognition is adapted, the desired speaker has to pass through the centre divide line for which the mixing situation is singular with respect to the desired speaker. Along with the time to switch to the correct channel, this prevents to reach the performance level obtained in the static setup while significantly improving over the non adapted channel selection case. The channel selection module therefore allows to robustify the performance observed in static settings in mobile environments.

5. CONCLUSIONS

A speech enhancement methodology has been presented that enhances noisy speech signals in subsequent processing stages using a two microphone setup. While the noise reduction stages including blind source separation and background denoising do not require any a priori knowledge about the speech or noise signals involved, a subsequent separated channel selection stage employs a word spotting procedure based on prior knowledge of desired speech content. The scheme's performance was evaluated in a mobile scenario where significant improvement over standard techniques such as spectral subtraction and beamforming was observed. Moreover the proposed scheme with adaptive output channel selection outperformed the equivalent scheme with fixed channel selection. The presented challenges and solutions are currently being investigated for higher order microphone setups.

6. REFERENCES

[1] Ephraim, Y., Van Trees, H.L., A Signal Subspace Approach for speech enhancement, *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251-266, July 1995

[2] Brandstein, M., Silverman, H., A Practical Methodology for Speech Source Localization with Microphone Arrays, Computer, Speech and Language, vol 11, no 2, pp. 91-126, 1997

[3] Hopgood, J.R., Rayner, P.J.W., Yuen, P.W.T, The Effect of Sensor Placement in Blind Source Separation, *Proc. IEEE Workshop on ASPAA*, New York, October 2001

[4] Araki, S., Makino, S., Mukai, R., Saruwatari, H., Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Null Beamformers, Eurospeech2001, vol.4, pp 2595-2598

[5] Visser, E., Lee, T.-W., Speech Enhancement using Blind Source Separation and Two-Channel Energy Based Speaker Detection, Proceedings of ICASSP 2003, I-836-839, Hong-Kong, April 2003

[6] Parra, L., Spence, C., Convolutive Blind Separation of Non-Stationary Sources, *IEEE Trans. on Speech and Audio Proc.*, vol 8., pp. 320-327, 2000

[7] Martin, R., Spectral Subtraction Based on Minimum Statistics, *Proc. EUSPICO'94*, pp. 1181-1185, 1994

[8] Attias, H., Source Separation with a Sensor Array Using Graphical Models and Subband Filtering, *Proceedings NIPS 2002*, pp. 1205-1212, 2003

[9] Hermansky, H., et al., Rasta-PLP Speech Analysis Technique, *Proc. ICASSP*, pp. 121-124, 1992