MOVING SOURCE SPEECH ENHANCEMENT USING TIME-DELAY ESTIMATION

Zohra Yermeche, Nedelko Grbić and Ingvar Claesson

zye@bth.se, ngr@bth.se, icl@bth.se
Blekinge Institute of Technology
School of Engineering
37225 Ronneby, Sweden

ABSTRACT

This paper presents a new constrained subband beamforming algorithm to enhance speech signals generated by a moving source in a noisy environment. The beamformer is based on the principle of a soft constraint calculated from an estimated source position. The soft constraint secures the spatial-temporal passage of the desired source signal in the adaptive update of the beamforming weights and guaranties the full rank property of the covariance matrix inverted in the update. This approach allows for an efficient adaptation of the beamformer to speaker movement by using a tracking algorithm for sound source time-delay estimation. The proposed method has the benefit of taking into consideration the discrepancies in the acoustical environment model as well as errors in the time-delay estimation. Evaluation in a real environment with a moving speaker in a hands-free situation shows up to 10 dB noise suppression and 20 dB interference suppression within the conventional telephone bandwidth. This is achieved with a negligible impact on target signal distortion.

1. INTRODUCTION

Microphone arrays in conjunction with digital beamforming techniques have been extensively exploited for speech enhancement in hands-free applications, such as conference telephony, speech recognition and hearing aid devices [1]. In a hands-free environment, microphones are placed at a remote distance from the speakers causing problems of room reverberation, noise and acoustic feed-back. Successful microphone array processing of speech should achieve speech dereverberation, efficient noise and interference reduction, and should also provide an adaptation capacity to speaker movement.

In the microphone array literature many algorithms address these issues separately. Beamforming techniques using optimal filtering or signal subspace concepts have been suggested [1, 2]. Many of these algorithms rely on voice activity detection (VAD). This is needed in order to avoid source signal cancellation effects [1], which may result in unacceptable levels of speech distortion. Methods based on calibration data have been developed to circumvent the need of a VAD [3, 4]. However, these methods require a re-calibration of the algorithm whenever the location of interest is changed. Blind source separation techniques have open the path to speech separation algorithms such as presented in [5]. Such algorithms do not require any specific knowledge about the source location. However, they present a high computational complexity making a real time implementation impractical.

This paper presents a new constrained subband beamforming algorithm to enhance speech signals generated by a moving source in a noisy environment without the use of a VAD. Performance of the proposed algorithm is evaluated on real data recordings conducted in an office environment with a moving source. The beamformer is based on the principle of a soft constraint calculated from an estimated source position rather than formed from calibration data as it is done in [4]. This approach allows for an efficient adaptation of the beamformer to speaker movement by using a tracking algorithm for sound source localization.

The time-delays estimated by the localization algorithm, corresponding to the speaker's location, are used directly in the beamformer through source covariance estimates. The objective of the beamformer is formulated in the frequency domain as a weighted recursive least squares solution. In order to track variations in the surrounding noise environment, the proposed algorithm continuously estimates the spatial information for each frequency band, based on the received data.

Similarly to blind speech separation methods, the proposed method has the benefit of taking into consideration the discrepancies in the acoustical environment model, such as microphone mismatches and gain variations, as well as errors in the source localization. This is achieved by specific modelling of uncertainties in the desired signal array response vector as an area expansion of the source. The computational complexity of the beamformer is substantially reduced by using a subband beamforming scheme which allows for a real time implementation.

2. PROPOSED ALGORITHM

2.1. Signal Model

The target speech source $s(t)$ is assumed to be a wideband source located in the near field of a uniform linear array of I sensors. The $I \times 1$ array input vector at a specific angular frequency $\Omega = 2\pi f$, when all the sources are active simultaneously, is

$$\mathbf{x}^{(\Omega)}(t) = \mathbf{x}_s^{(\Omega)}(t) + \mathbf{x}_i^{(\Omega)}(t) + \mathbf{x}_n^{(\Omega)}(t), \qquad (1)$$

where $\mathbf{x}_s^{(\Omega)}(t)$, $\mathbf{x}_i^{(\Omega)}(t)$ and $\mathbf{x}_n^{(\Omega)}(t)$ are the $I \times 1$ received microphone input vectors generated by the target source, the interfering sources and the ambient noise, respectively.

2.2. Soft Constrained Beamformer

The output of the beamformer, for a frequency Ω and at a sample instant n, is given by

$$y^{(\Omega)}(n) = \mathbf{w}_n^{(\Omega)^H} \mathbf{x}^{(\Omega)}(n)\,, \qquad (2)$$

where $\mathbf{w}_n^{(\Omega)}$ is the beamformer weight vector and $(.)^H$ stands for the Hermitian transpose. The proposed beamformer is deduced from a least squares formulation of the Wiener solution, with the array weight vector given by

$$\mathbf{w}_n^{(\Omega)} = \left[\mathbf{R}_s^{(\Omega)} + \hat{\mathbf{R}}_{\mathbf{x}}^{(\Omega)}(n)\right]^{-1} \mathbf{r}_s^{(\Omega)}\,, \qquad (3)$$

where

$$\mathbf{R}_s^{(\Omega)} = E\left\{\mathbf{x}_s^{(\Omega)}(n)\,\mathbf{x}_s^{(\Omega)^H}(n)\right\}\,,$$
$$\mathbf{r}_s^{(\Omega)} = E\left\{\mathbf{x}_s^{(\Omega)}(n)\,s^{(\Omega)^H}(n)\right\}\,, \qquad (4)$$

are the spatial source covariance matrix and the spatial cross covariance vector, respectively. The symbol $E\{.\}$ denotes the statistical expectation. The matrix $\hat{\mathbf{R}}_{\mathbf{x}}^{(\Omega)}(n)$ is the received signal covariance matrix estimate continuously calculated from observed data by

$$\hat{\mathbf{R}}_{\mathbf{x}}^{(\Omega)}(n) = \sum_{p=0}^{n} \lambda^{n-p} \mathbf{x}^{(\Omega)}(p) \mathbf{x}^{(\Omega)^H}(p)\,, \qquad (5)$$

where λ is a forgetting factor with the purpose of tracking variations in the surrounding noise environment.
The update of the beamforming weights is done recursively as in [4] by iteratively using the matrix inversion lemma, with the source covariance matrix $\mathbf{R}_s^{(\Omega)}$ constituting a soft constraint. The soft constraint acts as a spatial passband which moderates the weight vector fluctuations generated by the speech discontinuity. Additionally, it forces the full rank of the total matrix to be inverted when no noise, interference or speech is present.

Information about the speech location is put into the algorithm by calculating source covariance estimates for a given location of the source following

$$\tilde{\mathbf{R}}_s^{(\Omega)} = P^{(\Omega)} \int_{\Phi_s} e^{-j\Omega\boldsymbol{\tau}}\, e^{-j\Omega\boldsymbol{\tau}^H}\, d\boldsymbol{\tau}\,,$$
$$\tilde{\mathbf{r}}_s^{(\Omega)} = P^{(\Omega)} \int_{\Phi_s} e^{-j\Omega\boldsymbol{\tau}}\, d\boldsymbol{\tau}\,, \qquad (6)$$

where $P^{(\Omega)}$ is the received source power spectral density at frequency Ω and $\Phi_s = [\boldsymbol{\tau}_s - \frac{\boldsymbol{\Delta}}{2}, \boldsymbol{\tau}_s + \frac{\boldsymbol{\Delta}}{2}]$ is the time-delay range corresponding to an area expansion of the source. Here $\boldsymbol{\tau}_s = [\tau_1, \tau_2, ..., \tau_I]^H$ is the time-delay vector estimate of the direct path from the source position to the microphone array and $\boldsymbol{\Delta}$ is the delay-uncertainty vector.

2.3. Movement Tracking

An SRP-PHAT algorithm for sound source localization [1, 6] is used for speaker movement tracking in conjunction with the filtering operations. This information is exploited to update the spatial source covariance estimates when a movement of the source is detected. Fig. 1 illustrates the principle of time-difference of arrival (TDOA) estimation using a microphone array. The SRP-PHAT algorithm used to estimate the time-delay vector $\boldsymbol{\tau}_s$ is formulated as

$$\boldsymbol{\tau}_s = \arg\max_{\boldsymbol{\tau}} \left\{ \sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} G_{lk}(\omega)\, e^{jw(\boldsymbol{\tau}(l) - \boldsymbol{\tau}(k))}\, d\omega \right\} \qquad (7)$$

where

$$G_{lk}(\omega) = \frac{X_l(\omega) X_k(\omega)^H}{|X_l(\omega) X_k(\omega)^H|} \qquad (8)$$

is the normalized cross power spectrum of the sound signals received by the microphone pair l and k, and $X_i(\omega)$ is the Fourier transform of the signal received by microphone i.

2.4. Subband Beamforming

A multichannel uniform over-sampled analysis filter bank is used to decompose the received array signals into a set of narrow-band signals, prior to the beamformer filtering operations [4, 7], as depicted in Fig. 2. The outputs of the subband beamformers are reconstructed by a synthesis filter bank in order to create a time domain output signal. The analysis and synthesis filter banks constitute a modulation of two prototype filters, which leads to efficient polyphase realization [7]. The spatial characteristics of the input signal are maintained by using the same modulated filter bank at each microphone.

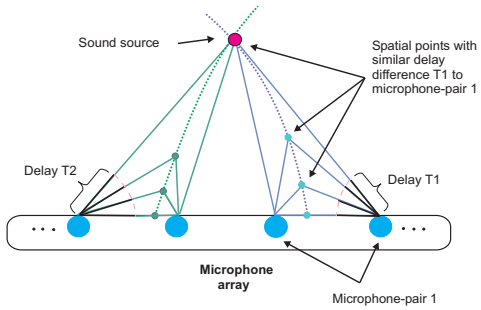

Figure 1: *Time-delay estimation of a point source signal relative to pairs of microphones.*

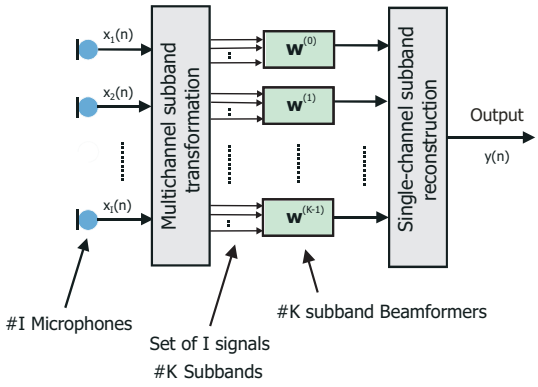

Figure 2: *Subband beamforming structure.*

3. SIMULATIONS AND RESULTS

Simulations include influence of background noise, hands-free interference as well as room reverberation. The data was acquired with a linear array of four microphones uniformly spaced with 5 cm spacing and was gathered on a multichannel DAT-recorder with a sampling rate of 12 kHz. The signal at each microphone was bandlimited to 300-3400 Hz. All simulations were performed with 64 subbands. The room used in the experiment is an office room of size ($3 \times 4 \times 3$ m), with the microphone array placed in the center of the room. A number of scenarios have been investigated with different positions of the target source speaker relative to the microphone array, see Fig. 3. The simulations were made with speech sequences of both male and female speakers.

Fig. 4 shows short-time power estimates of a speech signal, a corresponding single unprocessed microphone signal, and the output from the continuously processed signals. A major reduction of ambient noise at low frequencies can be seen. The performance of the algorithm is given in Fig. 5 as a function of the norm of a uniform delay uncertainty vector $\mathbf{\Delta}$ for a fixed position of the target speaker. The performance evaluation includes source speech distortion and suppression of both background noise

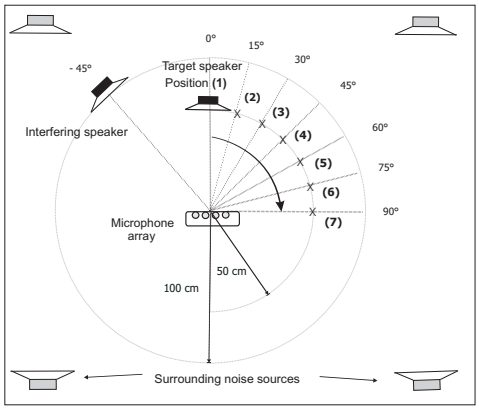

Figure 3: *Configuration of microphone array and sound sources in an office room. Source movement path represented by the numbered crosses, passed at a speed of one second per step.*

and loudspeaker interference for scenarios with different signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR). Up to 10 dB noise suppression and 22 dB interference suppression are obtained. It can be seen that the noise and interference suppression decrease for a relatively large estimated source spread (i.e., corresponding to a time-delay range above 1 ms). It is accompanied by an increase in speech distortion.

An error may be introduced by the tracking algorithm when evaluating the time-delay estimates for the speech source due to microphone mismatches and in low SNR environments. The algorithm performance is given in Fig. 6 as a function of the error in the TDOA estimation. The source covariance estimates are calculated for a uniform uncertainty vector norm $||\mathbf{\Delta}||$ of 750 μs. The proposed beamformer has proven to be robust to time-delay estimation errors. It presents a relatively small decrease in performance ($\sim$2 dB) with a time-delay error of 1 ms.

Another setup was defined for a moving target source along a circular path of radius 50 cm, centered at the reference point to describe a quarter of a circle. The numbered crosses depicted in Fig. 3 correspond to the speaker's positions where the update of the SRP-PHAT algorithm takes place. Results in Fig. 7 show good speech tractability with around 9 dB noise reduction while keeping an approximately constant level of speech distortion as for a non-moving source.

4. CONCLUSION

A new adaptive subband beamformer for moving source speech enhancement using time-delay estimation has been presented. The proposed beamformer is a recursive least squares algorithm using a soft constraint defined for a time-delay spread corresponding to an area expansion of the

speech source. A speech localization algorithm is combined with the beamformer to allow for speaker movement. The evaluation of the algorithm in an office room shows up to 10 dB noise reduction and 20 dB interference suppression. Furthermore, the algorithm provides good speaker movement tracking, while keeping an approximately constant level of speech distortion as for a non-moving source.

5. REFERENCES

[1] M. Brandstein, and D. Ward, "Microphone Arrays - Signal Processing Techniques and Applications," Springer, 2001.

[2] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech Enhancement Based on the Subspace Method," *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.

[3] S. Shahbazpanahi, A.B. Gershman, Z. Q. Luo, and K. M. Wong, "Robust Adaptive Beamforming for General-rank Signal Model," in *IEEE Trans. Sig. Proc.*, Vol. 51 , no. 9 , pp. 2257 2269, September 2003.

[4] Z. Yermeche, P. M. Garcia, N. Grbić, and I. Claesson, "A Calibrated Subband Beamforming Algorithm for Speech Enhancement," in *IEEE Sensor Array and Multichannel Sig. Proc. Workshop*, August 2002.

[5] J. P. LeBlanc, and P. L. De Lon, "Speech Separation by Kurtosis Maximization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1029–1032, May 1998.

[6] A. Johansson, N. Grbić, and S. Nordholm, "Speaker Localisation Using the Far-field SRP-PHAT in Conference Telephony," in *IEEE Int. Symposium on Intelligent Sig. Proc. and Comm. Sys.*, November 2002.

[7] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of Oversampled Uniform DFT Filter Banks with Delay Specifications using Quadratic Optimization," in *IEEE Int. Conf. Acoust. Speech and Sig. Proc.*, vol. VI, pp. 3633–3636, May 2001.

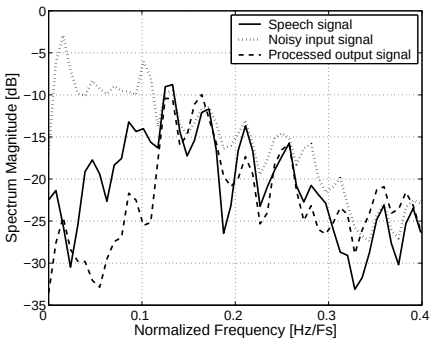

Figure 4: *short-time power estimates.*

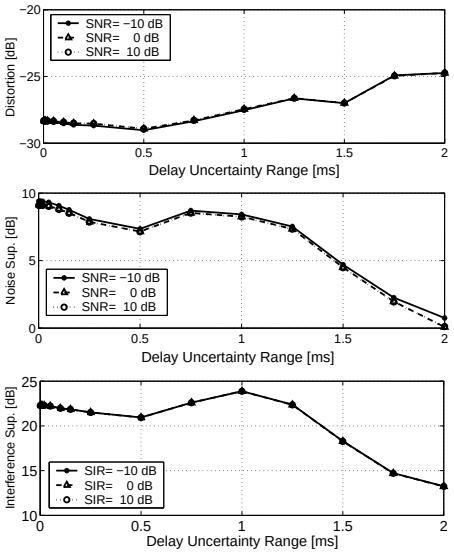

Figure 5: *Performance versus norm of delay-uncertainty vector* $||\Delta||$.

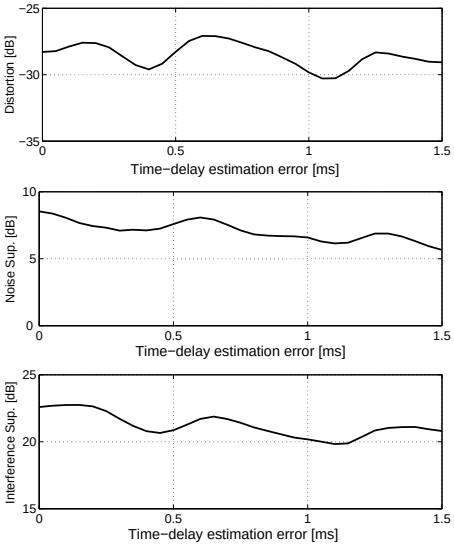

Figure 6: *Performance versus time-delay estimation error.*

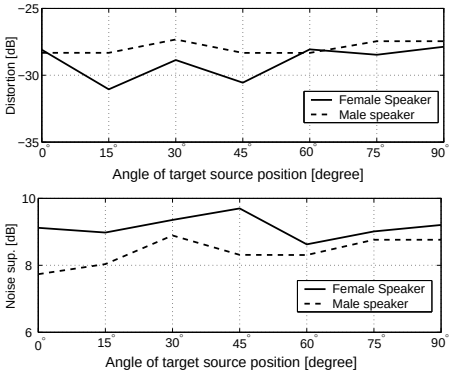

Figure 7: *Performance versus speaker angular position.*