SENSING PEOPLE - LOCALIZATION WITH MICROPHONE ARRAYS

Peter Noll, Markus Schwab, and Wiryadi

Technical University Berlin,
Communication System Group
{noll|schwab|wiryadi@nue.tu-berlin.de}

Abstract:

In this paper we present a real-time microphone array system which performs 3D source localization, multi channel speech enhancement and robust speech recognition. The acoustic source localization uses the SRP-PHAT method [1] to produce potential source locations. A clustering algorithm excludes outliers and enables a multi source tracking. The localizations are finally optimally filtered with an appropriate Kalman filter. The proposed speech enhancement, a weighted subarray Delay-and-Sum beamformer, is designed to cope with the problem of diffuse noise and changing speaker positions subject to minimization of the word error rate (WER) of an automatic speech recognition system (ASR). The proposed algorithm reduces the WER by more than 50 % compared to the WER of a single microphone signal.

1 Introduction

"Sensing People" [2] [3] denotes the description and analysis of persons in digital images and from acoustical signals with respect to their state, their activity, their identity and many other aspects. Intelligent cameras and sensors integrate sensing people functionalities for multimedia applications and industry electronics. They automatically extract personal data from sensor signals for many applications. For this purpose techniques are developed which allow the localization and the description of persons in video data from CCD-cameras, CMOS-cameras and in audio signals from microphone arrays. Microphone arrays yield information about the direction of acoustic sources. CMOS-cameras are used to capture the 3D profiles of a scene. All the data are integrated and further analyzed within a sensor fusion module. Figure 1 visualizes the concept of "Sensing People".

Acoustic source localization based on microphone arrays has been a mainstream research topic for over two decades. The solutions available in the literature can be coarsely classified into three broad categories [4]: those based on maximizing the steered response power (SRP) of a beamformer; those based on High-Resolution Spectral Estimation (HRSE) methods; and those based on Time-Difference Of Arrival (TDOA) algorithms. Steered Beamforming is a well-known method for deriving information on the source locations directly from a filtered linear combination of the acquired signals. Methods based on HRSE imply the analysis of the correlations between the acquired signals, while TDOA methods extract information on the source location through the analysis of a set of delay estimates.

Methods based on TDOA algorithms have two steps. First, they estimate the TDOAs. The most popular method for TDOA estimation is the cross correlation approach. A more robust estimation can be achieved with the generalized cross correlation GCC (e.g. GCC-PHAT) [5]. A more recent approach is the so called Adaptive Eigenvalue Decomposition (AED) algorithm [6]. The AED algorithm performs better than the GCC-PHAT algorithm in highly reverberant

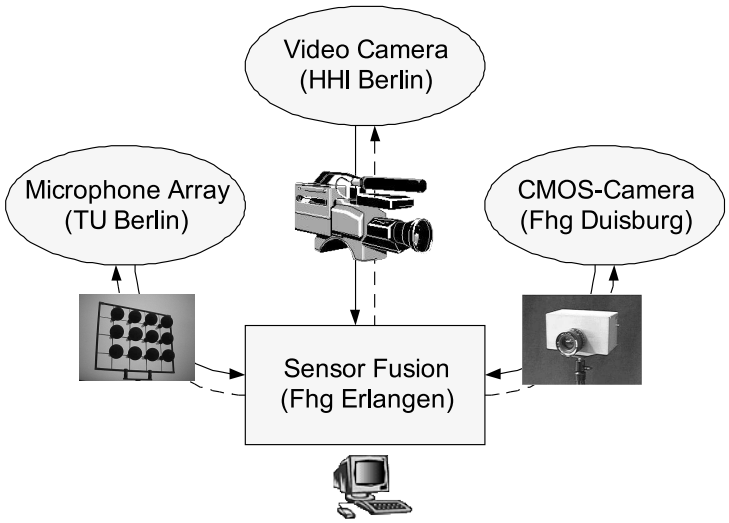


Figure 1 - Sensing People Concept and Partners

environments. The second step is to estimate the position of the source with the knowledge of the TDOAs.

Source localization methods of the second category are all based on the analysis of the spatial covariance matrix (SCM) of the array sensor signals. The SCM is usually unknown and needs to be estimated from the acquired data. Such solutions rely on high resolution spectral estimation techniques [7]. Popular algorithms based on HRSE are Minimum Variance Beamformer (MVB) [8] and MUltiple SIgnal Classification (MUSIC) algorithms [9]. These algorithms can be extended to wideband signals, e.g. speech, by transforming the signal into narrowband signal. Each narrow band signal can be processed individually (incoherent method) or an universal focusing SCM [10] can be generated to perform a coherent localization.

Roughly speaking, the optimal maximum likelihood (ML) SRP-based localization methods rely on a focused beamformer, which steers the array to various locations in space, and look for peaks in the detected output power [11]. In its simplest implementation, the steered response can be obtained through a Delay-and-Sum process performed on the signals acquired by two microphones. One of the two signals is delayed in order to compensate for the propagation delays due to the incidence direction of sounds. SRP based localization methods are very robust especially the newly proposed SRP-PHAT method [1].

In this paper we will present a real-time microphone array system which performs a localization with the SRP-PHAT method, Kalman filtering and clustering of the obtained positions, multi channel speech enhancement algorithm and speech recognition. Results will be given in a speaker tracking example, recognition error rates of the newly developed speech enhancement algorithm and the traditional Delay-and-Sum beamformer will be compared.

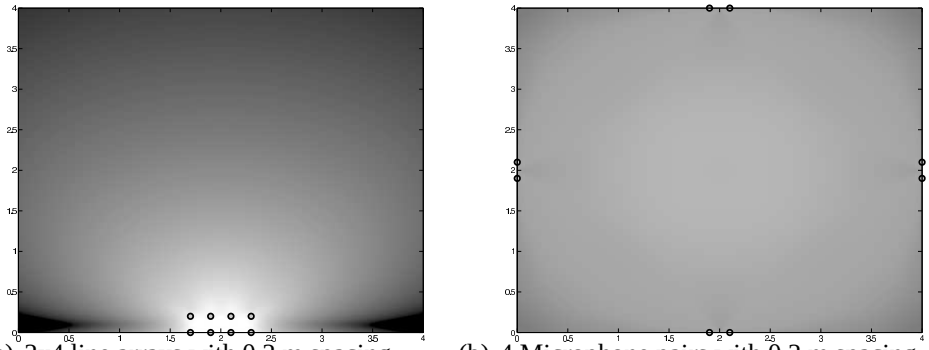
2 Sensor placement

Sensor placement is the first issue in the design of a microphone array. Acoustic source localization algorithms generally exploit the coherence between the acquired sensor signals. Thus, we will group the microphones into microphone pairs whose correlation will be used for the localization. The distance between microphones grouped to a pair should be small because the coherence of the microphone signals decreases with the distance between the sensor positions [12], [13]. For a TDOA based maximum likelihood localization algorithm, Brandstein et al. derived a predictor for the covariance of the expected localization error Δp depending on the

variances of the TDOA estimates, the source location, and the microphone positions:

$$cov(\Delta p) = (\mathbf{H}^T \mathbf{V} \mathbf{H})^{-1} \tag{1}$$

where $\mathbf{V}$ is a diagonal (N,N) matrix (N is the number of the defined microphone pairs). The diagonal elements are the reciprocal variances of the TDOA estimates $w_{ii} = \frac{1}{\sigma_i^2}$. $\mathbf{H}$ is a $(N, 3)$ matrix which relates the error of the localization Δp to the errors in the TDOAs $\Delta \tau$:
$\Delta \tau = \mathbf{H} \cdot \Delta p$



(a) 2x4 line arrays with 0.2 m spacing (b) 4 Microphone pairs with 0.2 m spacing

Figure 2 - Log-scaled predicted localization error, (a) one compact 8 microphone array with a spacing of 0.2 m between its microphones, (b) array with four microphone pairs each with a spacing of 0.2 m between its microphones)

In figure 2 the predicted location errors are depicted as two different array geometries. In the left figure the array geometry is a compact 8 microphone array and in the right figure the array geometry is more sparsely, i.e. a microphone pair is placed at each border of the sensed area. It is assumed that the variances for the TDOA estimation are only affected by the signal to noise ratio (SNR) at the microphones. Since the spacing between the microphones of a pair is small the same SNR can be assumed for both microphones.Therefore, the variances of the TDOA estimates are reciprocally proportional to the square of the distance r_i between the microphone pair and the sound source $\phi_i \propto \frac{1}{r_i^2}$.

The brightness in figure 2 is related to the predicted error at a given position. A brighter dot indicates a smaller predicted error. Figure 2 (a) shows that a precise localization is only possible close to the microphone array whereas in figure 2 (b) the predicted localization error is better balanced and a localization is feasible for all positions in the sensed area [4x4 m^2]. This 2D representation can easily expanded to 3D and the result will be the similar. The conclusion from this example is that the microphone pairs have to be widely distributed in the sensed area to be able to perform a 3D source localization.

3 System Overview

The audio system consists of three steps. First, the speaker is localized by the microphone array. For this purpose we use the SRP-PHAT algorithm [1]. An appropriate Kalman filtering improves the tracking results of a moving speaker. Second, enhancement of the noisy speech. The microphone signals are then time aligned and averaged according to the estimated speaker position and the microphone positions. After this Delay-and-Sum beamforming the signal is further enhanced by one-channel postfiltering. Third, an automatic speech recognition (ASR) system evaluates the enhanced speech and recognizes words.

The signal processing is efficiently done in the frequency domain. Let $x_i(n)$ denote the i^{th} time discrete microphone signal. The observed signals are then segmented into overlapping frames and zero padding is used to avoid the effect of circular convolution in the beamformer. Then

these frames are transformed into the frequency domain $X_i(k,l)$ using the FFT, where k and l are the frequency and time index, respectively.

4 Speaker Localization

For speaker localization, suitable microphone pairs are selected to contribute to the localization. For each microphone pair the normalized cross power spectrum is calculated and then transformed into the time domain to obtain the generalized cross-correlation with phase transform (GCC-PHAT) for each microphone pair:

$$R^{ij}_{GCC-PHAT}(n,l) = \frac{1}{N}\sum_{k=0}^{N-1}\Big(\frac{X_i(k,l)\cdot X_j^*(k,l)}{|X_i(k,l)\cdot X_j^*(k,l)|}\Big)e^{j\frac{2\pi nk}{N}} \ . \tag{2}$$

N is the frame length and n is the time shift in the generalized cross correlation. In the following we will drop the time index l for a more convenient representation. With the GCC-PHAT sequences from all used microphone pairs we can define a cost function depending on a virtual speaker position S:

$$P(S) = \sum_{pairs(i,j)} R^{i,j}_{GCC-PHAT}(\delta(S,i,j)) \ , \tag{3}$$

where $\delta(S,i,j)$ denotes the index corresponding to the time delay for the microphone pair i,j if a virtual sound source is located at position S. If M_i and M_j are the microphone positions, δ can be calculated as follows:

$$\delta(S,i,j) = \frac{|S-M_i| - |S-M_j|}{c} \tag{4}$$

where c represents the velocity of sound in air.

The speaker position is then estimated by maximizing of cost function $P(S)$ over the sensed area.

To improve the quality of the estimated speaker position we have implemented a Kalman filter. Only the x and y coordinates are Kalman filtered because we have implied that the changes in the z direction are small so that recursive smoothing for the z-coordinate is sufficient.

The motion model uses a simple Newtonian model with velocity and acceleration. The speaker state vector $\theta(l)$ consists of 6 elements:

$$\theta(l) = [x(l) \quad y(l) \quad \dot{x}(l) \quad \dot{y}(l) \quad \ddot{x}(l) \quad \ddot{y}(l)] \ . \tag{5}$$

The process noise is modeled by uncorrelated zero-mean noise. The covariance of the measurement noise is calculated as a function of source location, sensor positions, and the generalized cross correlations of the microphone pairs [13].

4.1 Beamformer

We use a Delay-and-Sum beamformer whose output can be computed in the frequency domain by using a complex steering vector:

$$\mathbf{W}^H(k) = (e^{j2\pi k/f_s\tau_1}, e^{j2\pi k/f_s\tau_2}, \ldots, e^{j2\pi k/f_s\tau_M}) \tag{6}$$

with $\tau_i = \frac{|S-M_i|}{c}$, $(.)^H$ denotes the hermitian transpose.

The output of the delay-and-sum beamformer in the frequency domain is then calculated by:

$$B(k,l) = \mathbf{W}^H(k,l)\cdot\mathbf{X}(k,l) \tag{7}$$

with $\mathbf{X}(k,l) = (X_1(k,l), X_2(k,l), \ldots, X_N(k,l))$.
The quality of the speech signal (e.g. SNR as a measurement) can differ significantly at different microphones due to their differing locations. Therefore, we regroup the microphones into subarrays where we assume that the speech quality for each microphone signal within a subarray is equal. The output of the subarray beamformer is then weighted according to the SNR. The SNR weighting factors are normalized so that the sum of the weighting factors is one. For the SNR estimation we use the a priori SNR SNR_{prior}, see section 4.2. The weighted outputs of the subarrays are then time aligned and added up to give the output of the weighted subarray delay-and-sum beamforming (WS-DSB). Figure 3 illustrates the structure of the WS-DSB. Shown are three subarrays, each with four microphones.

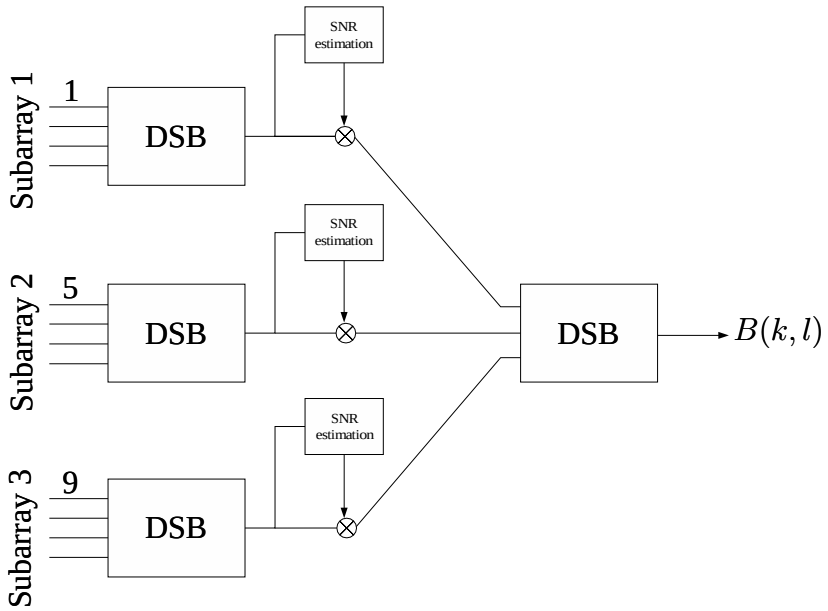


Figure 3 - Weighted subarray delay-and-sum beamformer (WS-DSB) with three subarrays, each with four microphones.

4.2 One-Channel Postfilter

The output of the beamformer $B(k,l)$ is supplied to a one-channel postfilter whose structure is shown in figure 4. Postfilter can efficiently suppress stationary background noise by estimating the power spectrum of the background noise and then optimally modifying the noisy speech spectra according to statistical methods.

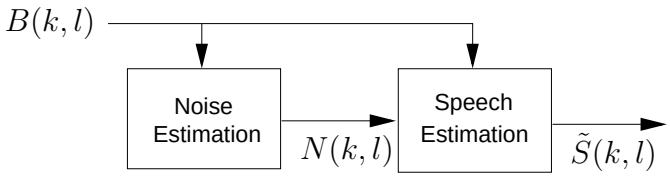


Figure 4 - Structure of one-channel postfi lter

A simple and robust noise estimation is realized by minima controlled recursive averaging (MCRA) as proposed by Cohen and Berdugo [14]. This algorithm takes advantage of the simplicity of recursive averaging and of the robustness of the minimum based noise estimation algorithm [15]. The noise estimate is given by averaging past spectral power values and using a

smoothing parameter that is adjusted by the signal presence probability in subbands. Presence of speech in subbands is determined by the ratio between the local energy of the noisy speech and its minimum within a specified time window.

The speech estimation follows the decision directed method and uses the log-spectral amplitude estimation rule as given by Ephraim and Malah [16]. The gain factor $G_{LSA}(k,l)$ is derived by:

$$G_{LSA}(k,l) = \frac{SNR_{prior}(k,l)}{1+SNR_{prior}(k,l)} \exp\Big(0.5 \int_{t=\nu(k,l)}^{\infty} \frac{e^{-t}}{t} dt\Big), \tag{8}$$

where the a priori SNR $SNR_{prior}(k,l)$, and $\nu(k,l)$ are defined as in [16].

Finally, the enhanced spectral components are obtained by

$$\tilde{S}(k,l) = G_{LSA}(k,l) B(k,l) \ . \tag{9}$$

5 Experimental Setup and Results

Figure 5 shows an overhead view of the experimental setup. The room has a size of 10x6 m^2, its height is 4 m and the room reverberation is moderate. We used 12 microphones subdivided into 3 subarrays with 4 microphones each, see figure 5. The microphones are marked with bold dots. For the subarrays 2 and 3 only two microphones can be seen in the overhead view since the four microphones in these subarrays are arranged in vertical rectangle. A result of a speaker tracking is shown. The small dots depict the true position of the speaker, hand labeled with a camera and a grid on the floor. The crosses mark the estimated position by the SRP-PHAT method and the line represents the improved results after Kalman filtering.

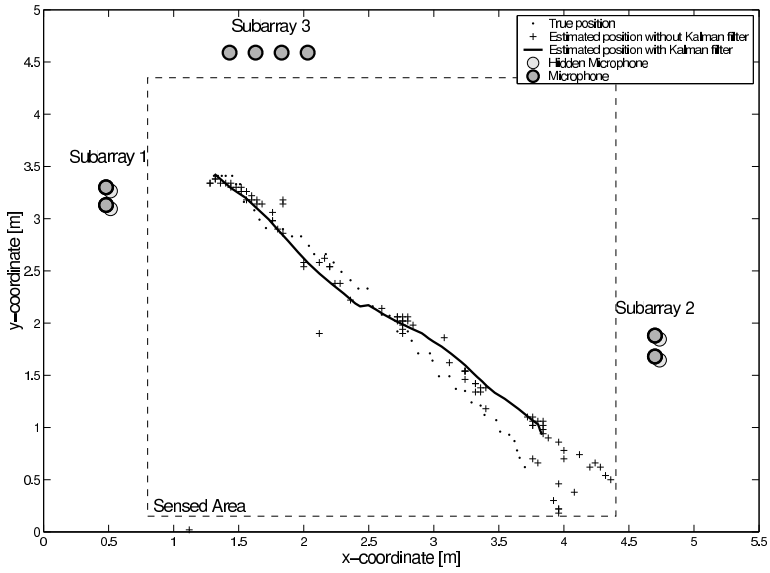


Figure 5 - Overhead view of the scenario with a speaker tracking example

In the next step, the described microphone array setup was used to obtain real multichannel recordings as the test database. The speakers were located in the sensed area (dashed rectangle in figure 5). After each sentence the speaker's position was changed. The recorded signals were corrupted by ambient noise such as computer fans and air conditioning. The signal-to-noise ratio varied from 6 dB to 13 dB depending on the channel and the speaker position. The test database consisted of 1274 sentences, containing 5 german digits per sentence spoken by eleven different speakers. The training database consisted of 600 sentences again containing 5 german digits per sentence but spoken by six different speakers. The recordings for the training database

were made with a close-talk microphone so that there is a mismatch between the training and test database.

The noise reduced power spectra $\tilde{S}(k,l)$ from equation 9 were used to calculate twelve standard mel frequency cepstral coefficients (MFCC) in a frequency range between 64 Hz and 8000 Hz. These twelve MFCC and the logarithmic frame energy together with the corresponding delta and acceleration coefficients resulted in a feature vector of dimension 39. The speech recognition was based on a whole word hidden Markov model (HMM). Each word was modeled by a HMM with 16 states and three gaussian mixtures per state. Two pause models, "sil" and "sp", were used which represented the pauses at the beginning/end and the pauses between the words, respectively.

A summary of word error results is given in table 1. Subarray 1, subarray 2, and subarray 3 are the outputs of a conventional Delay-and-Sum beamformer (DSB) using the microphone signals in a subarray. The overall DSB applies a conventional Delay-and-Sum beamformer using all microphone signals. The weighted subarray delay-and-sum beamformer (WS-DSB) is the algorithm proposed in section 4.1. The one-channel postfilter for the WS-DSB was applied to each subarray output.

	without postfilter WER [%]	with postfilter WER [%]
Subarray 1	24.07	19.94
Subarray 2	17.16	14.51
Subarray 3	18.51	14.39
overall DSB	16.28	14.09
WS-DSB	14.62	12.70

Table 1 - Word error rates (WER) in % for the three subarray DSB outputs, the overall DSB and the weighted subarray (WS)-DSB.

The word error rates for the unprocessed microphone signals 1, 5 and 9, belonging to the subarrays 1, 2 and 3, respectively, are shown in table 2. These word error rates serve as a baseline.

	Mic 1	Mic 5	Mic 9
WER [%]	29.40	25.42	28.67

Table 2 - Word error rates (WER) in % for three single microphone recordings

6 Conclusions

In this paper we have presented a real-time microphone array system for 3D speaker localization, speech enhancement, and speech recognition. The ability to track a speaker position is illustrated with an example. We have proposed a speech enhancement scheme based on a subarray structure. In order to measure the performance of the system we have compared the speech recognition results of several speech enhancement algorithms. The WS-DSB algorithm reduces the word error rate from the best microphone signal (5) from 25.42% to 12.70%. We also have shown that the proposed weighted subarray DSB performs better than the overall DSB. The WER can be improved by training the environmental conditions. This would lower the flexibility of the system since it has been trained to this specific acoustic conditions.

Literature

[1] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone Arrays, Signal Processing Techniques and Applications*, chapter 8, Springer Verlag, 2001.

[2] Fraunhofer HHI, "Sensing people homepage,"
URL: http://www.cebit2003.fraunhofer.de/servlet/is/4187/, 2003.

[3] P. Noll, "Microphone arrays in a sensing people project," 5th International Workshop on Microphone Array Systems - Theory and Practice, May 2003.

[4] M.S. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer, 2001.

[5] Charles. H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-24, no. 4, August 1976.

[6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *JASA*, vol. 107(1), pp. 384–391, January 2000.

[7] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice Hall, 1997.

[8] H. Krim and M. Viberg, "Two decades of array signal processing research. the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.

[9] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," in *IEEE Trans. Antennas Propagat.*, March 1986, vol. AP-34, pp. 276–280.

[10] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Aug. 1985, vol. ASSP–33, pp. 823–831.

[11] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," in *IEEE Trans. Information Theory*, Sept. 1973, vol. IT-19, pp. 608–614.

[12] M. Drews, *Mikrofonarrays und mehrkanalige Signalverarbeitung zur Verbesserung gestörter Sprache*, Ph.D. thesis, Technische Universität Berlin, 1999.

[13] M.S. Brandstein, J. E. Adcock, and H. F. Silverman, "Microphone array localization error estimation with application to optimal sensor placement," in *J. Acoust. Soc. Am.*, 1996, vol. 29(2), pp. 3807–3816.

[14] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," in *IEEE Signal Processing Letters*, January 2002, vol. 9, pp. 12–15.

[15] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. 7th EUSIPCO*, Sept. 1994, pp. 1182–1185.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE Trans. Acoust., Speech, Signal Processing*, April 1985, vol. ASSP-33, pp. 443–445.