MICROPHONE ARRAY DESIGN FOR ROBUST SPEECH ACQUISITION AND RECOGNITION

Julian Fernández, Eduardo Lleida, Enrique Masgrau
Dept. Ingeniería Electrónica y Comunicaciones
Universidad de Zaragoza, Zaragoza 50015, Spain
navajas@posta.unizar.es
http://www.cps.unizar.es

ABSTRACT

The aim of this paper is to study the use of a robust acquisition system based on a microphone array for speech related applications in real situations. A comparison is performed between two beamforming methods: the Delay and Sum beamforming (**DS**) and the Spatial Reference Optimal beamforming (**SRO**). Both of them are frequency domain designed, using harmonic spatial distributed microphones. The quality of the microphone array output signal for the different beamforming approaches is measured by using a continuous speech recognition system based on discrete HMMs. Results of the speech recognition system show the advantage of using a microphone array with the **SRO** beamforming as front-end.

Keywords: microphone array, beamforming, speech recognition

1. INTRODUCTION

Hands-free telecommunications, videoconference, speech recognition are some speech-related applications, which need a robust acquisition system when working in real acoustic scenarios with environment noise and reverberation. Close-talk microphones are a solution used typically to improve the signal to noise ratio of the speech. Another potential solution is the use of microphone arrays, which has the advantage of really hand-free operation[1,2].

The microphone array makes use of beamforming techniques to fight against the effects of the acoustics environments. In this paper, a comparison is performed between two beamforming methods: the Delay and Sum beamforming (**DS**) and the Spatial Reference Optimal beamforming (**SRO**). Both of them are frequency domain designed, using harmonic spatial distributed microphones. The purpose of the **SRO** beamforming is to minimize the output signal power of the array maintaining the output signal in the desired direction [3]. However, when working in a reverberate environment, the minimization criterion used by the **SRO** beamforming produces the cancellation of the desired signal with the reflections. This problem is addressed in the paper.

This paper is organized as follows. In section 2 we present an overview of the system, the speech recognition system and the speech adquisition system based on a microphone array. In section 3, we discuss the recognition experiments and the results. Finally we summarize our major findings.

2. SYSTEM OVERVIEW

2.1 The Speech Recognition System

The speech recognition system is a continuous speech recognition system based on discrete hidden Markov models driven by a stocastic language grammar. 25 context independent phones were trained using a clean data base collected with a close-talk microphone. Mel-cepstrum, first and second order differential parameters plus the differential energy were employed. A vector quantizer of 256, 128, 128 and 64 codewords were used. The speech recognition is working in a client/server architecture. Speech analysis is performed in the client. Speech parameters (VQ index) are sent to the server using the TCP/IP protocol. The speech decoder is running in the server in a high speed workstation.

2.2 The Speech Adquisition System

The speech adquisition system is composed by an microphone array working in the frequency-domain. A general structure of the frequency-domain beamforming is shown in Figure 1, where the broad band speech signal from each microphone is transformed into frequency domain using a FFT.

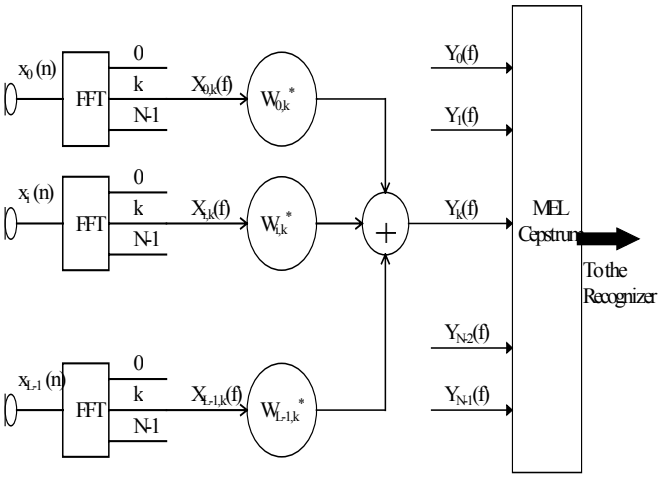

Figure 1. Broad band frequency-domain beam-former

Figure 2 shows the geometry of a 9 microphone array. Assuming a sampling frequency of 8 kHz, the total bandwidth of 4 kHz has been divided in three bands from 50 Hz to 1kHz (band I), 1kHz to 2 kHz (band II) and 2kHz to 4 kHz (band III) with a microphone distance of 0.16, 0.08 and 0.04 meters respectively. Each band is composed by 5 microphones, microphones 0,1,4,7 and 8 for band I, microphones 1,2,4,6,7 for band II and microphones 2,3,4,5,6 for band III.

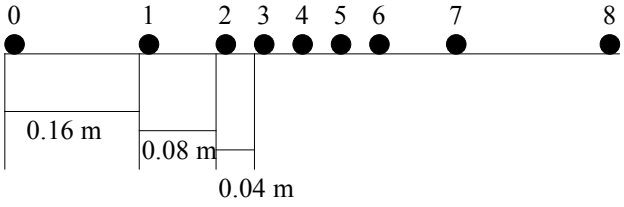

Figure 2. Microphone array geometry

Each frequency bin is processed by a narrow-band beamformer. Narrow band signals are weighted and summed to produce an output for each frequency bin. An overlap-add synthesis procedure can be used to get the speech signal, but in our approach, the output for each frequency bin is used to computed the MEL-scaled cesptrum used by the speech recognition system.

2.3 Optimal beamforming in reverberant environment

Given a narrow-band array of L microphones, the purpose of the optimal beamforming is to maximize the SNR without producing any distortion in the source signal. Suppose an acoustic scenario with a desired signal and M incorrelated interferences. The design of the array weights is formulated by introducing a set of constrains for the array output in the interference directions. The constraints force to cancel the array output in the direction of the inteferences, so the optimization problem is to minimize the mean output power while maintaining unity response in the desired direction and cancelling the mean output power in the interference directions. Defining the vector $\underline{f}_k=[1,0,...,0]^H$, the M+1 constraints can be formulated as

$$\underline{w}_k^H \underline{\underline{C}} = \underline{f}^H \qquad (1)$$

where the matrix $\underline{\underline{C}} = \left[\underline{s}_{d,k}, \underline{s}_{0,k}, \ldots, \underline{s}_{M-1,k}\right]$ is the matrix with the steering vectors associated with the desired direction and the M interferences and $\underline{W}_k=[w_{0,k}, w_{2,k}, ..., w_{L-1,k}]$ be a complex L-dimensional vector representing the weights of the beam-former for the *k*th narrow band. With these restrictions, the optimal weights are given by the expression [4]

$$\underline{W}_k = \underline{\underline{R}}_k^{-1} \underline{\underline{C}} \left(\underline{\underline{C}}^H \underline{\underline{R}}_k^{-1} \underline{\underline{C}} \right)^{-1} \underline{f} \qquad (2)$$

where $\underline{\underline{R}}_k = E\left[\underline{X}_k(f)\underline{X}_k^*(f)\right]$ is the autocorrelation matrix of the signal in the k-th narrow band.

However, when dealing with a real acoustic environment, a new problem is added, the reverberation. The reverberation produces the presence of coherent signals in different directions having a pernicious effect on the beamfomer. As the optimization criterion for the microphone array design is to minimize the output signal power maintaining a unit gain in the desired direction the result is a cancellation of the desired signal at the output. Assuming the knowledge of the direction of the desired signal, a modification could be done in the **SRO** beamforming to overcome the cancellation problem. As the cancellation is due to the presence of coherent signals (reflections of the desired signal) in different directions, the modification consists on performing the computation and adaptation of the beamformer response in the time intervals without desired signal. In this case, there is no coherent signals with the desired speech source and the beamformer diagram has a unit gain in the desired direction, minimizing the level of the interference signals and the non-directional noise. Another implementation issue of the **SRO** beamforming is the level of the sidelobes in those directions without signal. To control the level of these sidelobes, a virtual omnidirectional noise is introduced in the correlation matrix (adding a constant in the main diagonal) used by the minimization procedure. The signal to noise ratio between

the desired signal and the virtual omnidirectional noise is called **SVNR** and the signal to noise ratio between the desired signal and the interference is called **SIR**. When the **SVNR** is low, the **SRO** beamformer is closer to the **DS** beamformer due to the dominant main diagonal in the correlation matrix.

3. EXPERIMENTS AND RESULTS

Speech recognition experiments has been performed by simulating the acoustic scene and the microphone array. The test database consists on oral inquiries into a geographic information database. The vocabulary has 1250 words and the average number of words for sentence is greater than 9. The test material consists on 510 utterances from 12 speakers. The perplexity of the test database is 10. The word recognition rate in clean conditions is 91.68 %.

For each test database sentence, the transfer functions between the desired speech source and the microphones are simulated. The same procedure is done for the interference signal that will be composed by a mix of four signals from four different speakers. We assume that the direction of the desired source is known.

Figure 3 shows the room and the positions of the the microphone array and the speech sources (desired and interference). The room is 10m x 12m x 3m, and the microphone array is situated in (2.5m , 0 m , 1m). The desired speech source is in the broadside direction (2.5 m , 10 m ,1 m) and the interference is in (2.5+10*sin(30º)m, 10*cos(30º)m, 1m).

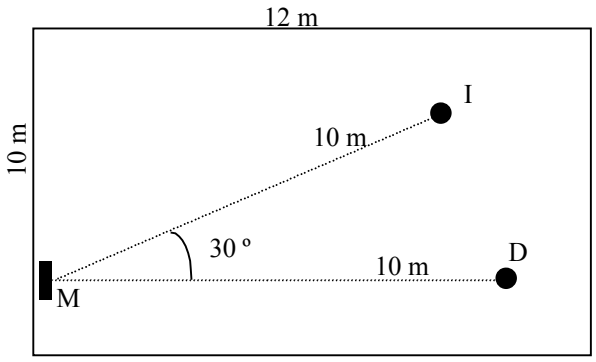

Figure 3. Room dimensions and positions

Assuming an ideal room without reverberation, table I shows the word recognition performance of the speech recognition system for different signal to interference ratios (**SIR**). Results are given when using only one microphone (**TASinP**), using the Delay and Sum beamforming (**TASDS**) and using the Spatial Reference Optimal beamforming (**TASRO**) with different signal to virtual omnidirectional noise ratios (**SVNR**). There is a fast degradation in the performance when using only one microfone, the delay-sum microphone array improve the performance over the one microphone approach and the **SRO** beamformer outperform both, mantaining a high word recognition rate (81.59 %) with a **SIR** of 5 dB.

SIR (dB)	TASinP (%)	TASD (%)
5	31.23	52.92
10	56.83	76.82
15	82.44	86.17

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	57.91	80.68	81.59	76.25	67.28
TASRO SIR=10dB	79.16	85.75	84.34	77.70	65.96
TASRO SIR=15dB	85.73	88.87	87.29	78.76	66.94

Table I. Word recognition performance for a room without reverberation.(TASinP: one microphone, TASD: delay-sum, TASRO: Spatial reference).

However, when working with a real room with reverberation, the performance of the **SRO** beamformer is degraded due to the cancellation between the signal in the desired direction and the reflections in different directions. Table II and III show the performance for the three kind of speech adquisition systems when the reverberation time is 0.2 and 0.4 secons respectivelly. It can be note the degradation in the performance of the **SRO** beamformer due to the reflections of the desired speech signal that are arriving in different directions. To solve this problem, the adaptation of the weights of the beamformer are done during the pauses of the desired speaker. In this situation, there is no desired signal and the bemaformer is conformed given a unit gain in the desired direction and minimizing the output for the rest of directions.

SIR (dB)	TASinP (%)	TASD (%)
5	27.76	54.40
10	56.28	76.40
15	80.86	85.85

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	57.23	80.79	69.66	20.62	6.20
TASRO SIR=10dB	77.75	86.50	69.69	17.52	5.17
TASRO SIR=15dB	86.45	88.17	69.83	12.74	3.93

Table II. Word recognition performance for a room with reverberation time 0.2s.(TASinP: one microphone, TASD: delay-sum, TASRO: Spatial reference).

SIR (dB)	TASinP (%)	TASD (%)
5	21.74	41.64
10	47.33	72.70
15	74.50	83.91

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	44.48	67.13	27.70	16.32	12.24
TASRO SIR=10dB	73.64	74.74	27.56	14.84	12.13
TASRO SIR=15dB	83.81	80.32	24.82	10.70	5.59

Table III. Word recogntion performance for a room with a reverberation time of 0.4s. .(TASinP: one microphone, TASD: delay-sum, TASRO: Spatial reference).

Tables IV, V and VI show the results when adapting the weigths of the beamformer in the desired speaker pauses for a room with a reverberation time of 0.2 , 0.4 and 1.6 seconds respectivelly. The improvement in the word recognition rate is notable. By using a **SVNR** of 15 dB and a **SIR** of 5 dB, the word recognition rate improve from 54.4 % for the **SD** to 89.14 % for the **SRO** with a reverberation time of 0.2 seconds and from 41.64 % for the **SD** to 86.36 % for the **SRO**, which means more than a 100 % improvement in the recognition rate and only a degradation of only a 5 % over the clean conditions. For a high reverberant room (1.6 seconds), the performance still is mantained, a 75.74 % word recognition rate for the **SRO** beamformer while the **SD** beamformer gives a 26.90 % word recognition rate. Also, the performance is maintained when going from a **SIR** of 15 dB to 5 dB. There is only a degradation of 6% in the word recognition rate. So it is clear the advantage of adapting the beamforming weights when no coherent signals with the desired signal are present in the acoustic scene.

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	57.96	85.04	89.14	88.28	87.75
TASRO SIR=10dB	76.27	88.81	90.25	90.31	86.97
TASRO SIR=15dB	78.68	83.82	86.57	84.91	81.30

Table IV. Word recogntion performance for a room with a reverberation time of 0.2s. adapting the weigths in the desired speaker pauses.

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	41.97	81.55	86.36	85.54	81.68
TASRO SIR=10dB	72.41	87.35	87.97	86.87	82.82
TASRO SIR=15dB	83.72	89.21	87.15	84.57	79.12

Table V. Word recogntion performance for a room with a reverberation time of 0.4s. adapting the weigths in the desired speaker pauses.

SIR (dB)	TASinP (%)	TASD (%)
5	18.74	26.90
10	20.96	56.28
15	59.53	72.58

SVNR (dB)	-15	0	15	30	60
TASRO SIR=5dB	28.93	58.48	75.74	77.52	76.23
TASRO SIR=10dB	60.38	74.52	79.48	78.32	75.47
TASRO SIR=15dB	73.02	79.40	80.24	76.83	75.32

Table VI. Word recogntion performance for a room with a reverberation time of 1.6s. adapting the weigths in the desired speaker pauses. (TASinP: one microphone, TASD: delay-sum, TASRO: Spatial reference).

4. CONCLUSIONS

In this paper, a robust speech adquisition system for speech recognition applications has been presented. A comparison between a delay-sum beamformer and a spatial reference optimal beamformer has been done in a reverberant environment. We have addressed the coherence problem in the estimation of the beamfomer weights for the spatial reference optimal beamformer. A practical solution has been proposed and tested based on the adaptation of the beamformer weights in the pauses of the desired signal. An improvement of more than 100 % over the delay-sum beamformer is obtained by using our solution for a wide range of reverberation times.

Acknowledgments

This work has been supported by the CICYT under contract TIC98-0423-C06-04 and CONSID-DGA P-64/96

5. REFERENCES

[1] Omologo M., Matassoni M., Svaizer P., Giuliani D. (1997), Microphone array based speech recognition with different talker-array positions. *Proceedings of ICASSP-97*, pp. 227–230.

[2] Grenier Y., Affes S. (1997), Microphone array response to speaker movements. *Proceedings of ICASSP-97*, pp. 247–250.

[3] Lleida E., Fernández J., Masgrau E. (1998), Robust Continuous Speech Recognition System based on a Microphone Array. *Proceedings of ICASSP-98*, vol 1. pp. 241–244.

[4] Jonhsons D.H., Dungeon D.E. (1993), Array Signal Processing: Concepts and Techniques. Prentice-Hall.