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 side-lobe 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

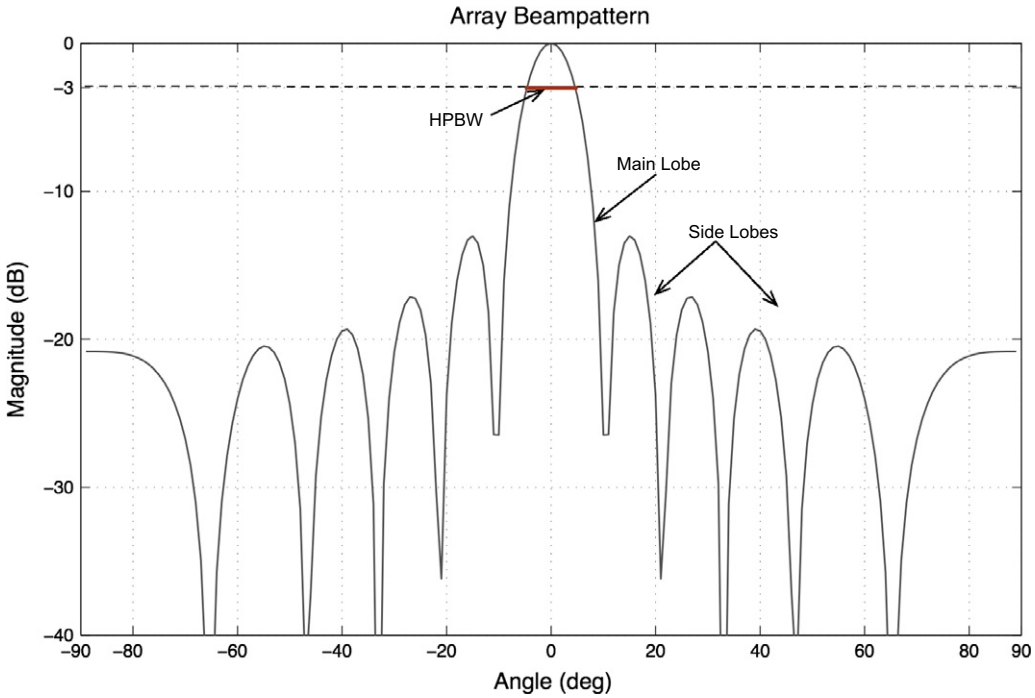

Fig. 1. An example beampattern for a single frequency. The corresponding array has $M = 11$ elements that are uniformly spaced at half-wavelength distance.

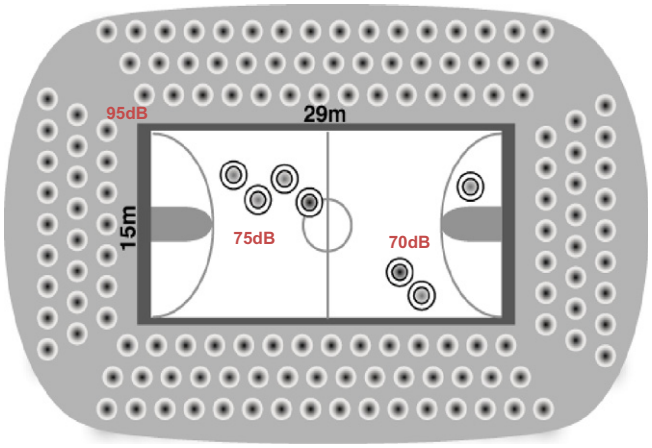

Fig. 2. Layout of a basketball field indicating typical distances and sound pressure levels.

cheering, and spatial SINR (players to audience ratio) of approximately −20 dB, the array should be designed with the average sidelobe level at least 20 dB below the level of the main lobe. In addition, a good spatial coverage (i.e. a large maximum steering angle) without significant changes in response is desirable. With a two-dimensional array the best coverage of the field is achieved if the array is placed parallel to the floor, at some suitable height.

For the sake of generality, we assume a rotationally invariant array response. To cover a basketball court, the array must cover approximately 15 m from the array center in all directions. Table 1 shows the radial coverage for some relevant heights and maximum steering angles. The maximum useful steering angle of the array will be given by the array's geometry and the required spatial resolution. In general, an increase in steering angle relative to the main response axis (towards the seating area in Fig. 2) will broaden the beam and at some point reach the scan limit. The scan limit is reached when the half-power attenuation disappears on one side

Table 1
Radial field coverage (meters rel. array axis). h = height (m), θ = steering angle (°).

	$h = 6$	$h = 8$	$h = 10$	$h = 12$	$h = 14$
$\theta = 40$	5	7	8	10	12
$\theta = 50$	7	10	12	14	17
$\theta = 60$	10	14	17	21	24

of the beam, beamwidth vs. steering angle is detailed in [14]. From Table 1, we see that we need useful steering angles between 50° and 60° to cover the basketball field. Since certain parts of the audience and the players can be spaced by only a few meters, the array must have good spatial resolution. For a spatial resolution of e.g. 1 m, an opening angle (half-power beamwidth) of 12° is required of an array operating from a height of 8–10 m.

When the array is operating from large distances, the sound pressure loss of 6 dB per distance doubling must be taken into account. For example, if the sound pressure level (SPL) of a source is 70 dB at 1 m distance, and the array is 20 m away from the source, the loss at the array will be $SPL_{loss} \approx 26$ dB (assumed point sources and free field propagation). The array gain should ideally compensate for this loss, if it is to be compared with the SNR of a hand-held microphone. The maximum array gain against noise that is uncorrelated across the microphones (e.g. sensor self-noise) is given in decibels by $10\log_{10}(M)$. A large number of microphones would be required to compensate for the distance loss alone, i.e. $M = 10^{0.1 \cdot SPL_{loss}}$ which corresponds to 398 microphones for the example of $SPL_{loss} \approx 26$ dB. In addition, in a real acoustic environment, with reverberations, non-isotropic noise field, and not true point sources, the attainable gain is reduced and the speech quality attainable with an array will not be as good as with a high-quality hand-held microphone. This is however not a realistic basis for comparison, since the common way of acquiring sound from a basketball field is by placing multiple shotgun microphones at the sideline. The performance of the set of distributed shotgun microphones then provides the actual benchmark with which to compare the speech enhancement capability of the microphone array.

3. Microphone array design

There are many choices for the design of a microphone array system and in this section we describe basic theoretical requirements and some of the practical limitations that we have to consider when dealing with speech acquisition in sports arenas.

3.1. Microphone array geometry

The number of sensors in the array and their spacing will be determined by the following requirements:

- Speech signals must be captured from large distances with good SNR.
- Intelligible speech signals occupy the frequency band from 300 Hz to 3500 Hz, and in a speech acquisition application the array should ideally have frequency-invariant response within the entire band in order to match the quality of the existing solutions.
- The array has to satisfy the spatial sampling criterion at all frequencies in the band of interest.
- The array must have a narrow mainlobe for good spatial resolution and low sidelobes for adequate noise attenuation, as discussed in the previous section.

In general, for a finite number of sensors, array patterns with low sidelobe levels have broader mainlobe. Due to this fact, fulfilling the listed requirements simultaneously for broadband signals, will require a potentially large number of sensors. This may in turn lead to an impractically large array (diameter of several meters).

In our design the initial requirement was portability of the array and hence the design limit on the maximum dimension of the two-dimensional array was approximately 2 m. With this in mind, microphones were exponentially spaced along one line for a better response to broadband signals [15,16], and also to avoid grating lobes that are a problem for broadband arrays with periodically spaced elements. Rotations of one line were then combined into a circular array for a rotationally invariant spatial response, resulting in an array comprised of 300 microphones and diameter of 2.1 m, as shown in Fig. 3D.

In order to even the spatial response across the frequency bands the array is divided in three sub-arrays. The subarray division is done by evaluating Eq. (3) for multiple configurations of the elements in the full array (2.1 m diameter), and by selecting the configurations giving the least frequency dependent responses. The final sub-array configurations are shown in Fig. 3A–C. Each sub-array covers a smaller frequency range while the sum of the bandpass-filtered sub-array outputs covers all frequencies of interest. The bandwidths of the sub-arrays are: [350,1700] Hz, [1700,2500] Hz, and [2500,f_{max}] Hz. The default value of f_{max} is 3.5 kHz, but can be extended if a slightly higher side-lobe level is acceptable. The sub-band filters are 96th order Hamming-window based, linear-phase filters, designed for a sampling rate of 33,075 Hz using Matlab. Filters are designed to have linear phase and flat frequency response in the pass-band. The combined filter response (sum of the sub-array filters) is shown in Fig. 4. Since the filters are digital, they can easily be changed to any desired response.

The theoretical spatial response of the array is shown in Fig. 5, with level curves illustrating the −3, −10, and −20 dB beamwidths.

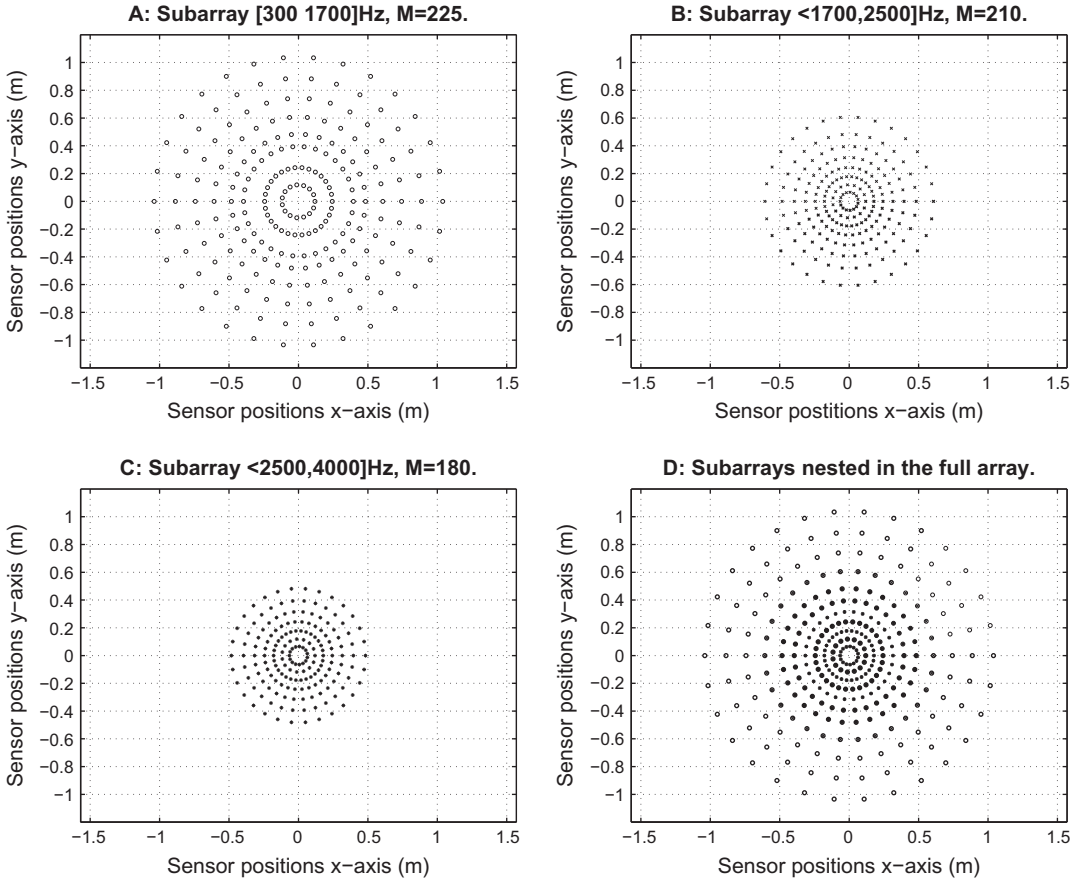

Fig. 3. Sub-array configurations and the full array.

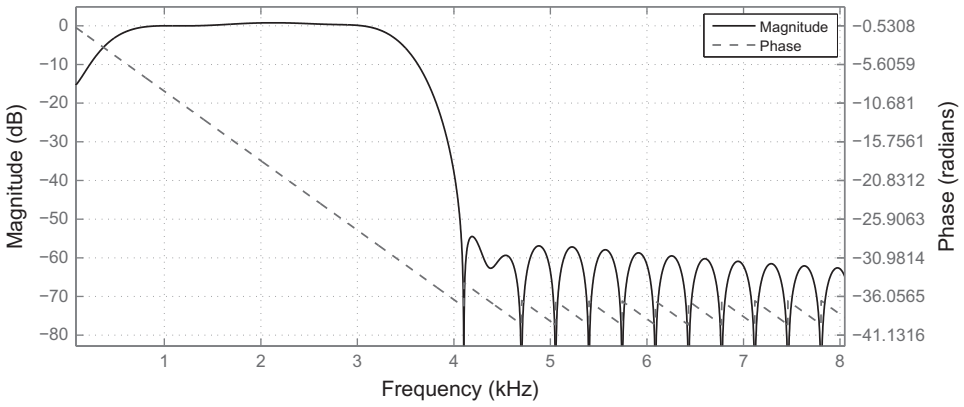

Fig. 4. Response of the sub-band filter combination.

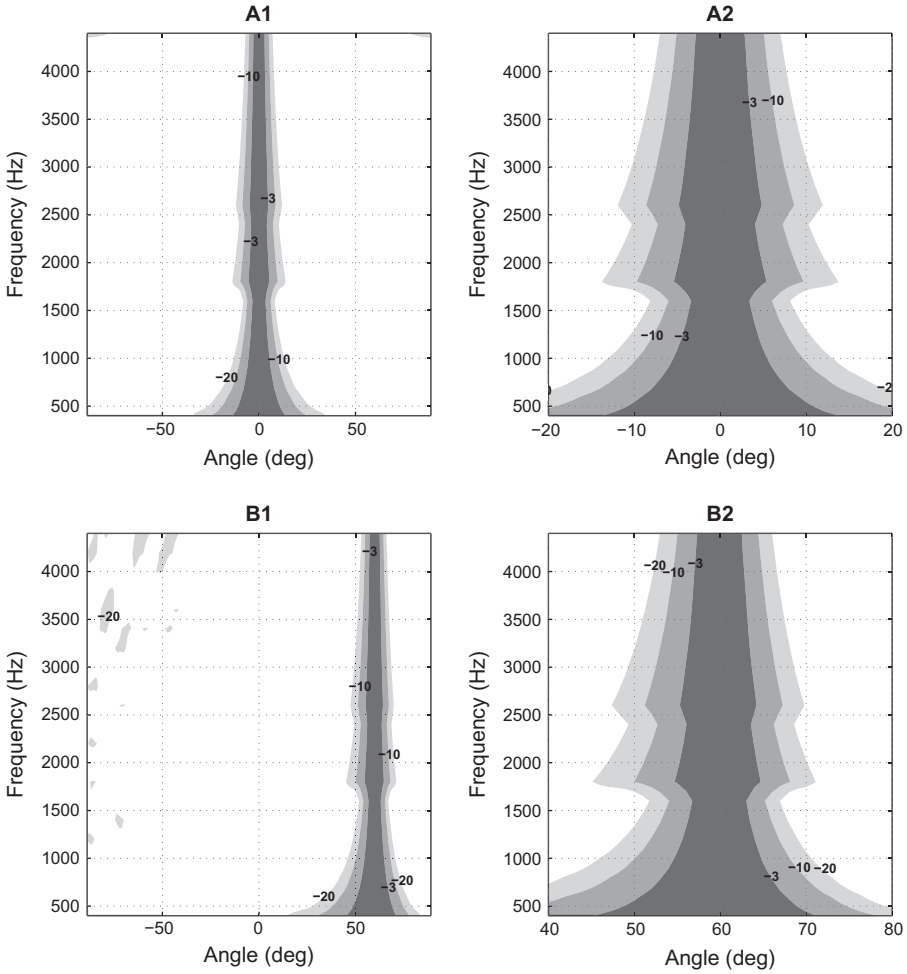

Fig. 5. Contours of the 3 dB, 10 dB, and 20 dB attenuation for varying frequencies. **A1:** Beam-width at the specified levels for a steering angle of 0°. All side-lobes are below −20 dB. **A2:** A1 zoomed in. **B1:** Beam-width at the specified levels for a steering angle of 60°. Notice that the side-lobes at the left side of the plot have increased to −20 dB level for this steering angle. **B2:** B1 zoomed in.

This array has a broadband response in the sense that the sidelobes are kept below a required level of −20 dB (except for small areas seen in Fig. 5B1). In addition, the −3 dB beamwidth does not vary with much more than 2° across frequencies above 1 kHz. For example, at 4 kHz the angular distance from the main response axis (in the center) to the curve for 3 dB attenuation is 1.4°, while at 1.5 kHz it is 3.5°, giving a difference of 2.1°. However, as seen in Fig. 5A2, the beamwidth increases for the frequencies below 1 kHz due to the limited aperture size, meaning that the array will work as a spatial low-pass filter with respect to noise outside the main lobe. While reduced dependence on frequency can be achieved [19,20], the response in Fig. 5 is considered as an acceptable compromise between the beam width and the side lobe levels, for evaluating the viability of the solution currently proposed. The need for any further improvements will emerge from the measurement investigation.

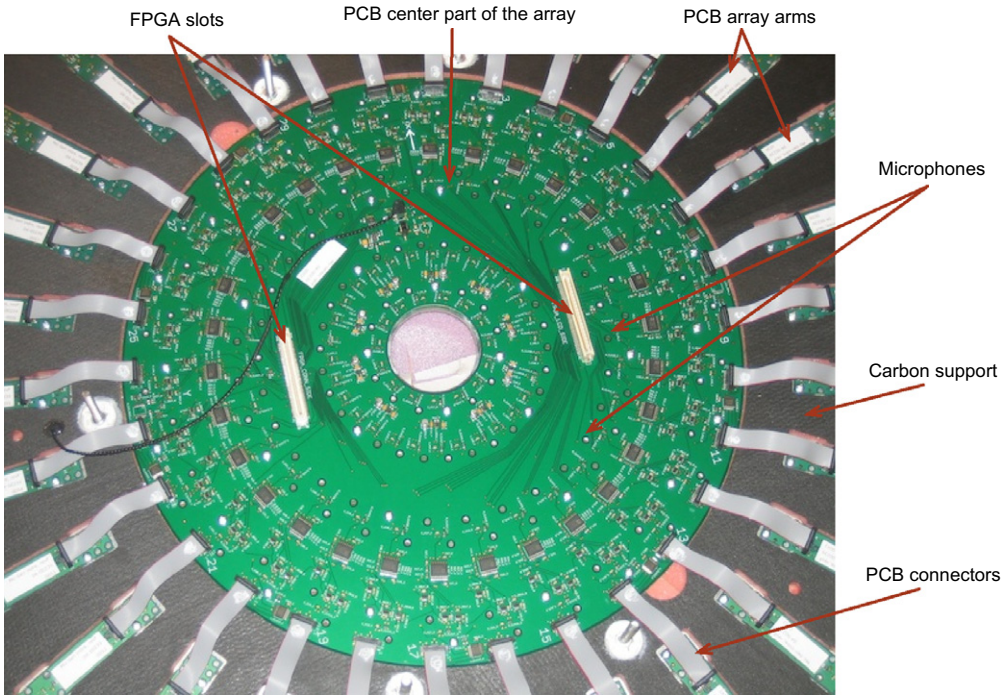

Fig. 6. Inside of the array. Printed circuit board with microphones and the FPGA connectors.

3.2. System architecture

The proposed array requires 300 microphones. The miniature MEMS microphones (footprint area 2.6 mm × 1.6 mm) are a cost-effective solution for array applications, and suitable for our design due to their flat frequency response (100 Hz to 10 kHz). In addition, MEMS microphones produced from the same silicon wafer are considered to have equal phase response. The signal-to-noise ratio of individual microphones is 60 dB.

Digital data from 300 microphones are sampled and transferred to the processing unit. The sampling process is considerably simplified using digital MEMS microphones with internal AD converters. The microphones are surface mounted on modular, printed circuit boards, as shown in Fig. 6. MEMS microphones are sensitive to mechanical load and direct contact between the microphone surface and the acoustic seal is avoided by placing elevated contact pads between each PCB and a carbon shield. A layer of an acoustically transparent and water repelling fabric covers microphone membranes for protection from moisture and contamination.

What we denote as the array unit in Fig. 7, is the physical disc containing the microphones, an FPGA, and a camera. The array unit is connected to the signal processing unit with a single cable, a multi-cable terminated with standard SMPTE Fiber HDTV connectors. Digital array data can in this way be transferred over large distances, e.g. to a remote control room, without loss or interference. The camera at the center of the array unit has the purpose of providing a live video used for beam steering. The details around visual beam steering and the beamforming are given in next section.

4. Signal processing and user interaction

The system is divided into channel specific processing (CSP) and beam specific processing (BSP). As illustrated in Fig. 7, the CSP is performed on an Xilinx Virtex-4 FPGA within the array unit, and consists of sampling of delta-sigma modulated data from the microphones and per-microphone demodulation to 16-bit samples. The BSP is performed on these 16-bit samples using a Mac Pro computer, and consists of user interaction and subsequent beamforming.

4.1. Channel specific processing

The MEMS microphones used in the array have integrated ADCs, and provide digital bitstreams created using 4th order delta-sigma modulators at high sampling rates. In our implementation, the microphones are clocked at a rate of 2.1 MHz and intended for a Nyquist rate of 33,075 Hz, which corresponds to an oversampling factor of 64. The theoretical SNR for a 4th order delta-sigma modulator operating at 64 times the Nyquist rate is 132 dB [17], which corresponds to a resolution of 22 bits per sample. This corresponds to a much lower noise level than the self-noise on most microphones, meaning that that quantization noise will not be the limiting factor for a single microphone, SNR-wise. Because the delta-sigma quantization noise is uncorrelated from microphone to microphone, it will be reduced by a factor $10\log_{10}(M)$ when M microphones are combined in an array.

Demodulation of the bitstream is carried out using a 5th order sinc-filter, followed by downsampling by a factor of 64. An efficient realization of this filter is a chain of five digital integrators, followed by downsampling, and finally a chain of five digital differentiators [18]. The register of each integrator/differentiator has a word length of 26 bits and is allowed to overflow naturally. The sinc-filter has been chosen for its simplicity, as implementing several hundred more general demodulation filters in this type of FPGA is simply not possible. A detailed treatment of demodulation and other topics in delta-sigma data conversion is found in [17]. Blocks of 16-bit samples from every microphone are multiplexed and transmitted as an UDP-packet over fiber optic gigabit Ethernet. The reduction from 22 to 16 bits per sample is due to the before mentioned analog SNR on each microphone.

4.2. Beam specific processing

All beamforming is performed in the time domain. For each of five possible beams, an operator specifies a focal point in the GUI

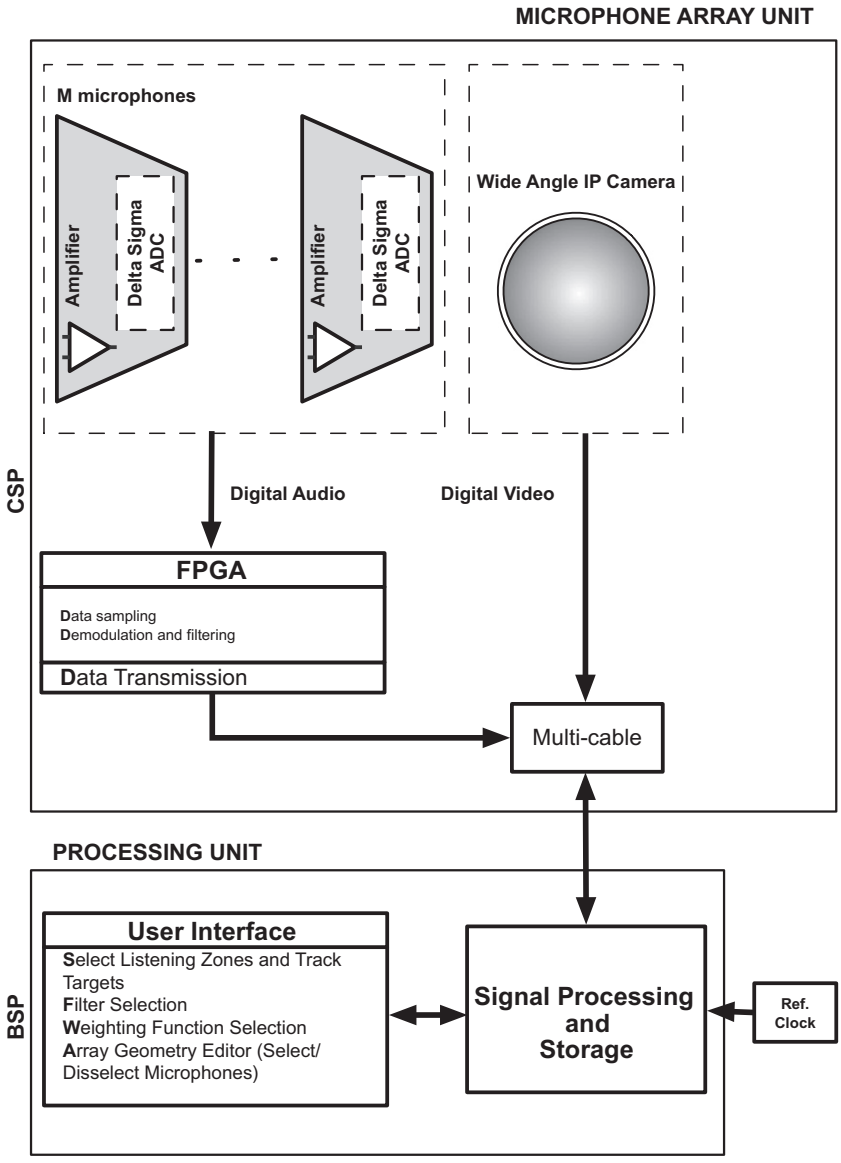

Fig. 7. System architecture.

shown Fig. 9. Whenever the focal point of a beam is moved, the system calculates a new set of delays. The delayed samples from each microphone are combined according to the previously described sub-array division, and processed as shown in Fig. 8. The output of each sub-array's beamformer is filtered using FIR-filters with 96 coefficients. The FIR-filters attenuate frequencies that are outside the bands that the sub-arrays are designed for. The filtered outputs of the sub-arrays are then summed to provide the final output. The system is designed to support multiple output channels with selectable beamforming and filtering options. Up to five channels can be beamformed simultaneously in real time. One channel can e.g. be used to track a moving source, while the other four channels can monitor stationary zones. Since the processing of the different channels is done in parallel, each channel has its own set of parameters (volume, weighting function for control of beamwidth, etc.).

4.3. Target selection

In the application in question, the selection of an area or a target of interest must be done manually, e.g. by a sound producer. One possible way of incorporating this with a microphone array system is by visual tracker, shown in Fig. 9. When the microphone array is arranged parallel to the ground, the overhead-view of the array's coverage is available at the control unit. Assuming flat ground and known height of the array, distances from the sources to the array are always known. The camera is calibrated to project the scene below the array correctly onto the screen, so that the visually chosen locations translate directly to the Cartesian coordinates. The coordinates are passed to the beamformer and the sound on the output of the system corresponds to the chosen locations. This approach avoids speech source detection problems and computationally intensive algorithms, but demands an operator. This is the major limitation in applications where manual tracking is not an option.

Video and audio are synchronized either by internal system clock, or time code input from an external device, meaning that replay can be synchronized with standard audio/video equipment. If all microphone data are stored, the same functionality is available in post-processing as described for real time processing, which is unique for microphone array systems.

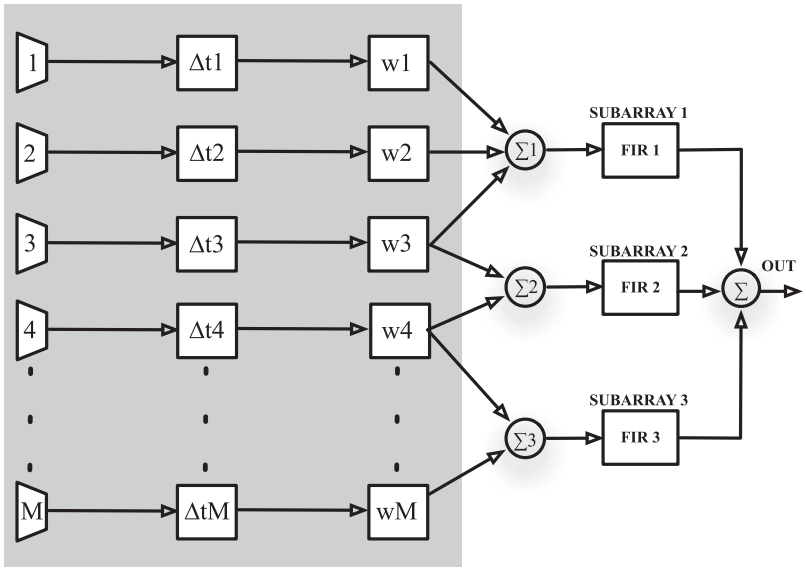

Fig. 8. Array processing scheme for a single beam.

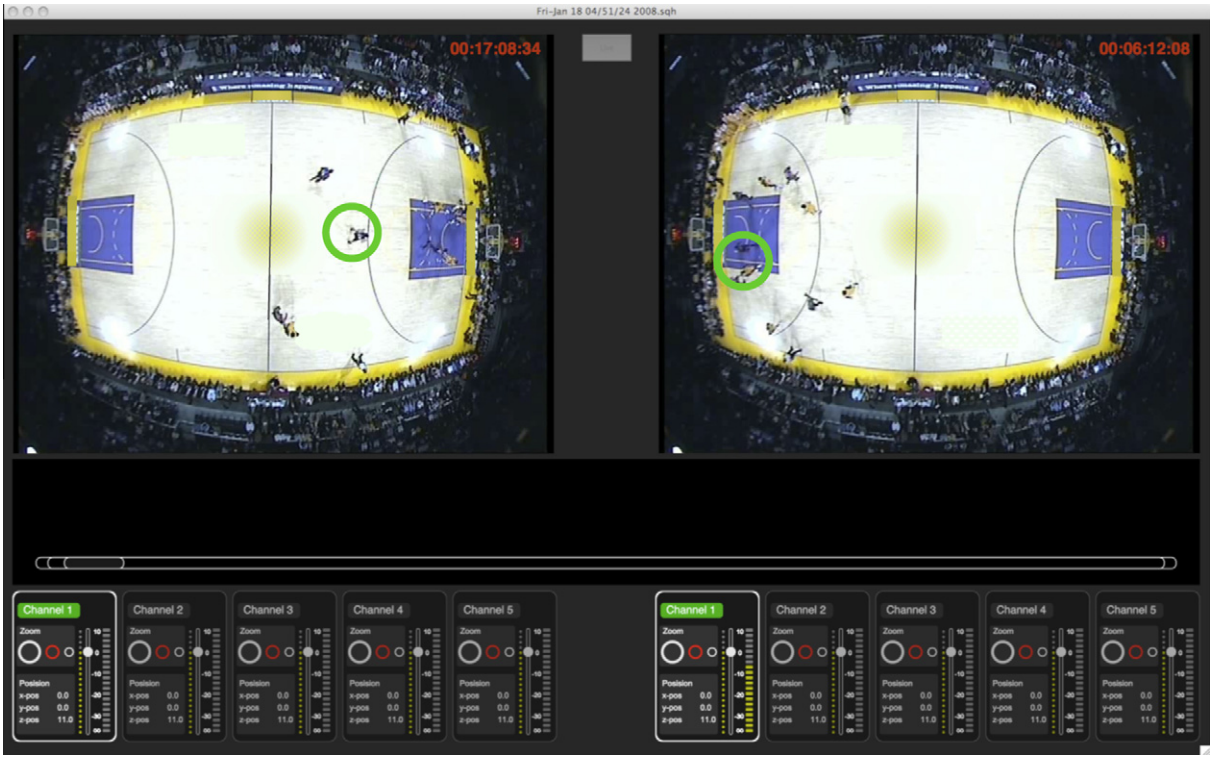

Fig. 9. GUI with the real-time processing at the left side, and replay at the right side.

5. Measurements

5.1. The anechoic chamber measurements

To confirm the actual spatial response of the array, we performed free field measurements in an anechoic chamber. The array was suspended from a system of hooks, as shown in Fig. 10. We measured the spatial response to sine waves (for frequencies within the speech bandwidth). The measured results and the corresponding theoretical beampatterns are shown in Fig. 11. The measured main-lobe widths are close to the theoretical beampatterns for all frequencies. The levels of the measured side-lobes are somewhat higher than expected, with the highest deviation of about 10 dB at 3 kHz. Some possible reasons for deviations from the predicted response are:

- SNR on individual microphones lower than expected.
- Phase errors:
  - Due to errors in the microphones or associated electronics.
  - Induced through incorrect steering or delay quantization.
  - Due to imprecise placement of the microphones on the array.
  - Due to the diffraction effects on the array surface.

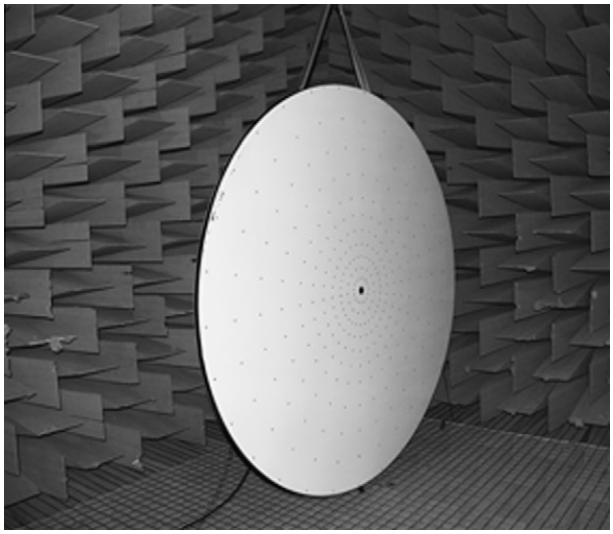

Fig. 10. Array in anechoic chamber.

- Amplitude errors, resulting in a "random" array weighting and a less than ideal sidelobe pattern.

By introducing one or more of these deviations into the theoretical beampatterns in Fig. 11, we would see similar effects as for the measured responses. However, determining the exact combination and magnitude of deviations is not trivial, and is outside the scope of this paper.

5.2. The sound field measurements on a basketball court

A second series of measurements was performed to analyze the performance of the array in a more realistic setting. Specifically, we wanted to verify that the directivity of the array is good enough to capture sound from the players at a sports arena in both high- and low-SNR situations.

To achieve this, the array was suspended above a basketball court, during the full course of a game. Sweeps were performed across the court and its immediate surroundings during intervals of intense cheering by the audience, typically after a team had scored. Only intervals where no signal of interest was present on the court were selected, to calculate the interference levels with which a potential signal would have to compete. Figs. 12 and 13 show the power of the interference $P_I(x,y,h)$ as a function of the focal point (x,y,h) on the floor, during two representative scenarios. Note that these plots are not intended as correct SPL estimates, but reflect the interference levels detected by the actual system for the corresponding points taking into account the directivity of the array. The SINR when focusing on a source at the position (x,y,h) would be given as:

$$SINR(x,y,h) = 10\log_{10}\left(\frac{\frac{N}{(x^2+y^2+h^2)}P_S(x,y,h)}{P_I(x,y,h)}\right), \quad (5)$$

where $P_S(x,y,h)$ is the power of the source of interest and h is the shortest distance from the array to the court. The factor in the numerator corresponds to the effects of array gain vs. path loss. As long as the number of elements is always larger than or equal to the square of the distance to the source, we can use the difference in dB between the source power and the power in Fig. 13 for a worst-case estimate of the SINR. In addition to high noise level from the audience, the occurrences of public address (PA) system reflections from the floor are seen in Fig. 13. When the array is mounted above the court the reflection points on the floor will act as sources of sound played from the PA system. This effect is present only during short periods of time and is generally not a major obstacle. If the sound from PA system is considered as noise, some of the possible counteracting methods (outside the scope of this article) would be:

- PA noise cancelation using a Wiener-filter with the actual (known) PA output as a noise reference.
- Positioning the loudspeakers of the PA system at positions and angles that do not induce strong reflections on the court.

To illustrate the array gain, we performed focusing on a speech source during an interval of moderate cheering, somewhat less than after a scoring. Fig. 14 shows the resulting signal, as compared to the signal captured by a single microphone on the array, illustrating the degree of noise reduction. Listening to a single microphone, the voices of the players are indistinguishable from the cheering and the ambient noise, while they are very clear in the array output. However, the sense of distance from the speech sources cannot be avoided. In addition, the voices of the players facing away from the array were more muffled than of those faced towards the array. This is not an effect caused by the processing, but by the acoustics and the fact that speech is directional and that this directivity varies with frequency. Another remark that can be made about the sound quality is that the ambient noise is low-pass filtered after array processing. Mixing the speech enhanced with the array system with a small amount of high-pass filtered ambient noise can give a more"natural" feel to the recordings, while retaining the intelligibility of the speech.

5.3. Microphone array compared to a shotgun microphone

We stated previously that the performance of the array should be comparable with the performance of the existing solutions. Having shown in the previous section that the array can provide sufficient SINR, we wanted to do a qualitative evaluation of the array. A comparison was carried out in a large room with similar acoustic properties as the basketball court using a high quality, omni-directional reference microphone, a shotgun microphone with 90 dB SNR, and the array. The room was quiet except for the noticeable ambient noise (air-conditioning system and echoes) and the speech source. We placed the reference microphone directly in front of the person speaking, while aiming the array and shotgun microphone at the person from distances of 8 m and 4 m respectively. Different persons of both genders participated as speech sources. The microphone array was originally band limited to [300, 3500] Hz. In this band, the sound quality of the microphone array was deemed clear but squeaky compared to the shotgun microphone, especially for certain male voices. By increasing the band to [100, 8000] Hz, the speech quality in the recordings done using the microphone array and the shotgun microphone were perceived as equal. The major difference between them was a superior ambient noise attenuation of the microphone array. As can be seen in Fig. 15, the array eliminates most of the ambient noise, which is present at the shotgun microphone. In situations where the ambient noise and SINR is even worse, the improvement offered by the array over a single shotgun microphone will only be better.

The initial drawback of the array was the limited frequency range. An extension of the frequency range to cover frequencies up to 8 kHz was found necessary. The array is not designed to give a good spatial response for frequencies this high, and some additional elements will need to be added. However, due to the aperiodic sampling of the array, pure grating lobes will not occur, just high side lobes. A larger problem is the limited performance for

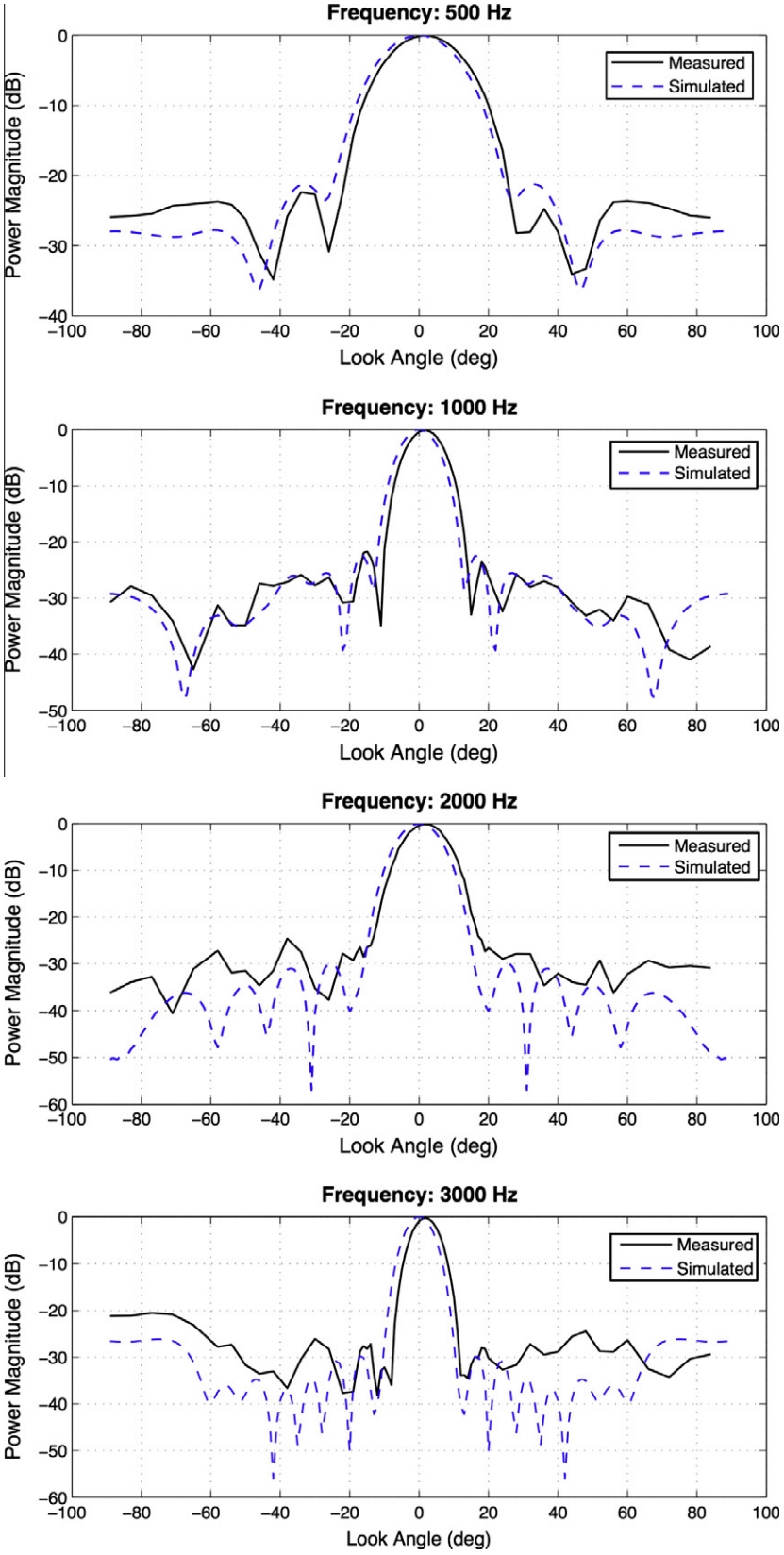

Fig. 11. Steered response measurements vs. theoretical curves.

low frequencies. The lower limit of the band cannot be set as low as 100 Hz in situations where there is a lot of noise, e.g. in the basketball scenario from the previous section. In these situations, the speech is experienced as "thin".

6. Conclusions

We proposed a large, portable microphone array solution for remote speech acquisition in live sports broadcasting. The system is a

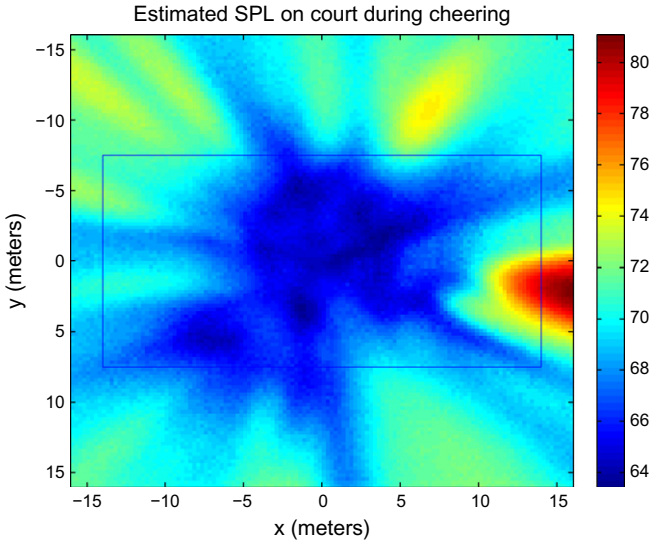

Fig. 12. Noise levels picked up by the array at the basketball court during cheering. The area corresponding to the court itself is marked by the rectangle.

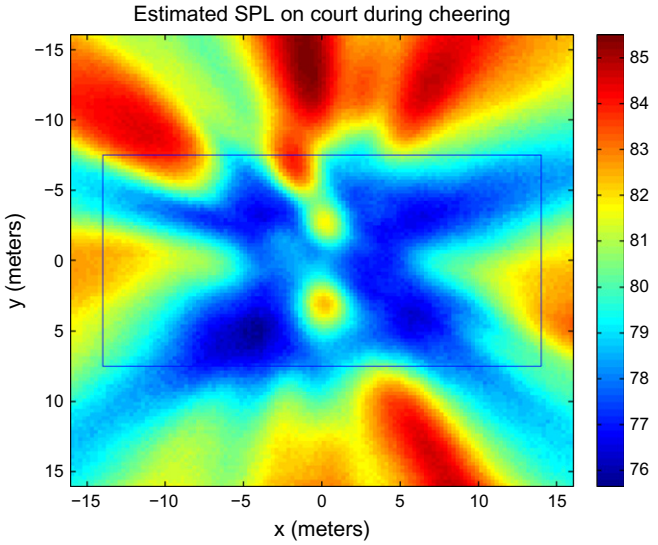

Fig. 13. Noise levels picked up by the array at the basketball court during cheering. The area corresponding to the court itself is marked by the rectangle.

complete solution, covering every step from sound capture to enhancement and processing via user interaction. Its design aimed to overcome particular constraints and difficulties associated with the requirements for speech capture in sports broadcasting, illustrated by the example scenario of a basketball court. The presented design of a large, broadband, MEMS microphone array and the suggested system architecture, opens up for the practical applications of large microphone arrays in fields beyond research and measurement. The performed measurements show that a well designed microphone array system can capture speech with good intelligibility from large areas, even in what can be considered as an extremely difficult acoustic environment. The feasibility of the array solution was demonstrated through estimation of worst-case interference power and by showing, by subjective evaluation, that it outperforms shotgun microphones in noise suppression, even when a shotgun microphone is placed close to a speech source in a controlled environment. Further research is necessary with respect to more efficient methods for joint audio-visual target tracking. Spatial filtering at frequencies below 600 Hz can be improved by e.g. adaptive beamforming methods. The advantages of microphone array systems like multi-channel output and spatial selectivity in post-processing are unique features that might be

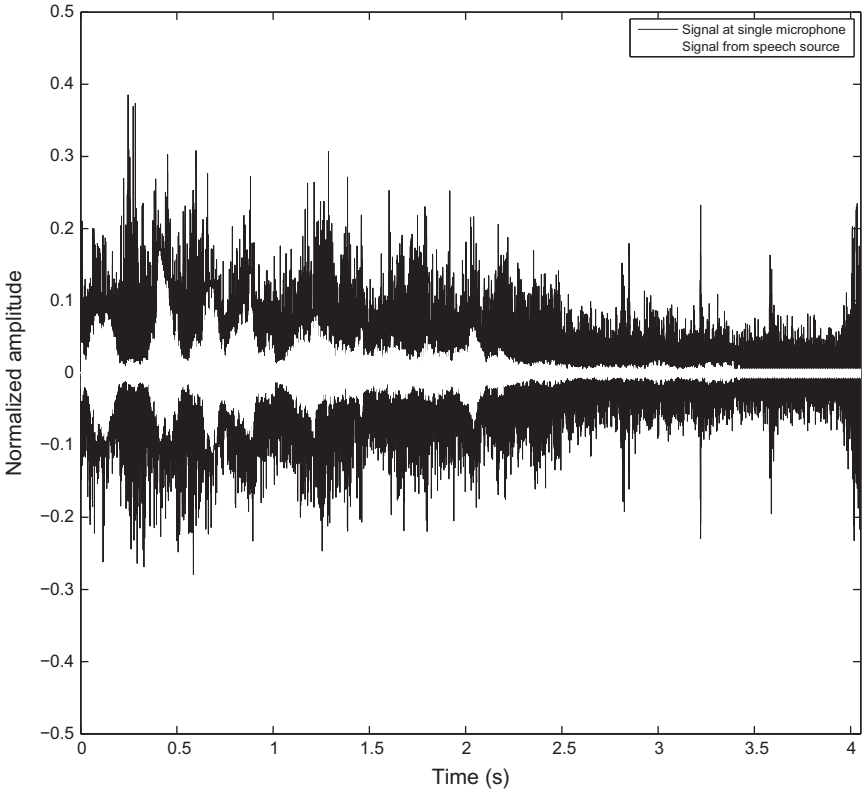

Fig. 14. Array output vs. single array microphone. Much of the spatial interference is gone in the array output.

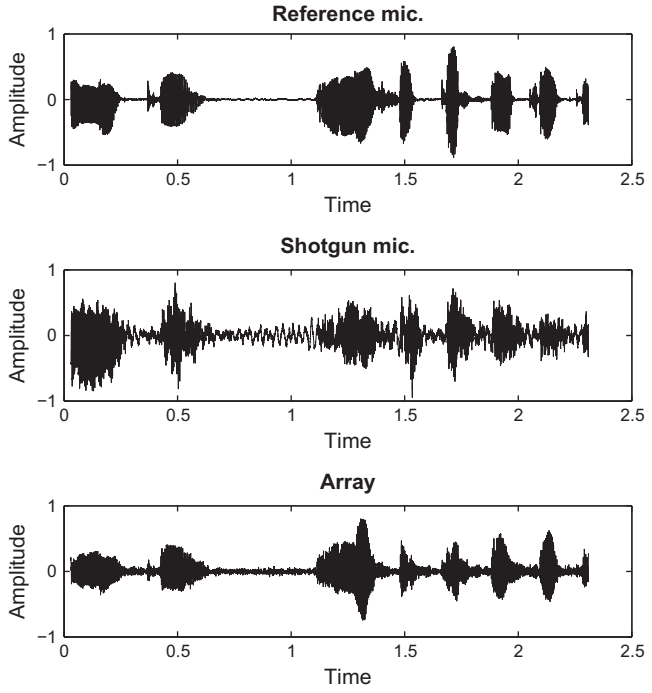

Fig. 15. Outputs of the reference microphone, the shotgun microphone, and the array. All three outputs are identically filtered and limited to [100, 8000] Hz.

desirable in the future for applications like entertainment and security. The quality of the speech captured by an array several meters away, cannot be as good as with a hand-held microphone, mostly due to the distance from the speech source and the acoustic effects that cannot be counteracted with an array. But, in applications where speech can not be captured by any other means, an array can provide desirable sound effects (shouting from the players, etc.) that can be mixed into a sound production and make it more vivid.

Acknowledgment

The authors thank Squarehead Technology for performing the engineering tasks, and SEAS for helping us with the anechoic chamber measurements. We would also like to thank Professor Sverre Holm for being the supervisor on this project. The research was partially funded by the Norwegian Research Council as part of the project "Broadband Audio Beam".

References

[1] Van Trees Harry L. Detection, estimation, and modulation theory IV: optimum array processing. John Wiley and Sons Inc.; 2003.
[2] Flanagan JL, Berkley DA, Elko GW, Sondhi MM. Autodirective microphone systems. Acustica 1991;73:58–71.
[3] Grenier Y. A microphone array for car environments. Speech Commun 1993;12(1):25–39.
[4] Dahl M, Claesson I, Nordebo S. Simultaneous echo cancellation and car noise suppression employing a microphone array. In: Proceedings of IEEE ICASSP'97, vol. I; April 1997. p. 239–42.
[5] Kiyohara K, Kaneda Y, Takahashi S, Nomura H, Kojima J. A microphone array system for speech recognition. In: Proceedings of IEEE ICASSP'97, vol. I; 1997. p. 215–8.
[6] Soede W, Berkhout AJ, Bilson F. Development of a directional hearing instrument based on array technology. J Acoust Soc Am 1993;94(2):85–798.
[7] Bradstein M, Ward D. Microphone arrays. Signal processing techniques and applications. Springer-Verlag; 2001.
[8] Silverman HF, Patterson WR, Flanagan JL, Rabinkin D. A digital processing system for source location and sound capture by large microphone arrays. In: Proceedings of the 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP '97), vol. 1, April 21–24; 1997. p. 251.
[9] Flanagan J, Johnston J, Zahn R, Elko G. Computer steered microphone arrays for sound transduction in large rooms. J Acoust Soc Am 1985;78(5):1508–18.
[10] Weinstein E, Steele K, Agarwal A, Glass J. LOUD: a 1020-node modular microphone array and beamformer for intelligent computing spaces. MIT/LCS Technical Memo, Cambridge; 2004.
[11] Silverman HF, Patterson III WR, Flanagan JL. The huge microphone array. IEEE Concurr 1999;7(1):32–47.
[12] Hafizovic I, Kjølerbakken M, Jahr V. System configuration for high quality audio capturing in a large microphone array. In: AES 31st international conference, new directions in high resolution audio; June 2007.
[13] Gray DA. Effect of time-delay errors on the beam pattern of a linear array. IEEE J Ocean Eng 1985;OE-10(3):269–77.
[14] Elliot RS. Antenna theory and design. Englewood Cliffs, New Jersey: Prentice-Hall; 1981.
[15] Ishimaru A, Chen YS. Thinning and broadbanding antenna arrays by unequal spacing's. IEEE Trans Antenn Propagat [legacy, pre-1988] 1965;13(1):34–42.
[16] Ishimaru A. Theory of unequally-spaced arrays. IEEE Trans Antenn Propagat [legacy, pre-1988] 1962;10(6):691–702.
[17] Norsworthy SR, Schreier R, Themes GC, editors. Delta-sigma data converters: theory, design, and applications. IEEE Press; 1992.
[18] Hogenaur EB. An economical class of digital filters for decimation and interpolation. IEEE Trans Acoust Speech Signal Process 1981;ASSP-29.
[19] Doles III J, Benedict F. Broad-band array design using the asymptotic theory of unequally spaced arrays. IEEE Trans Antenn Propagat 1988;36(1):27–33.
[20] Ward DB, Kennedy RA, Williamson RC. Theory and design of broadband sensor arrays with frequency invariant far field beam patterns. J Acoust Soc Am 1995;97:1023–34.