Two-channel speech denoising through minimum tracking

Two-channel speech denoising through minimum tracking

S. Srinivasan, K. Janse, M. Nilsson and W. Bastiaan Kleijn

A blind two-channel interference reduction algorithm to suppress localised interferers in reverberant environments is presented. The algorithm requires neither knowledge of source positions nor a speech-free noise reference. The goal is to estimate the speech signal as observed at one of the microphones, without any additional filtering effects that are typical in convolutive blind source separation.

Signal model: We assume a single speech source in the presence of a localised interference. For a two microphone system, the reverberant noisy observation can be written in the frequency domain as

$$\mathbf{Y}(\omega) = A(\omega)\mathbf{X}(\omega) + \mathbf{U}(\omega) \tag{1}$$

where $\mathbf{Y}(\omega) = [Y_1(\omega) \quad Y_2(\omega)]^T$ is the vector of observed microphone signals, $\mathbf{X}(\omega) = [S(\omega) \quad N(\omega)]^T$, $S(\omega)$ corresponds to the speech signal, $N(\omega)$ corresponds to the interference, and $\mathbf{U}(\omega) = [U_1(\omega) \quad U_2(\omega)]^T$ corresponds to the uncorrelated noise at the two sensors. $A(\omega)$ is the 2×2 mixing matrix. We assume that the speech and interference signals are statistically independent.

Following [1], we assume an unmixing matrix of the form (see *Note* at and of Letter)

$$W(\omega) = \begin{bmatrix} 1 & \alpha(\omega) \\ \beta(\omega) & 1 \end{bmatrix} \tag{2}$$

Let

$$\mathbf{Z}(\omega) \triangleq [Z_1(\omega) \quad Z_2(\omega)]^T = W(\omega)\mathbf{Y}(\omega) \tag{3}$$

Blind source separation (BSS) algorithms such as those in [1] and [3] estimate $W(\omega)$ such that $Z_1(\omega)$ and $Z_2(\omega)$ contain the individual separated signals, respectively. In the specific case of the noise reduction problem considered in this Letter, we only estimate $\alpha(\omega)$ such that $Z_1(\omega)$ is noise-free. The estimation of $\alpha(\omega)$ is discussed in the following Section. If both input channels contain a mixture of speech and interference signals, as is the case here, the estimated clean speech component $Z_1(\omega)$ is unique only up to a scaling factor [1]. This corresponds to an undesired filtering of the separated time-domain signal. The main contribution of this Letter is to compensate for this undesired effect. We propose a postprocessing step to ensure that the recovered signal is not only interference-free but also identical to the speech signal $a_{11}(\omega)S(\omega)$ observed at the microphone, where $a_{ij}(\omega)$ is the (i, j)th entry of the 2×2 mixing matrix $A(\omega)$ in (1). The postprocessing is discussed in the penultimate Section.

Denoising through minimum tracking: In this Section, we first reduce the problem of two-channel denoising to one of minimum tracking. The tracking itself can then be performed using well-known methods such as [4]. Let $\mathbf{Y}^n(\omega)$ and $\mathbf{Z}^n(\omega)$ denote the values of $\mathbf{Y}(\omega)$ and $\mathbf{Z}(\omega)$ during time intervals of speech absence. $\alpha(\omega)$ may be estimated by minimising the energy $\eta_1(\omega) = E[Z_1^n(\omega)Z_1^{n*}(\omega)]$, where superscript * denotes complex conjugate transpose and E is the statistical expectation operator. From (2) and (3), we have

$$Z_1(\omega) = Y_1(\omega) + \alpha(\omega)Y_2(\omega) \tag{4}$$

We can estimate $\alpha(\omega)$ by minimising $\eta_1(\omega)$, which amounts to minimising the energy of the interference component in the output signal. Thus we have,

$$\hat{\alpha}(\omega) = \arg\min_{\alpha(\omega)} \eta_1(\omega) = \frac{-R_{Y^n}^{12}(\omega)}{R_{Y^n}^{22}(\omega)} \tag{5}$$

$R_{Y^n}^{ij}$ corresponds to the (i, j)th element of $R_{Y^n}(\omega) = \mathrm{E}[\boldsymbol{Y}^n(\omega)\boldsymbol{Y}^{n*}(\omega)]$. $R_{Y^n}^{22}(\omega)$ is larger than zero and (5) is well defined if we assume $\mathrm{E}[U^2(\omega)U^{2*}(\omega)] > 0$, which can easily be validated in practice by adding a small amount of uncorrelated noise to the microphone signals.

$R_Y^{11}(\omega)$ and $R_Y^{22}(\omega)$ are both real quantities, and under an additive interference model, attain their minimum values when the speech signal is absent. Thus, by tracking the minimum of either $R_Y^{11}(\omega)$ or $R_Y^{22}(\omega)$, frequency bins that contain only interference can be identified, from which $R_{Y^n}(\omega)$ can be estimated. The minimum tracking is performed using the well-known minimum statistics algorithm [4], where a buffer of D past cross-spectral densities is maintained for each frequency bin and the minimum is tracked in this buffer. The buffer size D should be large enough to include non-speech regions and small enough to account for non-stationary channel conditions.

In the absence of the uncorrelated noise $\mathbf{U}(\omega)$, it is easy to see from (1) and (5) that optimally, $\hat{\alpha}(\omega) = -a_{12}(\omega)/a_{22}(\omega)$, where $a_{ij}(\omega)$ is the (i, j)th entry of $A(\omega)$. Using this optimal value in (4) cancels out the interference component and yields

$$Z_1(\omega) = \frac{a_{11}(\omega)a_{22}(\omega) - a_{12}(\omega)a_{21}(\omega)}{a_{22}(\omega)}S(\omega) \tag{6}$$

In the presence of the uncorrelated noise $\mathbf{U}(\omega)$, we have optimally,

$$\hat{\alpha}^{\mathrm{uncorr}}(\omega) = \frac{-a_{12}(\omega)}{a_{22}(\omega)} \frac{1}{\left(1 + \dfrac{\sigma_u^2(\omega)}{|a_{22}|^2\sigma_n^2(\omega)}\right)} \tag{7}$$

where $\sigma_u^2(\omega) = \mathrm{E}\{U_2(\omega)U_2^*(\omega)\}$ and $\sigma_n^2(\omega) = \mathrm{E}\{N(\omega)N^*(\omega)\}$. We see from (7) that we require $\sigma_n^2(\omega) \ll \sigma_n^2(\omega)$ to ensure that $\hat{\alpha}^{\mathrm{uncorr}}(\omega) \simeq \hat{\alpha}(\omega)$, which is valid in applications where a strong localised interference is to be suppressed. If the interferer is inactive, then $R_{Y^n}^{12}(\omega) = \mathrm{E}[U_1(\omega)U_2^*(\omega)] = 0$ so that $\hat{\alpha}(\omega) = \hat{\alpha}^{\mathrm{uncorr}}(\omega) = 0$ and (6) still holds.

Solving filtering ambiguity of blind source separation: The estimate of the speech signal in (6) corresponds to a filtered version of the original speech signal. Instead of this arbitrary filtering, it is desirable to obtain an estimate $a_{11}(\omega)S(\omega)$ which corresponds to the clean speech signal as observed at the microphone. One approach is to apply the minimal distortion principle [5] using the new unmixing matrix $W^{\mathrm{opt}}(\omega) = \mathrm{diag}(W^{-1}(\omega))W(\omega)$. In [3, 6], an equivalent postprocessing step is suggested where the separated signal is multiplied by $1/1 - \alpha(\omega)\beta(\omega)$. This, however, requires knowledge of $\hat{\beta}(\omega)$.

Instead, we propose to introduce an adaptive filter as shown in Fig. 1. The filter is adapted such that the expected energy of the residual signal is minimised and is implemented as a normalised least mean squares (NLMS) filter. The optimal solution for the filter is given by (assuming $\sigma_u^2(\omega) \ll \sigma_n^2(\omega)$)

$$\begin{aligned} H_{\mathrm{opt}}(\omega) &= \arg\min_{H(\omega)} E[|Y_1(\omega) - H(\omega)Z_1(\omega)|^2] \\ &= a_{11}(\omega)\frac{a_{22}(\omega)}{a_{11}(\omega)a_{22}(\omega) - a_{12}(\omega)a_{21}(\omega)} \end{aligned} \tag{8}$$

Using (6) and (8), the output $\hat{S}(\omega)$ of the adaptive filter (before the subtraction) becomes

$$\hat{S}(\omega) = H_{\mathrm{opt}}(\omega)Z_1(\omega) = a_{11}(\omega)S(\omega)$$

We note that the filter may be continuously adapted, even when the interference is active, as $Z_1(\omega)$ contains only the desired signal. The procedure described above addresses the filtering ambiguity of blind source separation (BSS). We note that our approach does not suffer from the permutation problem of BSS since we explicitly estimate the desired signal through the energy minimisation procedure.

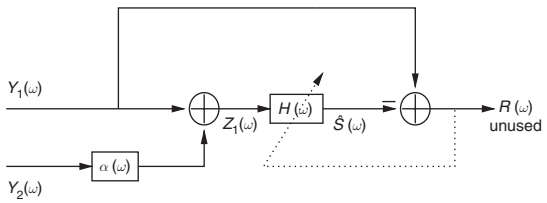

Fig. 1 *Estimating speech signal without filtering ambiguity of blind source separation ($\hat{S}(\omega)$ is desired estimate)*

Experimental results: Experiments were performed to validate the proposed method. Two omnidirectional microphones were placed 5 cm apart in an office room with a reverberation time of around 400 ms. Speech and interference signals were played from loudspeakers placed at two different locations (speech at $+45^\circ$ and interference at -45° relative to the centre of the microphone array). The speech and interference signals were recorded separately and then added together to obtain the noisy signal. A single 30 second-long speech sample, and two

different interference types, white noise and keyboard clicks, were used. The noisy signals were processed by the proposed method. Such a framework, in which the individual speech and interference signals are available, allows measurement of the improvement in the signal-to-interference ratio (SIR).

A sampling frequency of 16 kHz was used. $\hat{\alpha}(\omega)$ was obtained in the frequency domain using (5). A frame length of 1024 samples was used and the cross spectral density matrix was computed by averaging over five neighbouring frames. A 128-tap time-domain filter was then obtained by shifting, windowing (Hann) and truncating the inverse-DFT of $\hat{\alpha}(\omega)$. This filter was applied to $y_2(t)$, where the lower case symbol refers to the time-domain signal corresponding to the respective frequency-domain signal. The resulting signal was added to $y_1(t)$ to obtain the separated speech estimate $z_1(t)$. In the next stage, the compensation filter $H(\omega)$ (see Fig. 1) was realised as a 32-tap time-domain adaptive NLMS filter and was applied to $z_1(t)$ to obtain the desired signal $\hat{s}(t)$ in the time domain.

To obtain the SIRs, the $\hat{\alpha}(\omega)$ and $H(\omega)$ that were estimated using the noisy signals were applied separately to the clean speech and interference signals to obtain $z_1^s(t)$ and $z_1^n(t)$, respectively. The output SIR was then calculated as $SIR_{\text{out}} = 10\log_{10}\sum z_1^s(t)^2 / \sum z_1^n(t)^2$. The input signals were mixed such that the input SIR was 10 dB. The improvement in SIR due to processing (difference between output and input SIR) is reported in Table 1 for two different interference types, white noise and keyboard clicks. For comparisons, results obtained using the BSS method of [1] are also provided. An unmixing matrix $W_{\text{ref}}(\omega)$ was first obtained following [1]. For a fair comparison, the minimal distortion principle (MDP) described in [5] was then applied to compensate for the arbitrary filtering of the separated speech signal resulting in the unmixing matrix $\text{diag}(W_{\text{ref}}^{-1}(\omega))W_{\text{ref}}(\omega)$.

Table 1: Improvement in SIR (dB) corresponding to proposed method and reference method (BSS method of [1] followed by MDP approach [5])

Interference	Proposed	Ref. method
White noise	14.1	9.2
Keyboard clicks	16.9	8.9

We also measured the log spectral distortion between the clean speech signal and the enhanced signals, and the results are shown in Table 2. It can be seen that the proposed method results in lower distortion.

Table 2: Log spectral distortion (dB) corresponding to proposed method and reference method (BSS method of [1] followed by MDP approach [5])

Interference	Proposed	Ref. method
White noise	5.1	6.7
Keyboard clicks	4.1	6.2

Conclusion: A two microphone blind noise reduction algorithm is presented to suppress localised interferences. The method relies on minimum tracking to achieve the denoising and incorporates a final adaptive filtering step to compensate for the problem of arbitrary filtering that BSS techniques suffer from. Experiments show an improved signal-to-interference ratio and low signal distortion compared to the reference method.

Note: The form of the unmixing matrix in (2) assumes that both the microphone signals, $Y_1(\omega)$ and $Y_2(\omega)$, contain the interference signal. Such an assumption is valid in most practical microphone array configurations. For applications where only $Y_1(\omega)$ contains the interference signal during certain time intervals and at certain frequencies, the following unmixing matrix suggested in [2], which has the same degrees of freedom as the matrix in (2), may be employed:

$$\tilde{W}(\omega) = \begin{bmatrix} 1-\alpha(\omega) & \alpha(\omega) \\ \beta(\omega) & 1-\beta(\omega) \end{bmatrix}$$

For simplicity of notation, we retain $W(\omega)$ given by (2) as our unmixing matrix. As both $W(\omega)$ and $\tilde{W}(\omega)$ have the same degrees of freedom, the derivations in the remainder of this Letter can be applied to the case of $\tilde{W}(\omega)$ as well.

30 September 2009
doi: 10.1049/el.2010.2765

S. Srinivasan and K. Janse (*Digital Signal Processing Group, Philips Research Laboratories, High Tech Campus 36, Eindhoven, AE 5656, The Netherlands*)

E-mail: sriram.srinivasan@philips.com

M. Nilsson (*Skype Technologies, Stadsgården 6, Stockholm 116 45, Sweden*)

W. Bastiaan Kleijn (*School of Electrical Engineering, KTH Royal Institute of Technology, Osquldas v. 10, Stockholm 100 44, Sweden*)

References

1 Parra, L., and Spence, C.: 'Convolutive blind separation of non-stationary sources', *IEEE Trans. Speech Audio Process.*, 2000, **8**, (3), pp. 320–327
2 Srinivasan, S., Nilsson, M., and Kleijn, W.B.: 'Denoising through source separation and minimum tracking'. Proc. Interspeech, Lisbon, Portugal, September 2005, pp. 2349–2352
3 Gerven, S.V., and Compernolle, D.V.: 'Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness', *IEEE Trans. Signal Process.*, 1995, **43**, (7), pp. 1602–1612
4 Martin, R.: 'Noise power spectral density estimation based on optimal smoothing and minimum statistics', *IEEE Trans. Speech and Audio Process.*, 2001, **9**, (5), pp. 504–512
5 Matsuoka, K., and Nakashima, S.: 'Minimal distortion principle for blind source separation'. Proc. Int. Conf. ICA and BSS, San Diego, CA, USA, December 2001, pp. 722–727
6 Weinstein, E., Feder, M., and Oppenheim, A.V.: 'Multi-channel signal separation by decorrelation', *IEEE Trans. Speech Audio Process.*, 1993, **1**, (4), pp. 405–413