SPEECH ENHANCEMENT USING BLIND SOURCE SEPARATION AND TWO-CHANNEL ENERGY BASED SPEAKER DETECTION

Erik Visser, Te-Won Lee

Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093-0523
{visser,tewon}@rhythm.ucsd.edu

ABSTRACT

A speech enhancement scheme is presented integrating spatial and temporal signal processing methods for blind denoising in non stationary noise environments. In a first stage, spatially localized point sources are separated from noisy speech signals recorded by two microphones using a Blind Source Separation (BSS) algorithm assuming no a priori knowledge about the sources involved. Spatially distributed background noise is removed in a second processing step. Here, the BSS output channel containing the desired speaker is filtered with a time-varying Wiener filter. Noise power estimates for the filter coefficients are computed from desired speaker absent time-intervals identified by comparing only signal energy of separated source signals from the BSS stage. The scheme's performance is illustrated by speech recognition experiments on real recordings corrupted by babble noise and compared to conventional beamforming and single channel denoising techniques.

1. INTRODUCTION

Speech enhancement in real environments remains a challenging task. Single-microphone enhancement algorithms based on temporal information about the recorded signals are most frequently encountered. They often use a probabilistic framework with statistical models of a single speech signal corrupted by stationary Gaussian noise [1]. While reasonable performance is obtained when the noise is stationary, it deteriorates rapidly when noise power varies importantly or speech mixtures contain significant reverberation. Single channel denoising algorithms based on minimum statistics [2] and Voice Activity Detection (VAD) [3] have been developed to explicitly address non stationary noise. Spatial information about signal mixtures can be exploited by using multiple microphones. In beamforming [4] for example, an array of microphones with a known geometry is used to *suppress* interfering signals. Here, source localization can be performed as well and reverberation be handled with adaptive algorithms [4]. However, these methods usually rely on a priori information about the acoustical environment and sources involved. Also, large microphone arrays are required for good performance whose implementation is difficult and costly.

The number of microphones can be drastically reduced by using blind source separation (BSS) algorithms [5, 6]. The latter exploit spatial information about signal mixtures recorded at different microphone locations to explicitly *separate* interfering noise signals from the desired source signal without assuming any a priori source models. In the following a combined spatial/temporal speech enhancement approach based on BSS is developed.

2. SPEECH ENHANCEMENT SCHEME

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 $n(t)$

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

where P is the convolution order, $\mathbf{A}(\tau)$ is a $m \times m$ mixing matrix. A key distinction is made between spatially point sources $s(t)$ and distributed background noise $n(t)$. Assuming little reverberation, signals originating from point sources can be viewed as identical when recorded at different microphone locations except for an amplitude factor and a delay. The unmixing strategy would consist in finding these latter parameters for each source and summing up the realigned and scaled mixture signals. However background noise originates from a large number of spatially distributed sources resulting in no defined delay and amplitude difference between signals recorded at each microphone. Thus a background noise unmixing strategy poses a singular problem. These different spatial signal characteristics are addressed in subsequent stages of the speech enhancement scheme illustrated in Figure 1.

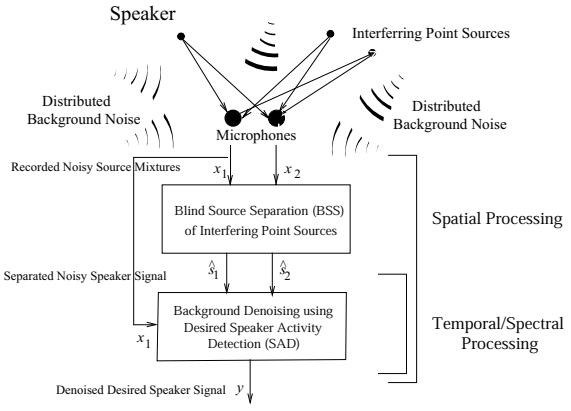

Fig. 1. Proposed Speech Enhancement Scheme

Spatial information about interfering point sources is processed in the blind source separation unit while the remaining stage removes distributed background noise by a mixed temporal/spatial processing approach.

3. BLIND SOURCE SEPARATION (BSS) OF INTERFERING POINT SOURCES

In recent years a number of algorithms have emerged implementing blind source separation of mixture signals by decorrelating their higher-order statistics [6]. However the second order decorrelation approach presented in [5] yielded the most consistent performance in our experiments. The Multiple Adaptive Decorrelation (MAD) algorithm [5] 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 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. The unmixing filter computation is executed in the frequency domain where $\mathbf{X}(\omega,t) \simeq \mathbf{A}(\omega)\, \mathbf{S}(\omega,t)$, $X(\omega,t)$ being the spectrogram obtained by consecutively computing the Short Time Fourier Transform of length T (where $T >> P$, the convolution order), of $x(t)$ at each time instant t in an overlap-shift fashion [5]. If 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 $\hat{\Lambda}_s(\omega,t) = \mathsf{E}\left[\mathbf{S}(\omega,t)\, \mathbf{S}^H(\omega,t)\right]$, we get $\mathbf{W}(\omega)$ from

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

The first constraint imposes that the filter length Q be much smaller than T to solve the frequency permutation problem [5]. Also scaling issues are solved by the second constraint fixing the diagonal elements of the filter matrices to unity. The final learning rule is $\Delta \mathbf{W}^*(\omega) \sim \left(\mathbf{W} \hat{R}_x(\omega,t) \mathbf{W}^H - \hat{\Lambda}_s(\omega,t)\right) \mathbf{W}(\omega)\, \hat{R}_x(\omega,t)$. It is noted that the second constraint in problem (1) ensures that the dominant speaker voice will be separated at the microphone position at which its amplitude is highest during most of the signal length [7], thereby determining the desired speaker containing output source. The approach has shown robust performance in a number of applications e.g. car environments [7]. Figure 2 illustrates how a desired speaker signal (digit utterance in a noisy office environment) is separated from an interferring point source by applying the BSS algorithm. However, both separated source files still contain the original baseline background noise. In the following, the background denoising stage (see Figure 1) is addressed.

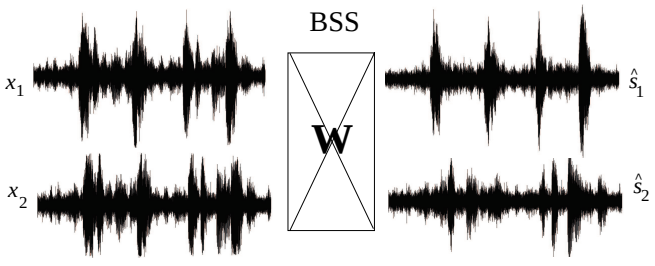

Fig. 2. Blind Source Separation (BSS) of Interfering Point Sources: Input recorded noisy source mixtures (left) and output separated noisy sources (right)

4. BACKGROUND DENOISING

The key to efficient denoising lies in accurately tracking non stationary noise power.

4.1. Standard Denoising Techniques

Two fundamental methods have emerged for determining time-varying noise power and are based on Voice Activity Detection (VAD) [3] or minimum statistics [2]. VAD approaches continuously track the measured noisy signal power and perform weighted noise power updates depending on the probability a speech interval has been detected. The drawback of these methods is that extensive a priori models are necessary to discriminate between speech/non-speech intervals and robustness is not guaranteed in the presence of speech containing disturbances like babble noise. 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 [2]. If the receding time horizons are chosen appropriately, robust and conservative denoising performance is obtained.

4.2. Desired Speaker Activity Detection (SAD)

Denoising performance can be significantly improved if time-intervals containing noisy desired speaker speech samples are differentiated from noise-only intervals when estimating non stationary noise power. In the following we propose a new robust, model-independent measure based on two channel information to detect desired speaker containing time-intervals. If the energy of separated BSS channel $\hat{s}_i(t)$ over a time frame T is given by

$$E_T\left(\hat{s}_i(t)\right) = \sum_{t}^{t+T} \left\|\hat{s}_i(t)\right\|^2 dt,$$

a two-channel energy ratio factor $\lambda(t)$ can be defined as

$$\lambda(t) = \quad (2)$$
$$\mathbf{exp}\left(\frac{-\nu * \mathbf{max}\left[E_T\left(\hat{s}_2(t)\right) - \xi \Delta E_T\left(x_2, \hat{s}_2\right)(t), \epsilon\right]}{\mathbf{max}\left[E_T\left(\hat{s}_1(t)\right) - E_T\left(\hat{s}_2(t)\right) + \xi\, \Delta E_T\left(x_2, \hat{s}_2\right)(t), \epsilon\right]}\right)$$

and computed over the whole signal length in an overlap-add fashion with shifting window of size T. The basic motivation for λ is given first before discussing the remaining terms in (2).

In the case of mixtures corrupted by distributed background noise only, the desired speaker will be isolated into BSS channel 1 and removed from BSS channel 2 which will solely contain distributed background noise. Hence, by neglecting the term in ΔE ($\xi = 0$) (as well as the **max** operation) and considering $\nu = 2$, the first-order Taylor expansion of $\lambda(t)$ yields

$$\lambda(t) = 1 - \frac{E_T\left(\hat{s}_2(t)\right)}{E_T\left(\hat{s}_1(t)\right)}.$$

The corresponding expression in the frequency domain with the noise variance in individual spectral subbands computed from the variance of $\hat{S}_2(\omega,t)$ and the noise plus speech variance from $\hat{S}_1(\omega,t)$ is equivalent to a Wiener filter coefficient. Similarly the corresponding expression for (2) is analog to a generalized Wiener filter gain function [1]. However, experiments have shown that reliable filters cannot be directly estimated from the ratio of spectral

subband channel energies. In fact, although the distributed background noise energy integrated over all spectral bands in a given time-interval of $\hat{s}_2$ is similar to the *overall* background noise energy in the same time-interval in $\hat{s}_1$, this is not true when *individual* spectral subbands are considered.

Instead the overall energy ratio (2) over a time frame T can be used to detect *desired speaker activity*. Indeed, since it is assumed that background noise energies are similar in each recorded mixture when microphones are positioned close enough and the overall energy is preserved from recorded to separated sources because of the scaling constraint in (1), the denominator in (2) is close to zero and hence λ tends to zero when the desired speaker is absent. If it is present, the energy in BSS channel 1 is much larger than in channel 2, the quotient in (2) tends to zero and thus λ to 1. In practice, BSS channel 2 may contain an interferring point source eliminated from channel 1. Therefore the ΔE term with

$$\Delta E_T\big(x_2, \hat{s}_2\big)(t) \quad = \quad \mathbf{max}\Big[\sum_{t}^{t+T}\Big(\|x_2(t)\|^2 \; - \; \|\hat{s}_2(t)\|^2\Big)\, dt, \epsilon\Big]$$

is introduced in (2) to robustify the detection of desired speaker by explicitly tracking energy changes from recorded channel 2 to separated BSS channel 2. Factor $\xi \;=\; \frac{\sum_t^{t+T}\|x_1\|^2\, dt}{\sum_t^{t+T}\|x_2\|^2\, dt}$ scales the energy change in channel 2 to a corresponding energy change in channel 1. The parameter ν allows to adjust the "sharpness" of speech/non speech interval delimitation.

The resulting $\lambda(t)$ is used to provide a probability measure for the speaker's presence. The noise estimate is given by

$$\Phi_n(\omega, T+1) \quad = \quad z_a * \Phi_n(\omega, T) + (1 - z_a) * X_c(\omega, T)$$

where z_a is a smoothing constant and $X_c(\omega, T)$ is the auto-correlation of $\big(1 - \lambda(t)\big)\; \hat{S}_1(\omega, t)$, t in time frame T. The current speech plus noise power estimate is obtained from the recurrence

$$\Phi_{s+n}(\omega, T+1) \quad = \quad z_g * \Phi_{s+n}(\omega, T) + (1 - z_g) * X(\omega, T)$$

where z_g is a smoothing constant and X is the auto-correlation of $\hat{S}_1(\omega, t)$, t in time frame T. The noise and speech+noise power estimates are used to compute the Wiener filter coefficients $g(\omega, T)$ for each frame. Finally the denoised speech spectrum is obtained from

$$Y(\omega, T) \quad = \quad g(\omega, T)\; \hat{S}_1(\omega, T)$$

with filter coefficients

$$g(\omega, T) \quad = \quad \sqrt{1 - \frac{\Phi_n(\omega, T)}{\Phi_{s+n}(\omega, T)}}.$$

It was observed that using $\lambda(t)$ directly from (2) resulted in too agressive denoising performance since the value of $\lambda(t)$ is not necessarily one at each local maximum and may decrease too rapidly near the edges of detected time-intervals, thereby cutting off speech parts. Hence $\lambda(t)$ is refined by replacing it by a sequence of Hanning windows with centers determined by the local maxima of $\lambda(t)$ from (2) and widths given by twice the distance between symmetric points around each maximum where $\lambda(t)$ reaches a certain threshold β. The resulting curve (see Figure 3) is smooth and sufficiently wide to avoid cutting off edges of desired speaker parts. The denoised speech signal is shown in Figure 5, case (f), where it is compared to other denoising approaches. Quantitative comparison to standard methods is presented in the next section.

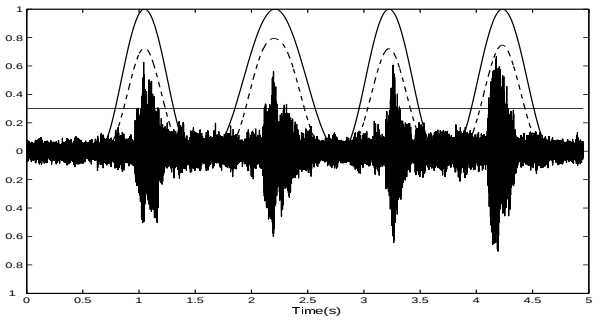

Fig. 3. Separated BSS channel 1 with corresponding λ (dashed line) from (2) and $\lambda(t)$ after refinement (full line); threshold β indicated by horizontal line (see text)

5. EXPERIMENTS

Recordings were taken in an $3m \times 4m \times 6m$ room with two directional microphones separated by 10 cm mounted on a desk. The desired speaker was sitting 30 cm from the microphone setup (closer to left microphone) and uttering continuous digit sentences while 4 loudspeakers positioned in each room corner were playing an identical sound file containing a mixture of babble and white noise to generate spatially distributed background noise. Also an additional loudspeaker was put at 30 cm distance from the right side microphone playing a prerecorded word sequence to create an interferring point source. The speech recognizer as well as a multiple noise condition database for training the HMM models was provided by the AURORA 2 benchmark dataset. The feature extraction front-end FE_v2_0 (AURORA 2) was used to compute 39 MFCCs (including energy, delta, delta-delta). The test database consisted of files recorded at different SNR dB levels (from -5 to 10 dB). 100 digit sentences, each containing a maximum of 4 digits, were recorded for each SNR case.

The proposed scheme was compared to standard speech enhancement methods like delay-and-sum beamforming and minimum statistics type denoising like Martin's algorithm [2]. In the spatial processing method of beamforming (BF), one mixture is delayed and summed to the other based on the desired speaker's known location to emphasize the desired signal amplitude by inphase summation. Whereas better beamforming methods exist, emphasis is put in this study on comparison of largely "blind" enhancement techniques. Martin's single channel type denoising algorithm (DN) was preferred over model-based VAD techniques for the same reason. In Table 1 and Figures 4 & 5, speech recognition on recorded files (REC) is evaluated against the conventional scheme (BF+DN), BSS, BSS followed by Martin's algorithm (BSS+DN), BSS+DN followed by Speaker Activity Detection (BSS+DN+SAD) and BSS followed by SAD (BSS+SAD).

First Figure 4 clearly indicates that speech recognition on the unprocessed recorded files leads to unacceptable performance even at high SNR. The conventional approach (BF+DN) yields unsatisfactory results as the interferring point source was not removed efficiently (see case (b) in Figure 5). Blind source separation alone (curve BSS and case (c) in Figure 5) lifts the recognition rate by at least 20 %. Finally curve BSS+SAD and case (f) in Figure 5 illustrates BSS followed by SAD yielding the best performance.

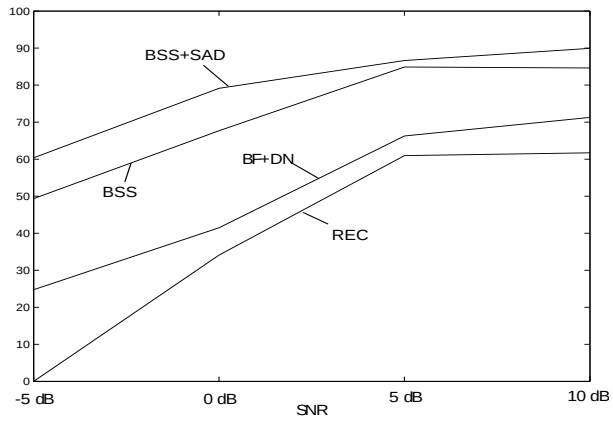

Fig. 4. Word recognition accuracy for standard (BF+DN) and proposed scheme (BSS+SAD) (REC=recorded, unprocessed case)

To compare the background denoising techniques, performance of all strategies involving BSS is analyzed in Table 1. For high SNR (5;10 dB), one obtains similar ($\leq$ 1% difference in) performance when processing signal $\hat{s}_1$ with conventional denoising (BSS+DN) or SAD schemes. Indeed the signal has been considerably enhanced by removing the interferring point source and the remaining background noise level is too low to cause significant speech deterioration. However, at difficult SNR (-5;0 dB) levels, highly non stationary background noise components cannot be sufficiently eliminated with the minimum statistics approach (see case (d) in Figure 5). On the contrary, the two channel information based SAD approaches achieve the necessary denoising in noise-only intervals (cases (e) and (f) in Figure 5) and outperform conventional denoising (BSS+DN) considerably ($\sim$ 5-10 % accuracy increase). This shows the benefit of non stationary noise estimates determined from two channel information over single channel, minimum noise power averaged over a long time interval. Finally, the superior performance of BSS+SAD over BSS+DN+SAD in low SNR cases suggests that less agressive denoising in digit containing time-intervals preserves more desired speech information and/or induces less artifacts. The best reference accuracy achieved was 90.84 % on 100 digit sentences recorded with the same microphone setup in the silent office environment. This reflects the effects of room reverberation, speaker and recording equipment different from the ones used in the AURORA 2 database.

SNR [dB]	-5	0	5	10
REC	0	34.10	60.96	61.71
BF+DN	24.75	41.48	66.25	71.28
BSS	49.34	67.68	84.89	84.63
BSS+DN	50.64	74.30	87.15	89.67
BSS+DN+SAD	56.96	77.10	86.65	89.67
BSS+SAD	60.37	79.13	86.54	89.92

Table 1. Word recognition accuracy (%) for various denoising schemes (see text for discussion)

6. CONCLUSIONS

A spatio-temporal speech enhancement scheme has been presented that enhances noisy speech signals in two subsequent processing stages using only two microphones and no a priori models about the speech and noise sources involved. First the desired speaker is separated from interfering, spatially localized point sources using a blind source separation algorithm. In a second step, spatially distributed background noise is removed using energy information from both separated BSS output channels to detect noise-only intervals and compute a non stationary noise estimate to design a time-varying Wiener filter. In speech recognition experiments carried out in a noisy office environment, the scheme was shown to yield significant enhancement over standard methods such as beamforming and single channel Wiener filtering based on minimum noise statistics. Denoising in the proposed scheme is independent of the background noise spectral content since detection of desired speaker speech activity is based on energy comparison between two channels only. As no a priori knowledge is used in the BSS stage either, the scheme is suitable for environment-independent speech enhancement and recognition tasks.

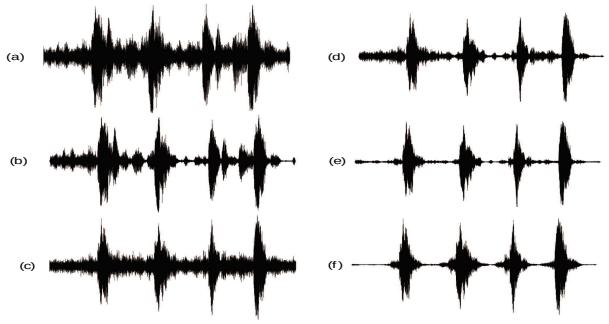

Fig. 5. Comparison of different denoising strategies on a recorded digit utterance example (transcript 0020): (a) REC, (b) BF+DN, (c) BSS, (d) BSS+DN, (e) BSS+DN+SAD, (f) BSS+SAD

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

[3] Sohn, J., Kim, N.S., Sung, W., A Statistical Model-Based Voice Activity Detection, *Sig. Proc. Let.*, vol 6., pp.1-3, 1999

[4] 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

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

[6] Bell, A.J., Sejnowski, T.J., An Information-Maximisation Approach to Blind Separation and Blind Deconvolution, Neural Computation, 7(6), pp. 1004-1034, 1995

[7] Visser, E., Otsuka, M., Lee, T.-W., A Spatio Temporal Speech Enhancement Scheme for Robust Speech Recognition, *Proceedings ICSLP2002*, pp. 1821-1824, September 2002