A NEW APPROACH TO MULTICHANNEL AUDIO SIGNAL ACQUISITION AND SUBBAND PROCESSING

Antonio Satué-Villar[1] , Juan Fernández-Rubio[2]

[1]Escuela Universitaria Politecnica de Mataro (EUPMT), Universidad Politécnica de Cataluña, Av. Puig i Cadafalch 101-111, 08303 Mataró, Spain, phone +34937574404, fax +34937570524, e-mail satue@eupmt.es
[2]Escuela Tecnica Superior de Ingenieros de Telecomunicacion de Barcelona (ETSETB), Universidad Politecnica de Cataluña, C. Jordi Girona 1-3, 08034 Barcelona, Spain, phone +34934016431, fax +34934016447, e-mail juan@gps.tsc.upc.es

ABSTRACT

This paper presents results obtained when processing a voice signal collected by several microphones using GSC structure with subband processing. The acquisition has been realized in two ways: by means of an audio acquisition card and a microphone array developed for the work and through an specific system (Mark III structure). We have evaluated the results with different subband transforms (in the adaptive branch of GSC), maintaining the microphones number constant and considering all or part of the adaptive coefficients. The aim of the structure is that, in an environment with several speakers talking in front of the array, the output is as close as possible to the signal of one of them (the one placed orthogonally to the array) and that it adapt quickly to possible environment changes. Thus, by means of arithmetic-harmonic sphericity distances computation quantitative results of a speaker recognition system will be defined.

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 to obtain the desired characteristics.

If the impinging signals are stationary, the weights could be fixed. As this is not the case in a more realistic situation, they must be adapted depending on incoming signal variations. Adaptive algorithms are one of the starting points for this work. All adaptive algorithms can be considered as an 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 is, better results are obtained. Wavelet transform does not only decorrelate input data but it defines non-constant widebands (in FFT bands are the same width), which seems more convenient for speech signals (it can work with octave bands). Thus, the beamforming problem is transformed to a set of less complex problems using narrow band signals. With such decomposition, different adaptation coefficients (μ) can be used.

When working with GSC structure, data can be transformed before adapting the weights. But if we consider that a blocking matrix is needed, we will put the wavelet filters (its specifications will depend on filter characteristics) in the matrix.

This paper deals with this structure behaviour in front of real signals captured in two different ways: with a multipurpose data acquisition card (ChicoPlus system, [1] [2]) and with specific system (Mark III system [3]). So, the exposition is structured as follows: in section 2 we present the working environment and in section 3 the subband decomposition challenges are shown. Section 4 is the main part of the work and it has two subsections: one for the ChicoPlus system and one for Mark III system. In both we describe the system and some results are presented. Finally, in section 5, some conclusions are derived.

2. WORK ENVIRONMENT

As mentioned above, we will use figure 1's structure. We can see that it is a GSC with the blocking matrix integrated inside the wavelet transform, since the wavelet has capacity to block signals by itself.

There are different families of wavelet transforms, each of them with its properties (each one has its characteristic matrix, D). Regularity is one of the main properties, and it represents the transform power to block determinated terms of input signal Taylor's series expansion.

The use of a wavelet transform implies convergence at higher speed (if adapted data are more correlated, adaptation must be faster) and a similar quality perceived by a speaker recognition system. Because of that, they are useful in systems whose goal is to identify speakers.

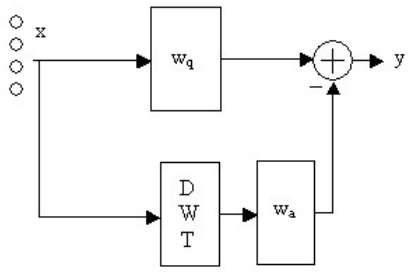

Figure 1: Wavelet-GSC structure

Depending on the wavelet used (Daubechies, biorthogonal,...) we will have one or another D matrix. Weight adaptation is performed as follows:

$$w_a(n+1) = w_a(n) + \mu \cdot y^*(n) \cdot D \cdot x(n)$$

where w_a are the adaptive branch weights, μ is the adaptation coefficient, y(n) is the output, D is the wavelet transform matrix and x(n) is the multichannel input.

The scenario in the simulations is:

- nt=8 taps by filter
- gaussian noise of mean 0 and variance 0.1
- incident signals are speech signals with 45000 samples (sampling frequency=8 KHz) with a mean power of 10000 (40 db)
- ne=4 microphones equal-spaced half wavelength (central frequency=2 KHz).

In the simulations we have used silence-activity detection.

When signals are acquired, arrays enable a spatial filtering of impinging signals so that we can maximize the desired signals and reduce the interfering ones. The problem is that the output array signal can appear as an echo and be captured again by the array. The structure presented does not use available signals as a reference for reducing echoes but this function is developed by the echo cancellers (AEC). There are different solutions for integrating echo cancellers and beamforming. In chapter 13 of [4] we have possible structures.

3. DATA COMPACTATION

An 8-tap structure has been used and has been evaluated the relative importance of the data matrix eigenvalues before the transform and after that. With that information we will see the kind of voice signal compactation performed by the different wavelet transforms. The data presented in table 1 corresponds to the mean of results obtained in several voice frames. This gives an idea of how much information is being wasted if we truncate coefficients.

		2 coef.	4 coef.	6 coef.	8 coef.
Without transform		75,27%	93,82%	99,56%	100%
Db4	Values	80,59%	97,51%	99,98%	100%
	Improv	5,32%	3,69%	0,42%	0%
Db6	Values	82,90%	98,36%	99,98%	100%
	Improv	7,63%	4,54%	0,42%	0%
Bior3.5	Values	75,39%	97,23%	100%	100%
	Improv	0,12%	3,41%	0,44%	0%
Bior2.2	Values	78,07%	95,88%	99,89%	100%
	Improv	2,8%	2,06%	0,33%	0%
Db2	Values	77,87%	95,32%	99,82%	100%
	Improv	2,6%	1,5%	0,26%	0%

Table 1: Transform compactation

To interpret these data we can see, for example, the Daubechies-4 item. If we do not transform and we take 8 coefficients, these carry out full information (100%). If we only take 4 coefficients, 93,82% of the information is carried. If we transform and take all the coefficients, we have full information again but if we only consider 4 of them, 97,51% is achieved. So, if we transform, we recover 3,69% of information.

In the tables we appreciate that transformed signals have more compacted information, so it can be said that truncating these signals is better than truncating the original ones.

4. PRACTICAL REALIZATION

4.1. ChicoPlus system-captured signals

The bus connecting the ChicoPlus card installed in the computer with the Breakout module is shown in figure 2. A second connection is that of each channel of Breakout module with each array microphone.

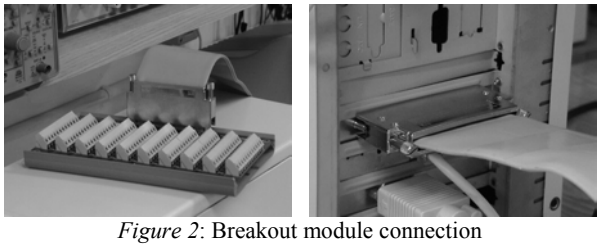

Figure 2: Breakout module connection

The number of each channel in the card of the Breakout module is serigraphied in order to know in which microphone the connections have to be done.

The signal coming out of each microphone is amplified by a pre-amplifier module. These amplifiers are made of operational amplifiers with integrated circuits and they are fed by a ±5 V power supply source.

In the experiment, three speakers (named A, B and C) are placed at the 0º, 50º and -30º angles. Each speaker reads two sentences and the signal captured is made by 4 microphones. The microphones used are very cheap but good enough for our purposes.

Acquisitions were made in 3 sessions; the speakers' positions were different in each session. In the results presented we use results of first session. In it, speaker A is at 0º, B at –30º and C at 50º. Seven different situations are taken into account, and each of them generates 4 files (4 microphones):

- First situation: speaker A reads sentences at 0º.
- Second situation: speaker A reads sentences at 0º while speaker B reads sentences at –30º.
- Third situation: speaker A reads sentences at 0º while speaker C reads sentences at 50º.
- Fourth situation: speaker A reads sentences at 0º while speaker B reads sentences at -30º and speaker C at 50º.
- Fifth situation: speaker B reads sentences at -30º (only speaker B).
- Sixth situation: speaker C reads sentences at 50º (only speaker C).
- Seventh situation: speaker B reads sentences at -30º while speaker C reads sentences at 50º.

Prior to these acquisitions, each speaker was placed at 0 degrees and read a text with a duration of 30 seconds. The files obtained after this capture will be useful for speaker recognition techniques in further work.

We will use the session 1 signals (speaker A at 0º, B at –30º and C at 50º) and particularly 3 of its situations (1: only speaker A; 2: speakers A and B; 3: speakers A, B and C). These signals will be introduced at wavelet-GSC algorithm of figure 1. For every environment 6 transformations have been defined: Daubechies-4, Biorthogonal 3.5, Biorthogonal 2.2, Daubechies-2, Daubechies-6 and without transforming.

We want to see that by doing transforms over the data the output signal quality is not noticeably degraded. We will use a simple method for speaker recognition. The arithmetic-harmonic sphericity distance [5] is evaluated

$$\mu\left(C_j, C_{test}\right) = \log\left[tr\left(C_{test} \cdot C_j^{-1}\right) \cdot tr\left(C_j \cdot C_{test}^{-1}\right)\right] - A$$

where A=2·log(m), C_{test} is the covariance matrix of the array output, C_j is the covariance matrix of the speaker to be tested and m is the matrix dimension (20 in all the experiments).

The Merit Factor is defined as follows:

$$Merit\ Factor = \frac{\mu_{IS} / \mu_{DS}}{\mu_{IE} / \mu_{DE}}$$

where

$\mu_{DE} = \mu$ (C_j target, array input)
$\mu_{IE} = \mu$ (C_j interferer, array input)
$\mu_{DS} = \mu$ (C_j target, array output)
$\mu_{IS} = \mu$ (C_j interferer, array output)

A Merit Factor equal to 1 means that the input is equal to the output (the array does not operate properly). A factor equal to N (N>1) means than the array has multiplied by N its capacity to recognize a speaker. The merit factors obtained (MF) are shown in tables 2, 3 and 4. For every case the merit factor is presented considering 8, 6 and 4 coefficients.

Input MF = 1,7090	8 weights	6 weights	4 weights
Daubechies-4	0,8402	0,8400	0,8398
Biorthogonal-3.5	0,8405	0,8403	0,8400
Biorthogonal-2.2	0,8396	0,8396	0,8398
Daubechies-2	0,8403	0,8401	0,8401
Daubechies-6	0,8396	0,8399	0,8399
Without DWT	0,8403	0,8402	0,8402

Table 2. Combination 1 in a real environment (room)

Input MF = 1,4421	8 weights	6 weights	4 weights
Daubechies-4	0,9024	0,9027	0,9024
Biorthogonal-3.5	0,9028	0,9028	0,9024
Biorthogonal-2.2	0,9023	0,9025	0,9025
Daubechies-2	0,9026	0,9024	0,9026
Daubechies-6	0,9028	0,9025	0,9025
Without DWT	0,9029	0,9028	0,9024

Table 3. Combination 2 in a real environment (room)

Input MF = 1,4012	8 weights	6 weights	4 weights
Daubechies-4	0,9383	0,9382	0,9384
Biorthogonal-3.5	0,9379	0,9381	0,9384
Biorthogonal-2.2	0,9383	0,9384	0,9384
Daubechies-2	0,9383	0,9384	0,9383
Daubechies-6	0,9380	0,9384	0,9384
Without DWT	0,9381	0,9381	0,9385

Table 4. Combination 3 in a real environment (room)

4.2. Signals captured with Mark III system

To verify the algorithms in a different environment some tests have been performed. For that, and with the collaboration of Signal Theory and Communications Department of the Catalonia University of Technology, it was possible to make some data acquisitions with the Mark III array installed in a laboratory. There were 3 speakers, one of them placed orthogonally with the array and the other two at –45º and +45º (approximately). We were specially careful with the first speaker's position as we appreciated in previous work the sensitivity of that parameter. As in previous tests, 4 microphones were used. 3 combinations were defined:

- Combination 1: Only the 0º speaker was present

- Combination 2: Only the 0º and –45º speakers were present
- Combination 3: Three speakers were present

The speakers were different from those in previous simulations so we also carried out an additional acquisition for each of them in order to extract characteristics parameters to obtain the merit factor in every case. These merit factors are shown in tables 5, 6 and 7. For every case, we present the merit factor taking 8, 6 and 4 coefficients. In these simulations we have also worked with 8 microphones and the effect of this increase is not significant.

Input MF 1,0667	ne=4			ne=8		
	8 w.	6 w.	4 w.	8 w.	6 w.	4 w.
Db-4	1,1125	1,1118	1,1121	1,1244	1,1255	1,1245
Bior-3.5	1,1133	1,1116	1,1120	1,1264	1,1247	1,1237
Bior-2.2	1,1129	1,1143	1,1119	1,1235	1,1251	1,1243
Db-2	1,1113	1,1129	1,1123	1,1231	1,1241	1,1253
Db-6	1,1140	1,1138	1,1143	1,1255	1,1246	1,1251
None	1,1138	1,1114	1,1118	1,1252	1,1244	1,1240

Table 5. Combination 1 in a real environment (room) with Mark III

Input MF 0,6498	ne=4			ne=8		
	8 w.	6 w.	4 w.	8 w.	6 w.	4 w.
Db-4	1,1685	1,1677	1,1679	1,1975	1,2018	1,2013
Bior-3.5	1,1708	1,1685	1,1684	1,1977	1,2015	1,2009
Bior-2.2	1,1697	1,1701	1,1679	1,1967	1,1986	1,2021
Db-2	1,1685	1,1688	1,1684	1,1970	1,1981	1,2024
Db-6	1,1702	1,1701	1,1700	1,1988	1,1982	1,1992
None	1,1693	1,1675	1,1682	1,1981	1,2014	1,2002

Table 6. Combination 2 in a real environment (room) with Mark III

Input MF 0,8389	ne=4			ne=8		
	8 w.	6 w.	4 w.	8 w.	6 w.	4 w.
Db-4	1,1085	1,1103	1,1106	1,0926	1,0914	1,0927
Bior-3.5	1,1088	1,1102	1,1108	1,0918	1,0911	1,0928
Bior-2.2	1,1072	1,1081	1,1109	1,0930	1,0934	1,0926
Db-2	1,1092	1,1094	1,1105	1,0923	1,0930	1,0918
Db-6	1,1081	1,1089	1,1089	1,0927	1,0937	1,0935
None	1,1085	1,1107	1,1107	1,0924	1,0916	1,0932

Table 7. Combination 3 in a real environment (room) with Mark III

5. CONCLUSIONS

For the signals obtained with ChicoPlus system the following conclusions can be derived:

- Generally, the merit factors are below 1. This is due to the pointing sensibility of GSC structure. Thus, if desired signal is not exactly at 0º, it comes into the adaptive branch and is cancelled with the signal of quiescent branch. Interferences are reduced, but also the desired signal.
- If we take the data within one combination, we can observe that they are very similar. This is logical, because the use of these transforms in the algorithm must not affect significantly the performance (it must affect speed convergence as commented previously).
- If we do not use all the weights, the results using transforms are better than the results without transforms (due to the capacity for compacting the data).

In the case of the signals obtained with Mark III array, merit factors are higher than the unit. Generally, increasing microphones leads to an increase of merit factor but it is not higher than 3%. In a reverberant environment, an increase in the microphone number does not imply a significant increase in performance. Reverberant environments are difficult to deal with and some solutions are proposed in the literature. So, while in [7] the knowledge that reverberant channels are low-pass for using a blind method to reduce the effect of these reverberations is taken in account, in [8] it is preferable to separate the reverberation problem from the conformation problem, placing an echo canceller system at the beamformer input. Even so, in [9] we can see that subband beamformers improve the results with reference to the methods that do not divide the signal into bands. Previous conclusions are also equally applied to this instance.

6. ACKNOWLEDGEMENTS

This work has been partially supported by the European Commission (FEDER) and Spanish/Catalan Government under projects TIC2003-08382-C05-02, TIC2003-05482, TEC2004-04526 and 2001SGR-00268.

7. REFERENCES

[1] Reference Handbooks of ChicoPlus card, Omnibus module and Armada software. Innovative Integration.

[2] www.innovative-dsp.com Web of Innovative Integ. enterprise

[3] Rochet Cedrick; "Documentation of the Microphone Array Mark III"; Information Access Division of the National Institute of Standards and Technology, sep. 2003; http://www.nist.gov/smartspace/toolChest/cmaiii/userg/Microphone_Array_Mark_III/

[4] Microphone arrays, Brandstein & Ward, Springer-Verlag 2001

[5] Bimbot F., Mathan L., "Text-free speaker recognition using an arithmetic-harmonic sphericity measure" *pp.169-172, Eurospeech 1991*

[6] Griffitths, Jim, "An alternative approach to linearly constrained adaptive beamforming", IEEE Trans. Antenna Propagation vol. AP.30 núm. 1 pp. 27-34, aug. 1982

[7] J.Liu, H.Malvar, "Blind deconvolution of reverberated speech signals via regularization", ICASSP 2001

[8] W.Herbordt, H.Buchner, "Computationally efficient frequency-domain combination of acoustic echo cancellation and robust adaptive beamforming", Eurospeech 2001

[9] W.Neo, B.Farhang-Boroujeny, "Robust microphone arrays using subband adaptive filters", ICASSP 2001