MULTIRESOLUTION MICROPHONE ARRAY FOR SPEECH SOURCE ACQUISITION AND TRACKING

Djamila Mahmoudi *Andrzej Drygajlo*

Signal Processing Laboratory
Swiss Federal Institute of Technology at Lausanne, CH-1015 Lausanne, Switzerland
e-mail: mahmoudi@lts.de.epfl.ch

ABSTRACT

This paper presents a new method of localization and tracking of a moving speech source in an adverse environment using a microphone array. To provide a good quality of communication, the problem of enhancing the acquired speech signal is also addressed. The source localization is applied to the sub-band decomposition and is achieved in two stages. First, the location of a coarse region where the speech source is present is detected. Then the multi-beamforming operation together with the discrimination function are used to pinpoint the speaker's location. Both stages are based upon the concept of examining the beam signal energy and its variation. Then, a postfiltering based noise reduction system is applied to attenuate the noise effect. The complete algorithm provides high noise reduction and relatively small errors in the source localization estimate. The system was developed in the multiresolution wavelet transform domain using a fast algorithm with short prototype filters. This results in a significant reduction in computational load and gives a minimum delay in sub-band processing.

1. INTRODUCTION

The performance of hands-free voice communication systems are usually disturbed by interfering sources such as ambient noise, especially when the system is designed to operate in an adverse acoustic environment like an office room.

The major advantage of a microphone array is its capability to electronically alter its direction of reception, while attenuating the interfering sources. This feature makes it very useful in situations involving moving sources. Therefore, it tends to replace the head-mounted microphone in many applications, especially in teleconferencing [1, 2] and voice communication using a personal terminal in a noisy environment [3, 4].

Three types of microphone array speech enhancement systems have been proposed: conventional or delay-and-sum beamforming (DSB), the noise reduction systems which consist of DSB including an additional post-filtering to improve the system output [3, 4], and the adaptive beamforming system [5]. On the one hand, the adaptive beamforming system has proved its utility, especially if the number of the noise sources is less than the number of the microphones. However, this system is highly sensitive to source localization errors and very costly in terms of calculations. On the other hand, the post-filtering based noise reduction system is efficient, but it suffers from a musical noise making an additional post-processing necessary.

All these systems can be applied to the moving speech source scenario by adapting the algorithm to every frame of the speech signal corresponding to a novel speaker's position. However, in all these cases, a high accuracy in the source localization is required.

Widely used source localization techniques include multi-beamforming based methods [1] and the correlation function based methods employing time-difference of arrival between the microphone signals [6]. The first type of methods collect various versions of the beamformer output when the array is steered in different directions. Unfortunately, these methods strongly depend on a discrimination function between a large number of the formed beams to select the source location, which is not easy to determine. Consequently, the resulting computational burden makes the operation unsuitable for real-time applications. The second method is based on the cross-correlation function. It computes a set of time delays between different microphone signals, or filtered versions of them, to estimate the direction of arrival (DOA) of the desired source [6]. Unfortunately, these methods fail in the reverberant environments. Furthermore, in case of a speech signal, this technique is limited by the presence of the pitch. The duration of the observation and noise level are also critical parameters.

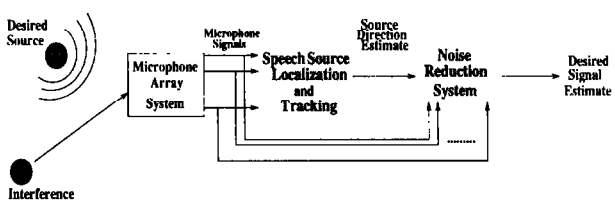

Figure 1. *Block diagram of the proposed system.*

The problem of central interest herein is that of estimating the DOA of the desired signal impinging on a receiving array. The basic structure of the proposed system is shown in Fig. 1. Obviously, the aim is to ensure good quality of the acquired signal and to incorporate the tracking capabilities.

Our work focuses on the localization and tracking of a moving speech source, relative to the microphone array,

in an adverse environment. Furthermore, the interfering sources are attenuated. The system is intended to be applied to voice communication through a personal terminal in an office environment. The broadband nature of the speech signal is taken into account in the microphone array design. The paper is organized as follows: in Sec. 2 we address the problem of designing a microphone array for the broadband speech signal. Sec. 3 presents the speech source localization and tracking procedure. The noise reduction method is discussed in Sec. 4. Finally, Sec. 5 describes the simulations setup, some results are also shown and conclusions are drawn.

2. MICROPHONE ARRAY SYSTEM

In the uniform array case, a large number of microphones is needed to avoid spatial aliasing while ensuring a sufficient directivity in the whole speech bandwidth. Thus, the number M of microphones and their spacing d are the critical parameters. Consequently, the linear nonuniform array appears to be a reasonable solution. In particular, a logarithmic distribution of microphones ensures the required array length and the microphone spacing with a small number of microphones [7]. Nevertheless, the sidelobes level is not negligible and can induce errors in the source localization estimate. To avoid such errors,

Sub-array	Microphones	sub-band (octave)
sb_1	M_1, M_5, M_6	250 Hz – 500 Hz
sb_2	M_1, M_4, M_5	500 Hz – 1000 Hz
sb_3	M_1, M_3, M_4	1000 Hz – 2000 Hz
sb_4	M_1, M_2, M_3	2000 Hz – 4000 Hz

Table 1. *Space-frequency decomposition.*

we superpose several sub-arrays and we decompose each microphone signal into frequency sub-bands. Then, each sub-band is spatially filtered by its appropriate sub-array. The adopted space-frequency decomposition is presented in Table 1. Observing this decomposition, we remark that it yields an octave-band structure (4 octaves) which can be easily achieved using the computationally efficient multiresolution wavelet transform (WT).

3. SPEECH SOURCE LOCALIZATION

By combining the outputs of the microphones, the array becomes sensitive to a predetermined direction, θ, and its spatial response is:

$$y(\omega, n) = \sum_{i=1}^{M} \alpha(i)\, x_i(n) e^{-\jmath \frac{\omega}{c} d_i \sin\theta} \qquad (1)$$

where c is the sound velocity, $x_i(n)$ and d_i are the WT coefficients of the i^{th} microphone output and its distance from the reference microphone, respectively. α is the weighting vector and ω is the signal frequency.

The array beam can be algorithmically steered in different directions. The multi-beamforming operation is very useful for tracking a moving speaker [1, 2]. Unfortunately, the search for a speaker with this technique is computationally intensive because of the large number of beams and the consequent complexity of the discrimination function.

In this paper, we propose a new two-stage localization method to reduce the number of the formed beams [7]. First, the region containing the desired speaker is detected. Then, only three beams are formed in different directions within the selected region and with a predefined spatial increment to locate the speaker (see Fig. 2).

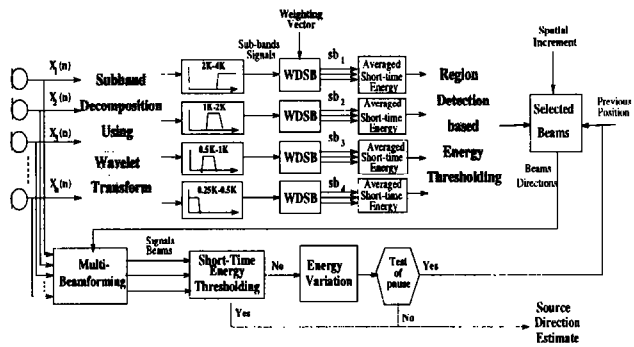

Figure 2. *Block diagram of the source localization and tracking algorithm.*

Both the region detection and the discrimination between these beams are based upon the concept of examining the energy contained in each frequency sub-band of the beams signals and its variation. Indeed, the speech signal is nonstationary. Thus, a great variation of its energy for successive blocks is noticed. This solution seems to be reasonable since the speaker is often closer to the acquisition system than the interfering sources and since these sources do not exhibit the same characteristics as the speech signal. In addition, to exploit the proposed decomposition achieved by wavelet transform, the process of source localization is applied in the time-frequency domain and the source localization accuracy is improved by combining the decisions provided by each sub-band beam signal. In this part, 5^{th} order Daubechies's prototype filter are used for the time-frequency decomposition.

3.1. Speaker's region detection

Several spatial responses can be obtained simultaneously by only manipulating the weighting vector α (see Eq. 1). The idea used here is to provide information about several regions in space. For this purpose, different weighting vectors, α_l, forming the lines of a fixed matrix, $\mathbf{A}$, are proposed. Here, each sub-array is composed of three microphones, yielding:

$$\mathbf{A} = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 0 & -0.5 \\ 0.5 & -0.5 & 0 \end{bmatrix} \qquad (2)$$

Each sub-array provides three beams $Y_l(n)$ ($l = 1, 2, 3$), corresponding to different regions in space. The power spectra on the octave scale $B_l = [B_l(1), B_l(2) \cdots, B_l(K)]$

is obtained by calculating the averaged short-time energy of L frames of the l^{th} beam as follows:

$$B_l(k) = \sum_{t=1}^{L} \sum_{n(k)} |Y_{l,t}(n)|^2 \quad k = 1, \cdots K \qquad (3)$$

where k is the band number and K is the total number of bands. The detection of the speaker's region is achieved by comparing the energy values of $B(k) = [B_1(k), B_2(k), B_3(k)]$ of the k^{th} octave. A threshold Tr is fixed with help of the input signal-to-noise (SNR) ratio.

$$\begin{cases} B_{max}(k) = max_l B(k) \\ B_{max}(k) \geq T_r \cdot B_{l_1}(k) \quad , \quad l_1 \in [1,2,3] \end{cases} \qquad (4)$$

where $B_{l_1}(k)$ is the 2^{nd} maximum value of $B(k)$. The l^{th} beam that verifies the two conditions given in Eq. 4 is selected as the region where the speaker is present. Since we have 4 octaves, 4 selections are obtained. The final decision is made with respect to the beam which has been selected more than 3 times. If two beams are selected the same number of times, a weight of 2 is attributed to the 2^{nd} octave ([250Hz, 500Hz]) which intervenes in the final decision. These regions are represented by $\phi = 0°, \pm 30°, \pm 60°$. We note that the distinction between $+\phi$ and $-\phi$ is based on the previous estimated speaker's location.

3.2. Speaker's localization

A predefined look direction ϕ, according to the estimated speaker's region, is determined. Three sub-array beams with ϕ and $\phi \pm \delta$ are formed. δ is the angular step chosen so that as these beams overlap sufficiently (e.g. $\delta = 15°$). We note that the wavelet decomposition is judged not necessary for this stage and the samples used here are the original samples. The short-time energy of the m^{th} beam signal y_m, is:

$$B_m(t) = \sum_{i=1}^{N} y_m(N(t-1)+i)^2 \qquad m = 1,\ 2,\ 3 \qquad (5)$$

where t and N are the index and the length of the frame. To obtain a large number of energy values, an overlapping of 75% between frames having a length of 256 samples. Given the duration of the observation of approximatively $45ms$, the normalized short-time energies of the three beams signals are computed. Then, for each frame, the maximum of the short-time energy, $B_m(t)$, is determined. Using two thresholds T_1 and T_2 where $T_1 > T_2$, the t^{th} frame will be selected to intervene in the decision if the following conditions are verified:

$$B_m(t) > T_1 \quad \text{and} \quad B_m(t) \geq T_2 \cdot B_{m_1}(t),$$

where $B_{m_1}(t)$ is the value that follows the maximum for the same frame. Then, the beam indexed by m, is a candidate. We note that T_2 allows to avoid the decisions in the pauses.

The operation is repeated for a predefined number of frames p. Finally, the beam which is selected more than $p/2$ times is chosen as candidate and it corresponds to the speaker's position. Otherwise, additional sets of p frames are called until the decision is taken. If the number of the selected frames is less than p, meaning that one of the conditions given above is not verified. This speech segment is assumed to be pause. In case of pause, the energy is relatively similar for all the beam signals. For confirmation, an energy variation is calculated and compared to a threshold T_p as follows:

$$H_m(t) = \frac{B_m(t)}{B_m(t-1)} \qquad H_m(t) \geq T_p \qquad (6)$$

T_p is the mean value of H_m for all the frames. If the speech segment is classified as pause, the estimated source direction corresponding to the previous segment is kept. Once the source position is found, the main beam is returned to the estimated location.

To pinpoint the speaker localization, a correct estimation of the speaker's region is required. For this purpose, the speech source and noise sources must be sufficiently separated.

4. NOISE REDUCTION

As shown in Fig. 3, the noise reduction algorithm is based on the conventional beamforming and Wiener filtering as done in [4, 3]. Here, a modified sub-band Wiener filtering process is developed and implemented in the time-frequency domain using a wavelet transform as shown in [8]. We note that the already existing noise reduction algorithms based on postfiltering have been developed in the time domain or using the Discrete Fourier transform (DFT). The major advantage of the wavelets is the significant reduction of the computational complexity without severely affecting performance. Indeed, the adopted wavelet transform is employed with lower-order filters (Haar filter) at lower rates providing a minimum delay in the filter banks.

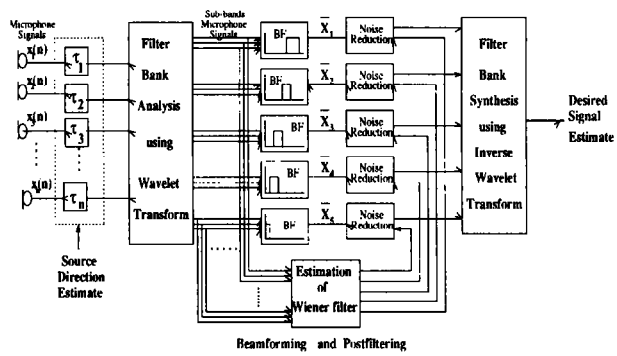

Figure 3. *Block diagram of the proposed noise reduction system.*

In the wavelet domain, the perfect diagonalization of the input autocorrelation matrix can not be expected. However, in most cases the high concentration of the significant coefficients around the diagonal can be pointed out. Thus, the autocorrelation matrix is approximated

by a diagonal matrix. This quite realistic assumption is done in order to obtain a reasonable compromise between the optimality of the Wiener filter and the computational load. By analogy to the transfer function of the Wiener filter proposed in Zelinski's and Simmer's methods [4, 3], the Wiener filter in the wavelet domain has the following expression:

$$W_k(n) = \frac{\frac{2}{M \cdot M-1} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} X_{i,k}(n) \cdot X_{j,k}(n)}{\frac{1}{M} \sum_{i=1}^{M} |X_{i,k}(n)|^2} \quad (7)$$

where $X_{i,k}(n)$ are the wavelet coefficients of the i^{th} microphone output signal in the k^{th} spectral sub-band.

5. EXPERIMENTAL RESULTS

A non-symmetric, logarithmically spaced array with 6 microphones and the smallest $d = 5cm$ is used. The array length is very suitable for communication terminals. The received signals has the telephone frequency bandwidth and are sampled with a sampling frequency equal to 48 kHz.

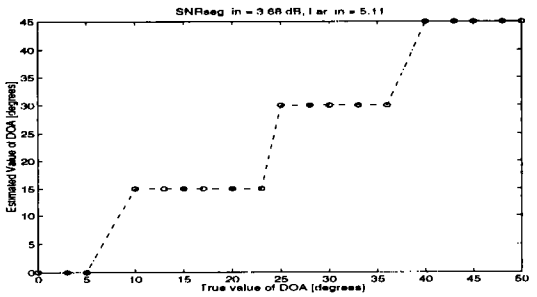

Figure 4. *Source localization results.*

To validate our approach for speech source localization, a number of simulations with different source position were performed, and the results are shown in Fig. 4. The estimated source direction is a multiple of 15°. The estimation error has an order of $0° - 7°$ which is acceptable comparing to the beamwidth. It was also found that the source localization procedure works correctly when the input SNR is larger than 8 dB. The new noise reductor algorithm was evaluated by performing simulations with segmental SNR decreasing down to 0 dB. This block gives a high attenuation of noise level (more than 15 dB of cancelation is achieved) and a good speech quality when a correct source localization estimate is found d. Fig. 5 shows the improvement in terms of Log-Area Ratio (LAR). We can also notice that the resulting speech signal is free from any musical noise.

6. CONCLUSIONS

In this paper, we have proposed a method based on a microphone array for tracking a moving speaker in an adverse environment, while enhancing the received signal. The method has the advantage to be simple. It exploits the spatial and the time-frequency decomposition of the array and the microphones signals, respectively, to solve two main problems: to obtain a sufficient directivity for the whole speech signal bandwidth and use efficiently of

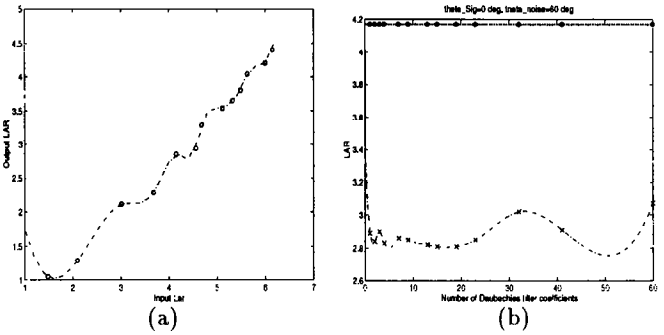

Figure 5. *Performance measure of the noise reduction system in terms of LAR (a) improvement of the system, (b) the gain for various Daubechies's prototype filter lengths.*

the energy criterion for the source localization procedure. In fact, we have noticed the good localization of the signal energy in the time-frequency plane. The postfiltering based on wavelets achieves a good noise reduction. In order to decrease the amount of computation, a critical subsampling with lower filter lengths are chosen for wavelet representation. Thus, the processing using the wavelets costs less than the one using the DFT.

Finally, we conclude that the usefulness of the proposed approach is proved by a good estimation of the speech source localization and a high attenuation of noise level. Further research should be conducted for the enhancement step to work under less restrictive assumptions like the diagonality of the input autocorrelation matrix.

REFERENCES

[1] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, and M. M. Sondhi, "Autodirective Microphone Systems", *Acustica*, vol. 73, pp. 58–71, 1991.

[2] W. Kellermann, "Self-Steering Digital Microphone Array", *in Proc. of ICASSP'91*, vol. 4, pp. 3581–3584, July 1991.

[3] R. Zelinski, "A Micropohone Array with Adaptive Postfiltering for Noise Reduction in Reverberant Rooms", *in Proc. of ICASSP'88*, pp. 2578–2581, 1988.

[4] K. U. Simmer and A. Wasiljeff, "Adaptive Microphone Arrays for Noise Suppression in the Frequency Domain", *in* 2^{nd} *Cost 229 Workshop on Adap. Algo. in Com., France*, pp. 185–194, Sep. 1992.

[5] O. L. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing", *Proceedings of IEEE*, vol. 60, pp. 916–935, 1972.

[6] H. F. Silverman and S. E. Kirtman, "A Two-stage Algorithm for Determining Talker Location from Linear Microphone Array Data", *Comp. Sp. and Lang.*, vol. 6, pp. 129–152, 1992.

[7] D. Mahmoudi, "Multiresolution Array Processing for Speech Source Tracking in Voice Communication Systems", *in Proc. of ICSPAT*, pp. 346–350, Oct. 1996.

[8] D. Mahmoudi, "A Microphone Array for Speech Enhancement using Multiresolution Wavelet Transform", to appear in Proc. of Eurospeech'97, Sept. 1997.