# ROBUST CONTINUOUS SPEECH RECOGNITION SYSTEM BASED ON A MICROPHONE ARRAY

*E. Lleida, J. Fernández, E. Masgrau*

Dpt. Electronics and Communications Engineering
Centro Politécnico Superior, University of Zaragoza
Maria de Luna 3, Zaragoza, SPAIN
{lleida,navajas,masgrau}@posta.unizar.es

## ABSTRACT

In this paper, a robust speech recognition system for videoconference applications is presented based on a microphone array. By means of a microphone array, the speech recognition system is able to know the position of the users and increase the signal-to-noise (SNR) ratio between the desired speaker signal and the interferences from the other users. The user positions are estimated by means of the combination of a direction of arrival (DOA) estimation method with a speaker identification system. The beamforming is performed by using the spatial references of the desired speaker and the interference locations. A minimum variance algorithm with spatial constraints working in the frequency domain is used to design the weights of the broad band microphone array. Results of the speech recognition system are reported in a simulated enviroment with several users asking questions to a geographic data base.

## 1. INTRODUCTION

One major problem of the speech recognition systems working in real situations is the need of the use a close-talk microphone to reduce the environmental noise. In some scenarios, as a information booth, a videoconference system,..., is annoying to use such microphones and a microphone array should be used. Microphone array allows to improve both the signal-to-noise ratio (SNR) and the mobility of the user [ 1,2 ]. SNR is increased by using an optimal beam-forming procedure and the mobility of the user is allowed by using a source location procedure working jointly with the beamformer.

Delay-and-sum beamforming is widely used because it is easy to design but it doesn't achieve a maximization of the SNR. In this paper, we propose to use a frequency-domain optimal beam-former by using spatial references to maximize the SNR. Spatial references are the location of the sound source and the interferences. The desired source and interference location are estimated by means of a combination of a speaker identification system and a direction of arrival estimation method. The frequency-domain beamforming is used as front-end of a continuous speech recognition system based on hidden Markov models (HMM).

This paper is organized as follows. In Section 2 we present an overview of the speech recognition system. In Section 3 we discuss the optimal frequency-domain beam-forming procedure. In Section 4 the source location procedure is described. In Section 5 we present and discuss the results. Finally we summarize our major findings and outline our future work.

## 2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the system. L microphones are arranged in an array structure. A source localization procedure estimates the direction of arrival of each signal present in the auditory space, given to each direction a label of desired signal or interference by using a speaker identification system. The speaker identification system use gaussian mixture models (GMM) to model each speaker and labeling the different detected signals directions [rose94]. These directions of arrival are used by the beamformer step to design a optimal beamformer by canceling the interferences and maximizing the SNR in the desired direction.

The speech recognition system is a continuous speech recognition system based on discrete hidden Markov models driven by a finite state 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.

The speech recognition system is an interface component of a videoconference system. Users involved in the videoconference can use the speech recognizer to retrieve information from a data base or to control the videoconference system. At the beginning of the session, the speaker identification system is trained with 30 seconds of speech. The speaker identification system jointly with the direction of arrival (DOA) estimation system is used to learn the initial position of every speaker in the acoustic space. After this initialization, the position is tracked every second with the DOA estimation and the speaker identification system is activated only in the case of an ambiguity in the tracking and labeling of the speakers. In this paper, we present results from the point of view of the speech recognizer, we don't evaluate the speaker identification system.

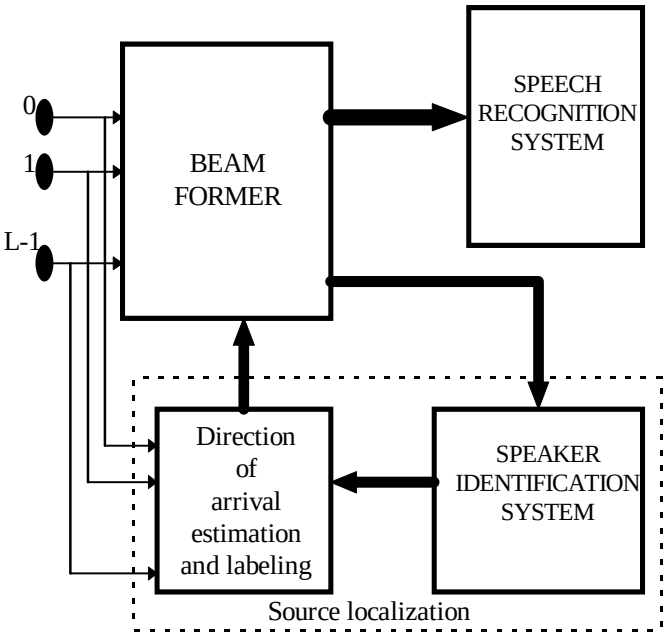

Figure 1. System block diagram

## 3. BEAMFORMING

### 3.1 Frequency-domain beamforming

A general structure of the frequency-domain beamforming is shown in Fig. 2, where the broad band speech signal from each microphone is transformed into frequency domain using a FFT.

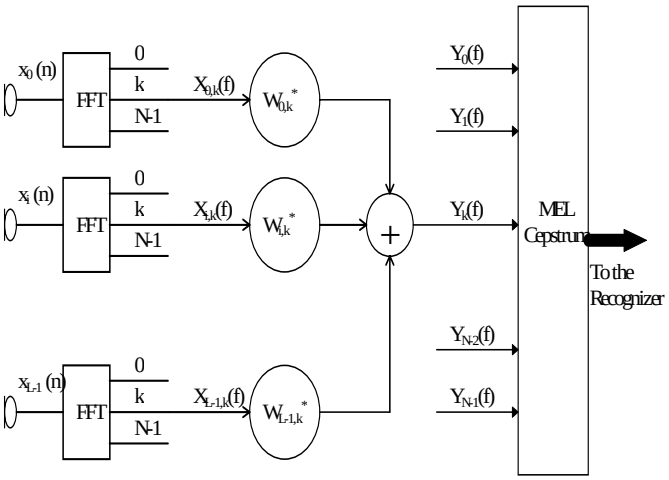

Figure 2. Broad band frequency-domain beam-former

A microphone array is working with a narrow-band signal when the total dimension (D) of the array holds the following relationship [4]

$$D << \frac{c}{\pi B} \qquad (1)$$

where B is the signal bandwidth and c is the sound velocity. In terms of the FFT order N and sampling frequency $f_m$, this condition holds

$$D << \frac{2 N c}{\pi f_m} \qquad (2)$$

On the other hand, the constraint to avoid spatial aliasing holds that the distance between microphones must be

$$d < \frac{\lambda_{min}}{2} = \frac{c}{2 f_{max}} \qquad (3)$$

For a speech signal with a sampling frequency of 8 kHz, these conditions imposes the following constraint in the microphone array design : d < 0.0425 meters and N>>37*D=37*(L-1)d, being L the number of microphones. However, as the resolution of the array is proportional to the relation D/λ, if d is fixed to hold the spatial aliasing constraint for the maximum frequency, then there is a significant lose of resolution at low frequencies. To overcome this problem a three section microphone array has been designed (harmonic array). 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. Figure 3 shows the geometry of a 9 microphone array. 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.

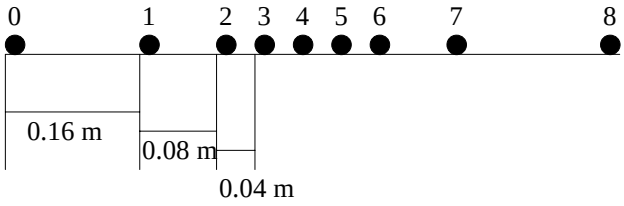

Figure 3. Microphone array geometry

Given the desired distance between microphones to fulfill the spatial aliasing constraint, the order of the FFT for each band must satisfy the narrow-band condition given in (2). Assuming a sampling frequency of 8 kHz, a FFT of order 512 for band I, 256 for band II and 128 for band III are used. As all the FFTs span the entire range of frequencies from 0 to 4 kHz, the set of frequency bins expanding the bands are selected as input to the beamformer ( band I is expanding by bins 4 to 64 of the 512, band II by bins 33 to 64 of the 256 and band II by bins 33 to 64 of the 128).

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 IFFT 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.

### 3.2 Optimal BeamForming

Given a narrow-band array of L microphones, the purpose of the optimal beam-forming is to maximize the SNR without producing any distortion in the source signal.

Let $\underline{W}_k=[w_{0,k}, w_{2,k}, ..., w_{L-1,k}]$ be a complex L-dimensional vector representing the weights of the beam-former for the *k*th narrow band, $X_k(f)=[X_{0,k}(f),X_{1,k}(f),...,X_{L-1,k}]^T$ the *k*th narrow band input vector signal and $Y_k(f)$ the *k*th narrow band output signal. The *k*th narrow band array output is given by

$$Y_k(f)=\sum_{i=0}^{L-1} X_{i,k}(f)\,W_{ik}^{*}=\underline{W}^{H}\,\underline{X}(f) \qquad (4)$$

where superscripts * and H denotes the complex conjugate and the complex conjugate transpose.

Denote $\underline{s}_{i,k}$ the steering vector associated with the direction $\theta_i$ for the *k*th narrow band as

$$\underline{s}_{i,k}=[\exp(j2\pi f_k\tau_0\theta_i),...,\exp(j2\pi f_k\tau_{L-1}\theta_i)] \qquad (5)$$

The weights are the solution of the following optimization problem : minimize the mean output power while maintaining unity response in the desired direction. This minimization procedure minimizes the total noise in the output signal but keeping the output desired signal power constant which means that the output SNR is maximized. The optimal weights are given by the expression [5]

$$\underline{W}_k=\frac{\underline{\underline{R}}_k^{-1}\underline{s}_{d,k}}{\underline{s}_{d,k}^{H}\underline{\underline{R}}_k^{-1}\underline{s}_{d,k}} \qquad (6)$$

where $\underline{s}_{d,k}$ is the steering vector associated with the desired direction and $\underline{\underline{R}}_k=E\left[\underline{X}_k(f)\underline{X}_k^{*}(f)\right]$. We will refer to this beam-former as optimal with spatial reference. An additional improvement in the SNR can be obtained by placing some constraints in the direction of the interferences.

Suppose an acoustic scenario with a desired signal and M interferences. The design of the array weights is formulated in the same terms as before introducing a new set of constrains for the array output in the interference directions. The new constraints force to cancel the array output in the direction of the inteferences, so the new optimization problem is to minimize the mean output power while maintaining unity response in the desired direction and canceling 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

$$\underline{w}_k^{H}\underline{\underline{C}}=\underline{f}^{H} \qquad (7)$$

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. With these new restrictions, the optimal weights are given by the expression [5]

$$\underline{W}_k=\underline{\underline{R}}_k^{-1}\underline{\underline{C}}\left(\underline{\underline{C}}^{H}\underline{\underline{R}}_k^{-1}\underline{\underline{C}}\right)^{-1}\underline{f} \qquad (8)$$

This solution allows to maximize the SNR when the number of interferences is less than or equal to L-2. A microphone array with L microphones has L-1 degrees of freedom and one has been used to set the constraint in the desired direction.

## 4. SOURCE LOCATION

The problem of the source location can be seen as a problem of finding the direction of arrival (DOA) of the acoustic waves to the microphone array. Once the direction of arrival has been detected, a label must be assigned to each direction as desired signal or interference. For this purpose, a speaker identification system trained with the voice of the system users is used. The speech signal is labeled as background, speaker interference and speaker desired. In this way, the source localization output is a direction (azimuth angle) with a label. The direction of arrival is estimated by using the band II microphone array (1 kHz to 2 kHz). The DOA is computed by accumulating the estimated DOA for each frequency bin in band II. The DOA estimation is based on the linear prediction method. This method estimates the output of one microphone using a linear combination of the remaining microphone outputs by minimizing the mean square prediction error. The DOA are estimated by looking for the maxims of the spatial spectrum defined as [5]

$$DOA(\theta)=\frac{\underline{u}^{H}\underline{\underline{R}}^{-1}\underline{u}}{\left|\underline{s}_{\theta}^{H}\underline{\underline{R}}^{-1}\underline{u}\right|^{2}} \qquad (9)$$

where $\underline{u}$ is a vector with all zeros except one element, which is equal to one and $\underline{s}_{\theta}$ is the steering vector associated with the direction $\theta$. This method performs well in low SNR environments with a high resolution.

## 5. RECOGNITION EXPERIMENTS

### 5.1 Databases

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[6]. The vocabulary has 200 words and the average number of words for sentence is greater than 9. The test material consists on 510 utterances from 12 speakers. A finite state grammar is used to drive the speech recognition decoder.

Two kind of interferences has been simulated, a white noise directional interference and a speech interference both with a 0 dB of SNR. Results are presented for one and two interference sources plus the desired signal. The desired signal is situated in the broadside direction (0º) and the interferences at 32º and -32º respectively.

Table I summarizes the word accuracy performance for the speech recognition system when the clean speech is used and when the noisy speech is collected with just one microphone.

| | Clean speech | White noise | Speech interference |
|---|---|---|---|
| Word Acc. | 99,3 % | 10,1 % | 34,8 % |

Table I. Word Accuracy performance of the baseline system

### 5.2 Recognition Experiments

Table II shows the word accuracy performances when using three different techniques to design the beamformer. The first algorithm is the delay-sum (DS) where the output of the microphones are delayed according to the direction of arrival of the desired speech signal and summed. The second one is a minimum variance estimation (MVE) without interference constraints (Eq. 6). The third one is a minimum variance estimation with interference constraints (MVE-IC) as in Eq. 8. Results are given when a priory knowledge of the situations of the users is known (+ priory label) and when the automatic location is used (+ location label). All results are given for a white noise (WN) interference and a speech interference (SI) both at 32 º.

| | DS | MVE | MVE-IC |
|---|---|---|---|
| WN + priory | 97,6 % | 99,1 % | 99,2 % |
| WN + location | 95,8 % | 96,9 % | 97,7 % |
| SI + priory | 64,9 % | 87,8 % | 97,7 % |
| SI + location | 64,0 % | 85,0 % | 97,0 % |

Table 2. Word accuracy using 3 beamforming algorithms with a priory knowledge or automatic estimation of the direction of arrival.

When the interference is a white noise, any of the three beamforming techniques give a big improvement in the performance. MVE-IC gives the best result with a 99,2 %, almost the same recognition rate of the clean speech. The use of automatic source location produces a slight decrease, less than 2 %, in the word accuracy performance. However, when a speech interference is used, the difference in the performance of the three techniques is highlighted. MVE-IC almost cancel completely the interference, maintaining the word accuracy at the level of the clean speech. Adding another speech interference in the acoustic scenario (-32º), doesn't change the performance of the system, keeping the word accuracy around 97 %.

A set of experiments were carry out to study the sensibility of the speech recognition performance when there is an error in the estimation of the direction of arrival. The desired source comes in the broadside direction (0º) and the speech interference is at 32º. Table 3 shows the word accuracy performance for different location errors. It can be seen that in the range of $\pm 5^{\circ}$ there is a decrease in the word recognition performance of less than 4 % which shows the robustness against small location errors.

| Location error | +10º | +5º | 0º | -5º | -10º |
|---|---|---|---|---|---|
| Word accuracy | 90.6 % | 96.4 % | 97,7% | 93.6 % | 79.7 % |

Table III. Word accuracy with source location errors

## 6. CONCLUSIONS

In this paper, a robust speech recognition system based on a microphone array has been presented. The microphone array has been designed to hold the spatial aliasing constrain maintaining a good spatial resolution. The microphone array has two important functions : 1) improve the SNR by optimal beamforming and 2) estimate the source direction of arrivals. Optimal beamforming is performed by applying the minimum variance estimation with spatial constraints. The estimation of the direction of arrival of the acoustic sources is done by means of the combination of a linear prediction DOA method and a speaker identification system. Speech recognition results has been given simulating a videoconference system with 2 and 3 speakers. MVE-IC beamforming technique allows to maintain the recognition performance of clean speech for a wide range of interference conditions, outperforming the simple delay-sum technique.

A future point to address is the tracking capabilities of the system when a more complex scenario is presented and the acoustic environment, reverberation, is taking into account.

## Acknowledgments

This work has been supported by the CICYT under contracts TIC95-0884-C04-04 and TIC95-1022-C05-02

## 7. REFERENCES

[1] M. Omologo, M. Matassoni, P. Svaizer, D. Giuliani, "Microphone array based speech recognition with different talker-array positions", Proc. ICASSP97, pp 227-230, 1997.

[2] Y. Grenier, S. Affes, "Microphone array response to speaker movements", Proc. ICASSP97, pp 247-250, 1997.

[3] D.A. Reynolds, R.C. Rose, "Robust Text-Independent Speaker Identification Using Mixture Speaker Models", IEEE Trans. On Speech and Audio Processing, pp 72-83, vol 3. Jan 1995.

[4] R.A. Monzingo, T.W. Miller, *Introduction to adaptive arrays,* Wiley&Sons, 1980.

[5] D.H. Jonhson, D.E. Dungeon, *Array Signal Processing : Concepts and Techniques,* Prentice-Hall, 1993.

[6] F. Casacuberta et al. "Development of Spanish Corpora for Speech Recognition Research", Proc. of Workshop on Int. Cooperation and Standarization of Speech Databases and Speech I/O Assesment Methods, Chiavari, 1991.