SPEECH ENHANCEMENT OF MULTIPLE MOVING SOURCES BASED ON SUBBAND CLUSTERING TIME-DELAY ESTIMATION

Zohra Yermeche, Nedelko Grbić and Ingvar Claesson

zohra.yermeche@bth.se
Blekinge Institute of Technology, School of Engineering, SE-372 25 Ronneby, Sweden

ABSTRACT

A new robust blind microphone array method to enhance speech signals generated by multiple moving sources in a noisy environment is presented. This approach is based on a two-stage scheme. A subband clustering time-delay estimation algorithm is first used to localize the dominant speech sources. The speech enhancement is performed in a second stage, based on the acquired spatial information, by means of a soft-constrained subband beamformer. The robustness of this structure is ensured by the spatial constraint constructed to include the discrepancies in the acoustical environment model as well as errors in the time-delay estimation. Such scheme also allows for an efficient adaptation of the beamformer to speakers movement. The proposed subband clustering approach for time-delay estimation exploits the sparseness of speech signals in the time-frequency domain to localize multiple speakers simultaneously. It also provides means to select the number of target sources. Evaluation in a real environment with moving speakers shows promising results.

1. INTRODUCTION

Microphone arrays have been extensively exploited for the enhancement of speech in hands-free applications, such as conference telephony, speech recognition and hearing aid devices [1]. Most existing methods aim at enhancing a target signal by attenuating interferences whether due to reverberation, background noise or jammers. In many of the above applications, it is however of interest to be able to listen to several active speakers simultaneously while attenuating interfering sound sources and background noise. In other words, the underlying objective is to mimic the human brain capability to focus on a discussion of interest in a crowded and noisy environment. This is commonly referred to as the cocktail party problem. Considerable research is being carried out on this topic spanning a large range of microphone array techniques. Among them, blind signal separation based on independent component analysis (ICA) prominently stands out. However, ICA algorithms have limited performances in the case of more sources than sensors. Different postprocessing approaches were suggested to remove the remaining interferences in the ICA-separated signals. A Kurtosis maximization solution was successfully used for the extraction of a single speech source drawn in babble speech [2]. Time-frequency masking was also suggested as a postprocessing alternative for the extraction of multiple speech sources [3].

In [4] we proposed a structure based on subband soft constrained beamforming and time-delay estimation (TDE) for blindly enhancing a dominant speech source in a noisy environment. In this paper, the method in [4] is extended to the case of extracting multiple dominant sources from a mixture with lower-power speech interferences and background noise. The desired dominant speech sources are assumed to have about the same power at the sensors. Thus, the first issue is to estimate the time-delay for each of these dominant sources. It is known that simultaneously active speech sources result in small overlap in the time-frequency domain [3, 5]. Based on this observation, a new subband structure for TDE is proposed to localize multiple dominant sources in adverse noise situations. This approach opens the path for a simple way to specify the number of dominant speech sources. The time-delay estimates are used directly in the subband soft-constrained beamformer through source covariance estimates.

2. PROPOSED METHOD

Lets consider a sound field consisting of multiple speech sources s_j, $j = 1, ..., S$, as well as background noise, impinging on a uniform linear array of I sensors. The array input vector $\mathbf{x}(n) = [x_1(n), \dots, x_I(n)]^T$, at sample instant n, when all the sources are active simultaneously, is

$$\mathbf{x}(n) = \sum_{j=1}^{S} \mathbf{x}_{s_j}(n) + \mathbf{x}_n(n). \qquad (1)$$

Here $[.]^T$ represents the transpose operator. Vectors $\mathbf{x}_{s_j}(n)$, $j = 1, ..., S$, and $\mathbf{x}_n(n)$ are the $I \times 1$ received microphone input vectors generated by the S speech sources and the ambient noise, respectively. The goal is to obtain an output signal $y(n)$ which corresponds to the direct-path com-

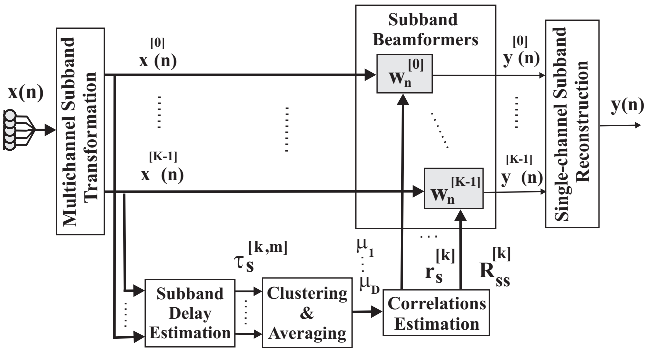

Figure 1: *Structure of proposed method.*

ponents of the D dominant sources ($D \leq S$) observed at the array. The structure of the proposed speech enhancement method is shown in figure 1.

2.1. Subband Decomposition

First, a subband decomposition of each microphone input signal $x_i(n)$, $i = 1, \ldots, I$, is achieved by means of a multichannel modulated uniform analysis filter bank [6]. This results in a set of K narrow band signals defined as

$$x_i^{[k]}(n) = \sum_{l=0}^{L-1} h(l)\, e^{j2\pi \frac{k}{K} l}\, x_i(n-l)\,, \qquad (2)$$

where $k = 0, \ldots, K{-}1$, is the subband index and $h(l)$, $l = 0, \ldots, L{-}1$, are coefficients of a low-pass prototype filter. The spatial characteristics of the input signal are maintained by using the same filter bank at each microphone. The subband decomposition allows the filtering operations to be performed in the frequency domain. The subband outputs of the subsequent spatial filters, $y^{[k]}(n)$, are reconstructed by a modulated uniform synthesis filter bank, in order to create a time-domain output signal, by

$$y(n) = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} h(l)\, e^{-j2\pi \frac{k}{K} l}\, y^{[k]}(n-l)\,. \qquad (3)$$

A resulting computational gain of the subband processing comes from the fact that the filtering of narrow band signals requires lower sample rates. Hence, in an efficient implementation, the frequency transform is followed by a decimation operation [6]. A simplified structure is made available through the use of a polyphase implementation. Another major outcome of subband decomposition is that the overlap between received speech signals decreases significantly in the time-frequency domain [3, 5].

2.2. Multiple Source Time-Delay Estimation

With a far-field model, the wave forms are viewed to impinge on the array as plane waves. The direction of arrival (DOA) of each source signal to a linear array is thus mapped to a time-difference of arrival (TDOA) for two adjacent microphones. In the following sections, the TDOA of the desired dominant speech sources is extracted from the received noisy signals and their number is determined. To make use of the sparseness of speech signals in the time-frequency domain, the TDE is performed in each subband individually. It is based on maximizing cross-correlations between the received subband short-time signals. Due to the relative high power of background noise at low frequencies, in comparison to speech [4], only K_h higher subbands ($K_h < K$) are used in the TDE. The TDE for the dominant speech sources is then accomplished by clustering the estimated TDOA for all K_h subbands and averaging the delays for each cluster. The averaging is additionally performed over M time slots of N samples to reduce the estimation errors.

2.2.1. *Subband Time-Delay Estimation*

A modified steered response power with phase transform (SRP-PHAT) algorithm for a far-field scenario [1] is used on the received subband short-time signals. The TDOA estimate, $\tau_s^{[k,m]}$, in each subband k and for time slot m is

$$\tau_s^{[k,m]} = \arg\max_{\tau} \sum_{p=1}^{I} \sum_{q=1}^{I} G_{pq}^{[k,m]}\, e^{j2\pi f_k \tau (p-q)}\,, \qquad (4)$$

where

$$G_{pq}^{[k,m]} = \frac{E\{x_p^{[k]}(n)\, x_q^{[k]}(n)^*\}}{|E\{x_p^{[k]}(n)\, x_q^{[k]}(n)^*\}|} \qquad (5)$$

is the normalized cross power spectrum of the signals received by the microphone pair p and q for subband k and time slot m. Here f_k is the central frequency of subband k, the symbol $E\{.\}$ denotes the statistical expectation and $(^*)$ is the conjugate operator.

2.2.2. *Clustering and Averaging*

The purpose of the time-delay clustering and averaging operations is twofold. First, it is used to set the number of target speech sources to enhance, and second to determine the TDOA of the direct path for each target source.

A K-means clustering algorithm [7] is used for partitioning the $K_h \times M$ gathered TDOA data $\tau_s^{[k,m]}$ into J disjoint subsets S_j containing N_j data elements. The criterion is to minimize the sum of cost functions $c(S_j)$, $j = 1, \ldots, J$, defined for each cluster S_j with element centroid μ_j as

$$c(S_j) = \sum_{\tau_s \in S_j} \mid \tau_s - \mu_j \mid\,, \quad \mu_j = \frac{\sum_{\tau_s \in S_j} \tau_s}{N_j}\,. \qquad (6)$$

Note, the index $[k,m]$ of the delay estimates $\tau_s^{[k,m]}$ has been dropped for convenience. This partitioning algorithm results in the minimization of the distance between elements in a cluster while maximizing the distance between elements in different clusters [7]. The TDOA estimates of the D target source signals are chosen as the cluster centroids μ_j corresponding to the clusters with a minimum predefined number of elements N_{th} and a variance below a specified threshold [3], according to

$$N_j > N_{th}\,, \qquad \frac{c(S_j)}{N_j} < var_{th}\,. \tag{7}$$

Without loss of generality the centroids corresponding to the desired time-delays are indexed as $\mu_1, \ldots, \mu_D$, and the related D dominant sources as $s_1, \ldots, s_D$.

2.3. Spatial Filtering

The time-delays estimated in section 2.2, for the D dominant speech sources are used directly in a subband soft-constrained beamformer through source covariance estimates [4].

2.3.1. Constrained Beamformer

The output of the beamformer, for subband k is given by

$$y^{[k]}(n) = \mathbf{w}_n^{[k]^H}\,\mathbf{x}^{[k]}(n)\,, \tag{8}$$

where $\mathbf{w}_n^{[k]}$ is the beamformer weight vector and where the symbol $(.)^H$ stands for the Hermitian transpose. The proposed beamformer is deduced from a recursive least squares formulation of the Wiener solution, with the array weight vector given by [4]

$$\mathbf{w}_n^{[k]} = \left[\hat{\mathbf{R}}_{\mathbf{x}}^{[k]}(n) + \mathbf{R}_s^{[k]}\right]^{-1}\mathbf{r}_s^{[k]}\,. \tag{9}$$

Here, $\hat{\mathbf{R}}_{\mathbf{x}}^{[k]}(n)$ is the received signal covariance matrix estimate, continuously calculated from observed data by

$$\hat{\mathbf{R}}_{\mathbf{x}}^{[k]}(n) = \sum_{p=0}^{n} \lambda^{n-p+1}\,\mathbf{x}^{[k]}(p)\;\mathbf{x}^{[k]^H}(p)\,, \tag{10}$$

where λ is a forgetting factor with the purpose of tracking variations in the surrounding noise environment. The matrix $\mathbf{R}_s^{[k]}$ and vector $\mathbf{r}_s^{[k]}$ are the spatial source covariance matrix and the spatial cross covariance vector, defined as

$$\begin{aligned}\mathbf{R}_s^{[k]} &= E\left\{\mathbf{x}_s^{[k]}(n)\,\mathbf{x}_s^{[k]^H}(n)\right\}\,,\\ \mathbf{r}_s^{[k]} &= E\left\{\mathbf{x}_s^{[k]}(n)\,\sum_{d=1}^{D} s_d^{[k]}(n)\right\}\,.\end{aligned} \tag{11}$$

The vector $\mathbf{x}_s^{[k]}(n) = [x_{s,1}^{[k]}(n), \ldots, x_{s,I}^{[k]}(n)]^T$ is the desired signal vector corresponding to the sum of the direct-path signals for the D dominant sources. It is defined as

$$x_{s,i}^{[k]}(n) = \sum_{d=1}^{D} a_{d,i}\,e^{-j2\pi f_k \tau_{d,i}}\,s_d{}^{[k]}(n), \tag{12}$$

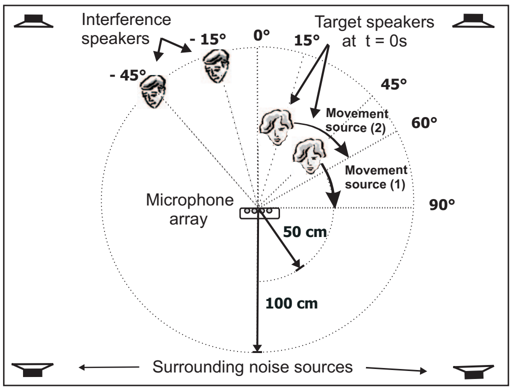

Figure 2: *Configuration of microphone array and sources.*

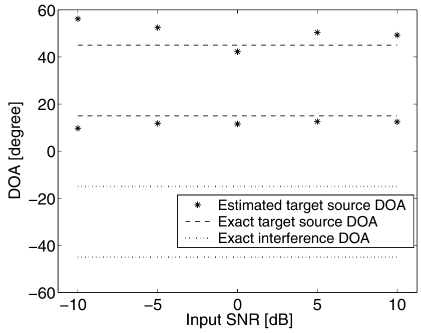

Figure 3: *Estimated DOA vs. SNR for scenario at* $t = 0s$.

where $a_{d,i}$ and $\tau_{d,i}$ are the signal power attenuation factor and the time-delay of the direct path from s_d to the i^{th} sensor [4]. The source covariance matrix $\mathbf{R}_s^{[k]}$ in (9) constitutes a soft-constraint, which acts as a spatial passband and reduces the weights fluctuations generated by the speech pauses. additionally, it forces the full rank of the total matrix to be inverted when no noise or speech is present (i.e., when $\hat{\mathbf{R}}_{\mathbf{x}}^{[k]}(n)$ is ill conditioned).

2.3.2. Spatial Source Correlation Estimation

Information about the target speech sources is used in the algorithm by calculating TDOA-based source covariance estimates, independent of the received signals power spectrum, as

$$\begin{aligned}\tilde{\mathbf{R}}_s^{[k]} &= \sum_{d=1}^{D}\int_{\Phi_{d,k}} e^{-j2\pi f_k \tau\,\mathbf{v}}\;e^{j2\pi f_k \tau\,\mathbf{v}^T}\,d\tau\,,\\ \tilde{\mathbf{r}}_s^{[k]} &= \sum_{d=1}^{D}\int_{\Phi_{d,k}} e^{-j2\pi f_k \tau\,\mathbf{v}}\,d\tau\,.\end{aligned} \tag{13}$$

Parameter $\Phi_{d,k}$ is the time-delay range in subband k, corresponding to an area expansion of s_d, and is defined as

$$\Phi_{d,k} = \left[\mu_d - \frac{\Delta_k}{2}, \mu_d + \frac{\Delta_k}{2}\right], \quad \Delta_k = \varphi\, f_k^{\frac{1}{4}}, \tag{14}$$

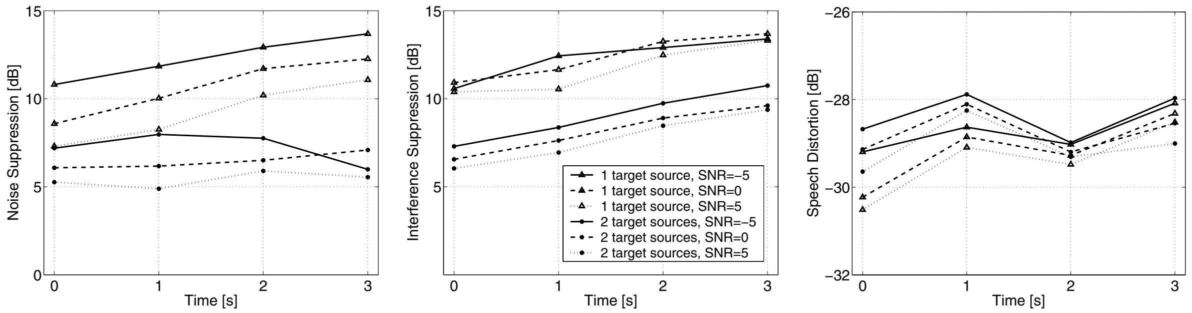

Figure 4: *Performance of proposed method with movement of target sources (velocity $15^o/s$ at 50 cm radius, TDE update every $1s$). Two settings: single dominant source (1), two dominant sources (1) and (2). The two male interfering sources are active at all time. The signal to interference ratio SIR=3 dB.*

where φ is a constant. The vector $\mathbf{v} = [0, \ldots, I-1]^T$. A frequency dependent delay-uncertainty, Δ_k, is chosen to adjust the beamformer's opening for each subband in order to reduce the sources' speech distortion. The rationale behind this definition is that the low-frequency bands contain essentially noise power, thus, a small constraint area allows for more noise suppression in these subbands. On the other hand, a bigger constraint area in the upper bands secures the passage of the desired speech signals relatively undistorted.

3. EXPERIMENTS

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 5 kHz. All simulations were performed with 64 subbands. The room used in the experiments is an office room of size (3 × 4 × 3 m), with the microphone array placed in the center of the room (see figure 2). The background noise is generated by four loudspeakers located in the room corners and emitting pink noise. Simulations also include influence of interference from two male speakers, as well as room reverberation resulting from the two target female speakers. The DOA estimates obtained by the subband SRP-PHAT algorithm of section 2.2 for different input signal to noise ratio (SNR) is given in figure 3. The scenario used was at $t\ =0$ s (see figure 2). The delay-uncertainty constant was experimentally set to $\varphi = 4.5 \times 10^{-6}$. It can be seen that a good DOA estimation is achieved using the proposed method. Simulations were performed for moving target speakers relative to the microphone-array, as depicted in figure 2. Performance measures of the proposed speech enhancement algorithm are given in figure 4 for different SNR values. Results are plotted for settings with a single target source as well as with two target sources. A considerable level of noise and interference suppression (around 8-10 dB) was obtained with a good preservation of speech integrity. The speech distortion is relatively invariant to speaker movement while the achieved suppression increases with the distance between target sources and interferences.

4. CONCLUSION

A new robust speech enhancement method for multiple moving sources in a noisy environment has been presented. The structure consists of a subband-based clustering time-delay estimation algorithm followed by a subband soft-constrained beamformer. This approach allows for the detection and enhancement of multiple dominant speakers in a mixture of interferences. Additionally, The use of a frequency dependent constraint region opens the path for a trade off between noise suppression and speech intelligibility.

5. REFERENCES

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

[2] S. Y. Low, R. Togneri, and S. Nordholm, “Spatio-Temporal Processing for Distant Speech Recognition,” in *IEEE Int. Conf. Acoust. Speech and Sig. Proc.*, vol. I, pp. 1001–1004, May 2004.

[3] H. Sawada, R. Mukai, S. Araki and S. Makino, “Real-Time Blind Extraction of Dominant Target Sources from Many Background Interference Sources,” in *Int. Worshop on Acoust. Echo and Noise Control*, September 2005.

[4] Z. Yermeche, N. Grbić, and I. Claesson, “Moving Source Speech Enhancement Using Time-Delay estimation,” in *Int. Worshop on Acoust. Echo and Noise Control*, September 2005.

[5] Ö. Yilmaz, and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” in *IEEE Trans. Sig. Proc.*, Vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] P. P. Vaidyanathan, “Multirate Systems and Filter Banks,” Prentice-Hall, 1993.

[7] C. M. Bishop, “Neural Networks for Pattern Recognition,” Oxford University Press, 1995.