TIME-FREQUENCY TRANSFORMS AND BEAMFORMING FOR SPEAKER RECOGNITION

Antonio Satué-Villar1, Juan Fernández-Rubio2
1Escuela Universitaria Politécnica de Mataró (EUPMT)
2Escuela Técnica Superior de Ingenieros de Telecomunicación de Barcelona (ETSETB)
Universidad Politécnica de Cataluña (UPC), Barcelona

e-mail: 1satue@eupmt.es , 2juan@gps.tsc.upc.es

ABSTRACT

In this paper we study the advantages of a data transformation using time-frequency transforms (wavelets, in our case) in a wideband beamforming system. For this purpose, we simulate a target signal and interfering jammers that are impinging upon a microphone array in a noisy environment. The target and interfering signals are obtained from a speech database. The simulations presented show that the wavelet transform reduces the time of convergence, even though there is not an unique optimal wavelet filter; it depends on the speech signal. In this paper it can be also inferred that the wavelet transform do not affect substantially to the quality of the output signal (evaluated with a speaker recognition method).

1 INTRODUCTION

In order to acquire speech signals with a microphone in a noisy and/or interfering signals environment, the reception beampattern of the microphone should be appropriated; it should have a maximum in the direction of the incoming speech signal and minimize all the interferences. To solve this problem, several microphones can be grouped and their beampatterns combined with some specific weights (w) to obtain the desired characteristics.

If the impinging signals are stationary, the weights could be fixed. As this is not the situation in a more realistic situation, they must be adapted based on incoming signal variations; adaptive algorithms are one of the starting points for this work.

All adaptive algorithms can be considered as approximation of the optimum Wiener solution. It is well known that the solution strongly depends on the dispersion of the eigenvalues of the input data. As a way to obtain better results incoming signal should be uncorrelated as much as possible. The more decorrelated the incoming signal are, better results are obtained. Wavelet transform does not only decorrelate input data but it defines non-constant widebands (in FFT the bands are of the same wide), what seems more convenient for speech signals (it can work with octave bands). Thus, the beamforming problem is transformed in a set of less complex problems using narrow band signals. With such decomposition, different adaptation coefficients can be used.

The paper is organized as follows. In the following section the GSC structure is reviewed. In section 3 the improvement of GSC results using wavelet transform is justified. In section 4, some experiments based on computer simulations are performed. The paper concludes in section 5 with a final result discussion.

2 GSC STRUCTURE

This section starts presenting the GSC structure, proposed by Griffiths and Jim [1]. This structure has an efficient hardware implementation.

As it can be seen in figure 1, there are two branches: - An upper branch, named quiescent, with weight vector \( w_q \) which makes use of the constraints.
- A lower branch, that consists of a blocking matrix (\( B \)) which blocks the target signal, followed by an unconstrained beamforming (\( w_a \) weights).

The lower branch only processes the interfering signals, which are subtracted to the upper branch. Thus, the output signal is free of interferences. It is very important to remark that the look-direction of the desired signal and microphone beampattern have to be known. On the contrary, the lower branch can not block completely the desired signal. The desired signal is then processed by the constrained beamforming and subtracted from the upper branch. Thus, the very well known signal cancellation phenomenon occurs.

Chen and Fang [2] applied the LMS algorithm in the frequency domain to the Griffiths-Jim GSC beamformer,
and showed that the mean square error (MSE) of both structures is the same, but the frequency domain provides faster convergence.

Figure 2. Chen-Fang beamformer

In this paper a GSC structure is outlined, applying the DWT of the data (wavelet domain). So, a wavelet transform block instead of FFT is inserted between the matrix B and the weights \( w_a \).

3 WAVELET TRANSFORM

Wavelet transform represents an improvement of Fourier transform, as it adds one more dimension. The Fourier transform of a one-dimensional signal provides a physical meaning: espectral content. The wavelet transform of a one-dimensional signal yields a two-dimensional function where, besides of spectral content, it gives temporal information. For more information about wavelet transform, see references [3] and [4].

There are different families of wavelet transform, each one with its properties. The more usual are Daubechies and biorthogonal families. These wavelets allow perfect reconstruction of the signal. There are different wavelets in every family, depending on the filters of the transformation. For example, db2 and db4 are two different wavelets of Daubechies family.

A wavelet is characterized by two filters (low pass filter LPF and high pass filter HPF). The wavelet coefficients are the even outputs of these filters. In the figure 3 we have the example for a wavelet with a 4-coefficient LPF and HPF, for a 8-bit input. Thus, 5 LPF coefficients and 5 HPF coefficients are obtained.

Figure 3. Obtention of DWT coefficients (example)

Figure 3 shows an one-level transform. Two-level transforms can be done transforming separately the low and high pass coefficients and so on. In figure 4 it can be seen the coefficients resulting from a 2-level transform that only iterates the LPF coefficients.

Figure 4. Wavelet decomposition (2 levels)

With classic methods the weights are adapted with a LMS algorithm with an unique \( \mu \).

\[
\begin{align*}
  w_j(n+1) &= w_j(n) + \mu \cdot y^*(n)x_j(n) \\
  \mu &= \frac{\alpha}{c + x_j^*x_j} 
\end{align*}
\]  

(1)

(2)

where \( \alpha \) and \( c \) determine the speed of the adaptation.

With the structure of figure 5, every kind of weights \((w1, w2, w3)\) with 3 different values of \( \mu \) can be adapted, accelerating the convergence.

\[
\begin{align*}
  w_j(n+1) &= w_j(n) + \mu_i \cdot y^*(n)x_j(n) \\
  \mu_i &= \frac{\alpha}{c + x_i^*x_i} 
\end{align*}
\]  

(3)

(4)

Figure 5. Weighting of DWT coefficients

4 EXPERIMENTS

In [5] a monosensor structure was presented which adapts some weights making use of a wavelet transform, but with a synthetic input signal. In this paper this structure is applied to speech signals (figure 6).
In figure 7 the MSE with DWT and without DWT are represented (nt=18 in both cases). It can be observed that in the first case the convergence is faster than in the latter, due to a more effective processing.

The DWT monosensor structure is extended to the GSC structure, as in figure 8. The data are decorrelated before the weights, in order to increase the convergence speed.

The scenario in the simulations is:
- ne=15 microphones equi-spaced half wavelength (central frequency=2 KHz).
- nt=8 taps by filter
- gaussian noise of mean 0 and variance 0.1
- incident signals are speech signals composed of 45000 samples (sampling frequency=8 KHz), incoming at 0º (target) and -30º (interfering) both with a mean power of 10000 (40 db)

It can be seen that using DWT, the speed of convergence is faster. It is interesting to see that the use of DWT does not degrade the quality of output signal. Therefore, a simple method of speaker recognition is used. The arithmetic-harmonic sphericity distances [6] are evaluated. It implies the computation of a covariance matrix for each speaker (target and interferer in our case), and the following measure distance:

$$\mu(C_j, C_{new}) = \log |tr(C_{new} C_j^{-1})| - tr(C_j^{-1})| - 2\log(m)$$  \hspace{1cm} (5)$$

where $C_{new}$ is the covariance matrix of the output of the array, $C_j$ is the covariance matrix of the user to be tested and $m$ is the dimension of the matrices (20 in all the experiments).

$$\text{Merit Factor} = (\mu_{15} - \mu_{20}) - (\mu_{15} - \mu_{20})$$  \hspace{1cm} (6)$$

where
- $\mu_{15} = \mu (C_j, \text{target, input array})$
- $\mu_{20} = \mu (C_j, \text{interferent, input array})$
- $\mu_{15} = \mu (C_j, \text{target, output array})$
- $\mu_{20} = \mu (C_j, \text{interferent, output array})$

A Merit Factor equal to 0 means that the input is equal to the output (the array does not operate properly).

The wavelets used are: db4, bior3.5 and bior2.2. In the table 1 are shown the coefficients for the LPF and HPF filters of these wavelets.

<table>
<thead>
<tr>
<th>Wavelet</th>
<th>LPF</th>
<th>HPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>db4</td>
<td>{-0.01, 0.03, 0.03, -0.18, -0.02, 0.63, 0.71, 0.23}</td>
<td>{-0.23, 0.71, -0.63, -0.02, 0.18, 0.03, -0.03, -0.01}</td>
</tr>
<tr>
<td>bior3.5</td>
<td>{-0.01, 0.04, 0.05 -0.26, -0.07, 0.96, -0.07, -0.26, 0.05, 0.04, -0.01}</td>
<td>{-0.17, 0.53, -0.53, 0.17}</td>
</tr>
<tr>
<td>bior2.2</td>
<td>{-0.17, 0.35, 1.06, 0.35, -0.17}</td>
<td>{0.35, -0.70, 0.35}</td>
</tr>
</tbody>
</table>

Table 1. Coefficients of the filters
Only 8 coefficients are taken to maintain equal the number of weights that have to be adapted using or not the DWT.

Table 2 shows the “Merit Factor” for two different couple of targets and interfering speakers (experiments A,B). References [7] and [8] shows that adaptation during the silences it’s not good. Simulations adapting the weights only when the signal is over a threshold are also performed. The results are:

<table>
<thead>
<tr>
<th></th>
<th>experiment A</th>
<th>experiment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptation during the silences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>db4</td>
<td>0.259</td>
<td>0.0988</td>
</tr>
<tr>
<td>bior3.5</td>
<td>0.258</td>
<td>0.1008</td>
</tr>
<tr>
<td>bior2.2</td>
<td>0.2603</td>
<td>0.1</td>
</tr>
<tr>
<td>With FFT</td>
<td>0.2612</td>
<td>0.1</td>
</tr>
<tr>
<td>Without transf.</td>
<td>0.2638</td>
<td>0.1</td>
</tr>
<tr>
<td>Not adaptation during the silences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>db4</td>
<td>0.2559</td>
<td>0.1036</td>
</tr>
<tr>
<td>bior3.5</td>
<td>0.2586</td>
<td>0.1043</td>
</tr>
<tr>
<td>bior2.2</td>
<td>0.2552</td>
<td>0.1041</td>
</tr>
<tr>
<td>With FFT</td>
<td>0.2614</td>
<td>0.104</td>
</tr>
<tr>
<td>Without transf.</td>
<td>0.2593</td>
<td>0.1052</td>
</tr>
</tbody>
</table>

Table 2. Merit Factor

The advantages of wavelet transform is the possibility of reducing the complexity of the structure. So, taking only 6 coefficients (as LPF as possible and some HPF, as shows figure 9), the results are practically the same as shown in table 3.

<table>
<thead>
<tr>
<th></th>
<th>experiment A</th>
<th>experiment B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptation during the silences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>db4</td>
<td>0.2568</td>
<td>0.1048</td>
</tr>
<tr>
<td>bior3.5</td>
<td>0.2551</td>
<td>0.1027</td>
</tr>
<tr>
<td>bior2.2</td>
<td>0.2584</td>
<td>0.1008</td>
</tr>
<tr>
<td>With FFT</td>
<td>0.2612</td>
<td>0.1</td>
</tr>
<tr>
<td>Without transf.</td>
<td>0.2622</td>
<td>0.106</td>
</tr>
<tr>
<td>Not adaptation during the silences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>db4</td>
<td>0.2608</td>
<td>0.1059</td>
</tr>
<tr>
<td>bior3.5</td>
<td>0.2596</td>
<td>0.1045</td>
</tr>
<tr>
<td>bior2.2</td>
<td>0.2585</td>
<td>0.1053</td>
</tr>
<tr>
<td>With FFT</td>
<td>0.2614</td>
<td>0.104</td>
</tr>
<tr>
<td>Without transf.</td>
<td>0.254</td>
<td>0.1049</td>
</tr>
</tbody>
</table>

Table 3. Merit Factor truncating the coefficients

5 CONCLUSIONS

Wavelet transform is an interesting alternative to frequency domain and temporary domain GSC beamforming. The complexity of GSC structure can be reduced by decreasing the number of DWT coefficients with not significant performance degradation. Moreover, DWT outperforms the conventional methods so it has faster convergence.

Also, we can see that truncating the coefficients (without adapting during the silences) we have better results using DWT instead of transforming.

ACKNOWLEDGEMENTS

This work has been supported by CICYT under contracts TIC2000-1669-C04-02 and TIC2001-2356-C02-01.

6 REFERENCES