Applied Acoustics 73 (2012) 132–143

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Design and implementation of a MEMS microphone array system for real-time speech acquisition

Ines Hafizovic [a,b,*], Carl-Inge Colombo Nilsen [b], Morgan Kjølerbakken [a], Vibeke Jahr [a]

[a] Squarehead Technology AS, Gullhaug Torg 3, 0484 Oslo, Norway
[b] University of Oslo, Department of Informatics, Gaustadalleen 23B, 0316 Oslo, Norway

ARTICLE INFO

Article history:
Received 30 July 2009
Received in revised form 25 July 2011
Accepted 26 July 2011
Available online 24 August 2011

Keywords:
Microphone array
Speech acquisition
Signal processing

ABSTRACT

Despite many attractive features and the potential for capturing sound in challenging acoustic environments, arrays with a large number of microphones have for a long time been discarded as a practical solution for speech acquisition. This is, among other reasons, due to the high production and computational costs. Only a few realizations of large microphone array systems have been documented, mainly for research and instrumentation use. The advent of MEMS microphones and computationally powerful off-the-shelf hardware has created new possibilities for microphone array development. We investigate a real life application, specifically the case of live sports broadcast, and the requirements that a such application imposes on a microphone array system. We present a system architecture of the first large (300 element circular array with a diameter of 2 m) MEMS microphone array system. In the proposed system, the latest technological advances are utilized to create a user-friendly array control interface. The array's performance is examined in an anechoic chamber and on a crowded basketball field, and finally compared with existing solutions. The results illustrate the potential of a large MEMS microphone array as part of the technological development in sound acquisition for entertainment and security applications.

1. Introduction

The problem of remote speech and audio acquisition in noisy environments is a challenging one when body-worn microphones are not an option. A representative situation is audio acquisition in live sports broadcasting. The region of interest is usually a large field onto which no recording equipment can be placed. The athletes themselves can, for practical reasons, not be equipped with microphones. Local coverage of certain areas close to the edges of the field can be achieved by distributing single shotgun microphones, or by using parabolic reflectors. These give, however, only sporadic coverage. The major restriction is fixed spatial selectivity in post-processing, meaning that the audio missed in real time due to misaiming is lost forever. An array of microphones is electronically steerable, can provide multiple beams from different directions simultaneously, and in addition offers the same functionality in post processing as in real time. Furthermore, arrays can be designed with a much better directivity than what is attainable with the best shotgun microphones. These features are very attractive, but still not fully exploited in the field of audio acquisition. An appreciable amount of research on sensor arrays has been done over the past decades, and the techniques are well documented in [1]. Array technology plays a leading part in applications like telecommunication, sonar, and medical ultrasound imaging. In speech acquisition applications, small microphone arrays used at close proximity are predominant. The research topics that have gained the most interest here are speech enhancement for teleconferencing [2], hands-free telephony in cars [3,4], speech recognition [5], and hearing aids [6]. A detailed review of microphone array techniques and applications is given in [7].

A solution for remote sound acquisition in large rooms with a large number of microphones (distributed around the room) is described in [8,9], and a research project documenting a large microphone array (1024 elements) for speech recognition is reported in [10]. Also, in these research projects concerning large microphone arrays, most of the attention is given to automated speech recognition, blind source separation and speaker tracking methods. For a large array with hundreds of microphones, these methods are computationally demanding and not suitable for the problem of real time speech acquisition where multiple speakers are present but only a few are of interest, like in the case of sports broadcasting.

* Corresponding author at: Squarehead Technology AS, Gullhaug Torg 3, 0484 Oslo, Norway. Tel.: +47 93038365.

E-mail addresses: ines@sqhead.com (I. Hafizovic), carlingn@ifi.uio.no (C.-I.C. Nilsen), morgan@sqhead.com (M. Kjølerbakken), vibeke@sqhead.com (V. Jahr).

0003-682X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2011.07.009

The use of microphone arrays in the entertainment industry has been suggested by Silverman et al. [11], but to our knowledge there have been no consumer-oriented large microphone array systems designed for a specific real-life application. Sound acquisition in real acoustical environments where traditional methods and equipment fail to perform is an important problem to solve and we have investigated the feasibility of a large (300 elements), two-dimensional microphone array as a solution to this problem. In this paper, we propose a user-friendly microphone array system, incorporating state-of-the-art MEMS microphones and off-the-shelf processing technology. We present the premises for the system design, its theoretical performance, and measurements performed in an anechoic chamber and at a live sports event. A brief description of prior work done by the authors on microphone arrays can be found in [12].

In Section 2.1, we explain the basics of the relevant array theory, and Section 2.2 outlines the environment for which the array is designed and the restrictions that must be considered. Sections 3 and 4 present the design process, and the array prototype with its theoretical capabilities. The measurements performed on the array, in ideal and real-life environments are shown in Section 5. The results are compared to the theoretical expressions, and explanations for the deviations are suggested. A conclusion regarding the applicability of microphone arrays for the above mentioned applications based on these results is presented in Section 6.

2. Background

2.1. Sensor arrays and beamforming

The term beamformer is generally reserved for the algorithm used to combine signals from multiple sensors into one or more outputs. A simple but robust beamformer is the delay-and-sum (DS) beamformer. As the name of the beamformer indicates, the sensor outputs are delayed to be in phase (steered) for a given direction, and then summed and averaged. The steering direction, which will be denoted by $\vec{p}_s$, usually coincides with the direction of arrival for the sound of interest, e.g. a speaker.

The output $y[n]$ of a DS beamformer at time instant n, applied to an array comprised of M elements (i.e. microphones) placed at positions $\vec{p}_m,\ m = 0, 1, \ldots, M-1$, and steered towards a source located at a point $\vec{p}_s$, can be described as a weighted sum of sensor outputs:

$$y[n] = \sum_{m=0}^{M-1} w_m x_m[n - \Delta_m], \tag{1}$$

where x_m is signal at sensor m sampled at rate f_s and w_m is the weight applied to the mth sensor. Δ_m is the delay (in number of samples for microphone sampling rate f_s) for the mth microphone when steering in the direction of $\vec{p}_s$, and is given as:

$$\Delta_m = \left\lfloor \frac{|\vec{p}_m - \vec{p}_s| f_s}{c} \right\rceil, \tag{2}$$

where c is the speed of sound, which is assumed to be 340 m/s. Eq. (2) gives the microphone delays in terms of the number of samples, which is subsequently rounded to the nearest integer multiple of the sampling period. This rounding gives a delay error that can manifest itself as a steering error. For the error to be negligible, the microphone data should be sampled at a rate f_s that is at least 10 times higher than the highest frequency we wish to beamform. This is just a rule of thumb, more details on the effect of f_s on the beamformer performance are given in e.g. [13]. One way of reducing delay errors is to increase the sampling rate f_s by interpolation prior to beamforming.

When the signal from element m is delayed by the number of samples described by Eq. (2), signals coming from the position $\vec{p}_s$ are temporally aligned, and we say that the array (the beam) is steered towards $\vec{p}_s$. By summing across elements, we can enhance the signal of interest and simultaneously filter out noise from directions $\vec{p}_i \neq \vec{p}_s$ with a spatial FIR-filter with a continuous time impulse response:

$$\begin{aligned} h_m(t) &= \sum_{m=0}^{M-1} w_m \delta\left(n - \Delta_m + \Delta_m^i\right), \\ \Delta_m^i &= \frac{|\vec{p}_m - \vec{p}_i| f_s}{c}, \end{aligned} \tag{3}$$

The weights w_m are either set to uniform weighting given by $w_m = \frac{1}{M},\ m = 0, 1, \ldots, M-1$, or given by some weighting function. There is a vast amount of weighting functions to choose from, e.g. Chebyshev, Taylor, Hamming, etc. The choice depends on the desired response, e.g. on the required side-lobe level or the main-lobe width. Regardless of the choice of weighting function, we will in further discussion assume that

$$\sum_{m=0}^{M-1} w_m = 1, \tag{4}$$

meaning that weights are normalized and therefore yield a distortionless response. In other words, any signal coming from the steering direction $\vec{p}_s$ will be unaffected by the array if Eq. (4) is satisfied.

The directivity of the array is often evaluated by investigating the *broadband beampattern*. This the array's response to a sine wave of frequency f approaching from an angle θ, when the array is steered to an angle of 0° plotted as a function of f and θ. The broadband beampattern shows what attenuation can be expected for a signal of a certain frequency, arriving from a certain angle. The resolution of the array is often determined from the beampattern by looking at the main lobe width, more specifically the distance between the angles for which the array yields 3 dB attenuation. This is commonly referred to as the half-power beamwidth (HPBW). An example beampattern, for a single frequency, is shown in Fig. 1, where side lobes, main lobe, and half power beamwidth are marked. To avoid strong contributions from directions away from $\vec{p}_s$, the array must satisfy the spatial sampling criterion saying that the distance between elements must be smaller than half the wavelength of the highest frequency received by the array. If this is not satisfied, the spatial aliasing will occur and manifest itself as *grating lobes* in the beampattern. A grating lobe is a sidelobe that is equally high as the main lobe.

2.2. Problem analysis

In this section we outline the nature of a selected problem (basketball court) and the requirements it imposes on the array performance. A basketball court scenario is considered to be a representative of other challenging areas of application (e.g. security and surveillance) where we have to deal with multiple speech sources, some considered as signals and other as noise.

Fig. 2 shows the dimensions of a basketball court and the approximate distribution of noise and signals. The sound from the audience will most of the time be considered as noise, while the players at the court are considered as speech sources of interest. Noise (shouting from the audience, music, and announcements from a PA system) normally has a higher level than the game. Noise levels are time varying, and typically highest when important action is taking place on the court, but the relative SINR (signal-to-interference-and-noise ratio) is nearly constant since the players adjust their vocals to compensate for the increased noise. Based on the measured sound pressure level (SPL) of 95 dBA during