Speech Enhancement and Source Separation based on Binaural Negative Beamforming

A. Álvarez, P. Gómez, R. Martínez, V. Nieto and V. Rodellar
Universidad Politécnica de Madrid – Facultad de Informática
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, SPAIN
Tel.:+34.91.336.73.84, Fax: +34.91.336.66.01, e-mail: pedro@pino.datsi.fi.upm.es

Abstract

Negative Beamformers are well known for their high angular selectivity, which makes them potentially suitable for speech enhancement applications in noisy backgrounds and for directional source separation. On the other hand, Spectral Subtraction is a well-known method for removing noise from a noise-corrupted speech signal. The scheme that is proposed in this paper combines both techniques in order to obtain large gains in the SNR at a reasonable low computational cost. This method may be used to eliminate or enhance a specific signal using a binaural array. The fundamentals of the technique are reviewed, and a structure to control and improve its angular selectivity is presented. Results obtained in a real situation are also commented. Applications of this technique may be found in Security Systems, Domotic Control and also, to improve Speech Recognition.

1. Introduction

Speech Processing and Recognition are important technologies to produce Intelligent General User Interfaces. In many situations, multiple speakers or speech/sound sources may be active at a time within the same environment, as implied in Figure 1.

Using that configuration, speakers need not to use or wear proximity microphones. In addition, it would be possible that several speakers would be active at a time without producing cross-talk if different conversations were to be monitored or used as inputs for Speech Recorders or Recognizers [8]. The key to allow this flexible use of speech in cocktail party scenarios is Array Beamforming [4].

Through this paper the use of Negative Beamformers as the one in Figure 2 is proposed, based on their rather narrow negative beams and lower number of array elements required, as compared with classical beamformers [6]. The output signal produced by the Negative Beamformer is then used to feed a subsequent Spectral Subtractor module, in order to restore the signal previously tracked and cancelled.

2. Negative Beamforming

The Negative Beamformer (NBF) introduced in Figure 2 shows a Transfer Function in the Frequency Domain, which may be formulated as [5]:

\[ Y(\alpha, \delta) = 2e^{-j(\delta-\delta)^2} \left( (1 - 2\beta) \cos \alpha \sin \delta / 2 - \sin \alpha \cos \delta / 2 \right) \]  

where:

\[ \alpha = \omega \xi = 2 \pi f \]  

\[ \delta = \omega T = 2 \pi k \tau = 2 \pi k f_s \]  

\[ \phi \] being the angle of arrival, \( \xi \) half the array travel time, \( f \) the frequency of the signal, \( k \) the delay order, \( d=2D \) the microphone distance and \( f_s \) the sampling frequency.

The module of (1) when plotted against \( \phi \) and \( \delta \) may be seen in Figure 3 and Figure 4.
The adaptation parameter $\beta$ may be automatically adjusted to focus on different incoming sources. This aspect is especially advantageous when the position of one or several sources is known, as in the situation contained in the Figure 1. The relation between the adaptation parameter $\beta$ and the angle of arrival $\phi$ is given by the following expression:

$$
\beta = \frac{1}{2} \left( 1 - \frac{\tan \left( \frac{2\pi D}{c} \sin \phi \right)}{\tan(\delta/2)} \right)
$$

(4)

As it may be noticed, the value of $\beta$ which produces highest degree of cancellation depends not only of the value of $\phi$, but of the signal frequency $f$, as well (Figure 5). This property of the *Negative Beamformer* implies that for broad-band signals (e.g. speech), the spectra of interest should be divided into different bands, throughout the use of bandpass filters (see Table 1), and replicating the beamformer cell, previously presented.

The structure implementing this strategy is shown in Figure 6. The signal from each microphone is first band-limited using one of the 20 bandpass filters, previously mentioned. Each two corresponding channels are then fed to the bank of *Negative Beamformers* and a *Linear Combiner* produces the total output signal, whether canceling or enhancing a given source, as desired.

![Figure 3. Module of the Negative Beamformer Transfer Function for d=5 cm, f=11,025 Hz, $\beta$=0.50 and k=1.](image)

![Figure 4. Module of the Negative Beamformer Transfer Function for d=5 cm, f=11,025 Hz, $\beta$=0.25 and k=1.](image)

![Figure 5. Dependencies among angle of arrival for a sound source and, their associated $\beta_j$ values for the 20 frequency-bands, as defined in Table 1.](image)

![Figure 6. Algorithmic structure to track a specific sound source.](image)

Table 1. Bank of filters applied to the input signals $x_a$ and $x_b$ (see Figure 6). The 20 filter transfer functions are linearly separated below 1000 Hz and exponentially separated above that frequency.

<table>
<thead>
<tr>
<th>Band #</th>
<th>Frequency Range</th>
<th>Band #</th>
<th>Frequency Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100 - 200 Hz</td>
<td>11</td>
<td>1123 - 1284 Hz</td>
</tr>
<tr>
<td>2</td>
<td>200 - 300 Hz</td>
<td>12</td>
<td>1284 - 1470 Hz</td>
</tr>
<tr>
<td>3</td>
<td>300 - 400 Hz</td>
<td>13</td>
<td>1470 - 1699 Hz</td>
</tr>
<tr>
<td>4</td>
<td>400 - 500 Hz</td>
<td>14</td>
<td>1699 - 1980 Hz</td>
</tr>
<tr>
<td>5</td>
<td>500 - 600 Hz</td>
<td>15</td>
<td>1980 - 2326 Hz</td>
</tr>
<tr>
<td>6</td>
<td>600 - 700 Hz</td>
<td>16</td>
<td>2326 - 2752 Hz</td>
</tr>
<tr>
<td>7</td>
<td>700 - 800 Hz</td>
<td>17</td>
<td>2752 - 3276 Hz</td>
</tr>
<tr>
<td>8</td>
<td>800 - 900 Hz</td>
<td>18</td>
<td>3276 - 3921 Hz</td>
</tr>
<tr>
<td>9</td>
<td>900 - 1000 Hz</td>
<td>19</td>
<td>3921 - 4715 Hz</td>
</tr>
<tr>
<td>10</td>
<td>1000 - 1123 Hz</td>
<td>20</td>
<td>4715 - 5512 Hz</td>
</tr>
</tbody>
</table>

One important problem in using *Negative Beamformers* is the Individual Source Tracking, which consists in finding the correct angle of arrival for a source present in the input among some others using estimates of the power of the beamformer output signal $y(t)$ [1][2]. The solution proposed in this system, is to track the presence of the different sources in the frequency domain [1][7]. A measure called *Groove Aspect Factor* is introduced to determine if there are one or more sources currently active at the corresponding band [6]. The criterion is based in the formula:
\[ c_j = \frac{2 |P_j| Q_{20}^2}{|P_j| (0.0)^j + |P_j| (0.0)^0} \]  
where \(|Y_j(\beta)|\) is the module of the \textit{Negative Beamformer Transfer Function} as given by (1), \(1 \leq j \leq 20\) being the index of the corresponding filter bank, and:

\[ \beta_{\text{in}} = \arg \min_j \{|Y_j(\beta)|\} \tag{6} \]

Once individual sources have been located for a single band, a \textit{band cross-mapping} will infer their presence in other bands, and help in estimating their angle of arrival and \textit{particular steering factor} \(\beta\).

### 3. Spectral Subtraction

To implement the filtering in the spectral domain [3] (Figure 7), the \textit{Negative Beamformer output} \(z_{\text{m}}\) is used as the primary signal, and one of the Microphone Inputs \(x_{\text{ab}}\) (or even another \textit{Negative Beamformer output}) is used as the reference one.

\[
\begin{align*} 
Z &= Z(m) = |F \{z(n)w(n)\}| \tag{7} 
X &= X(m) = |F \{x(n)w(n)\}| \tag{8} 
\end{align*}
\]
where \(M\) is the size of the window used, \(w(n)\) is the window function, and \(n\) and \(m\) are the time and frequency indices.

In a first step the relationship between the power spectra of the \textit{Negative Beamformer output} and the reference signal is calculated for every frequency channel:

\[
\gamma_m(m) = \left| \frac{X_m(m)}{Z_m(m)} \right|^2, \quad 0 \leq m \leq M/2 - 1 
\tag{9}
\]

Now, the ratio is weighted using a logarithmic law before subtracting:

\[
\gamma'_m(m) = \log(\gamma_m(m)), \quad 0 \leq m \leq M/2 - 1 
\tag{10}
\]

\[
|S_m(m)| = |X_m(m)|^2 - \gamma'_m(m), \quad 0 \leq m \leq M/2 - 1 
\tag{11}
\]

This subtraction is combined with a spectral flooring technique to limit the presence of artificial tones. Besides, the phase of the enhanced signal is recovered from the reference trace \(x_{\text{ab}}\):

\[
\varphi_{\text{aa}}(m) = \varphi_{\text{ab}}(m), \quad 0 \leq m \leq M/2 - 1 
\tag{12}
\]

### 4. Results

In a practical experiment shown in Figure 8 through Figure 12 a pair of simultaneous speech traces are produced using a couple of loudspeakers. The dispositions of the speech sources and the pair of microphones are shown in Figure 8.

![Figure 8. Framework for the recording of a two-source signal. Two signal sources (loudspeakers) \(s_1(t)\) and \(s_2(t)\) are placed on the same plane relative to the array microphone (m1 and m2 separated d=2D).](image)

The left source \(s_1(t)\) corresponds to the signal represented in Figure 9 (French sentence /un loup s’est jeté immédiatement sur la petite chèvre/ produced by a male speaker). The right source \(s_2(t)\) corresponds to a double production of the French sentence /ce petit canard append à nager/ uttered by the same speaker (see Figure 10).

![Figure 9. Power spectrum of an utterance of the French sentence /un loup s’est jeté immédiatement sur la petite chèvre/ produced by a male speaker.](image)

![Figure 10. Power spectrum of an double utterance of the French sentence /ce petit canard append à nager/ produced by the same male speaker.](image)
The power of signal $s_1(t)$ is similar to the power of $s_2(t)$. Figure 11 contains the power spectra of the combined signal captured by microphones $m_1$ and $m_2$ separated 5 cm. The separation between $s_1(t)$ and $m_1$ is 200 cm, and the distance between $s_2(t)$ and $m_2$ is also 200 cm. Finally, the arriving angles $\phi_1$ and $\phi_2$ are equal to 22.5°.

The low number of sensors required for the system operation (only two microphones), is of most importance since system performance is similar to larger microphone arrays, but at a much lower cost.

This technique may be extended to follow different sources once the best steering factor is determined for every signal. That fact is consequence of source-tracking processes, which are run in parallel with cancellation related tasks.

Important application fields of the system may be found in clean speech monitoring, and noise removal for Robust Speech Recognition Systems.

5. Conclusions

The combination of the herein-proposed methods (Negative Beamformer Filtering and Frequency-Domain Spectral Subtraction) results in a high degree of source enhancement or cancellation at a reasonable computational cost. As these structures are rather selective in the angular domain, signal enhancement factors up to 30 dB may be achieved.

The Spectral Subtraction process is then applied to restore only the portion of the signal corresponding to $s_1(t)$ (Figure 12.a) or the signal $s_2(t)$ (Figure 12.b).

6. Acknowledgments

This research is being supported by Project TIC99-0960 (Programa Nacional de las Tecnologías de la Información y las Comunicaciones), Project 07T/0001/2000 (Plan Regional de Investigación de la Comunidad Autónoma de Madrid), and a Contract between the Centre Suisse d'Electronique et de Microtechnique and Universidad Politécnica de Madrid.

7. References


