Development and Evaluation of a Digital MEMS Microphone Array for Spatial Audio

Anastasios Alexandridis*†, Stefanos Papadakis*, Despoina Pavlidi*† and Athanasios Mouchtaris*†
*FORTH-ICS, Heraklion, Crete, Greece, GR-70013
† University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013

***Abstract*—We present the design of a digital microphone array comprised of MEMS microphones and evaluate its potential for spatial audio capturing and direction-of-arrival (DOA) estimation which is an essential part of encoding the soundscape. The device is a cheaper and more compact alternative to analog microphone arrays which require external—and usually expensive—analog-to-digital converters and sound cards. However, the performance of such digital arrays for DOA estimation and spatial audio acquisition has not been investigated. In this work, the efficiency of the digital array for spatial audio is evaluated and compared to a typical analog microphone array of the same geometry. Our results indicate that our digital array achieves the same performance as its analog counterpart, thus offering a cheaper and easily deployable device, suitable for spatial audio applications.**

I. Introduction

Technological advances in the last few decades have led to a significant reduction in the cost of microphones and loudspeakers. Today, a standard loudspeaker array, such as the 5.1 surround system, is available to most consumers. As a result, immersive audio capturing and reproduction techniques are gaining attention for applications, such as immersive collaboration, telepresence, gaming, and recording and reproduction of multichannel music. However, capturing the soundscape with spatial information is a challenging task, especially in scenarios with a large or varying number of sound sources.

Spatial audio capturing methods typically employ a parametric model to encode spatial attributes of the soundscape, such as the directions-of-arrival (DOAs), or the locations of the sound sources [1]. Microphone arrays have been used as a robust solution to estimate these spatial attributes [2]–[5]. Array processing in the time-frequency domain is utilized to encode the soundscape using DOA estimates—such as in [2], [3]—and estimate information about the diffuse sound part, as in [4], [5]. A parametric technique for spatial audio using microphone arrays that operates on the spherical harmonic domain is also presented in [6]. Operating on the time-frequency domain, our previously proposed method utilizes the parametric model and source separation—through beamforming and post-filtering on the microphone array signals—to encode the spatial properties of the soundscape [7], [8].

This research has been funded in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 644283, Project LISTEN, and in part by EU and Greek national funds through the National Strategic Reference Framework (NSRF), grant agreement No 11ΣΥΝ_6_1381, Project SeNSE.

To date, microphone arrays are constructed using analog—usually omnidirectional—microphones. As such, they require amplifiers and a multichannel sound card which performs the analog-to-digital (AD) conversion and offers the interface to connect with a computer. All this necessary equipment makes the array too expensive, cumbersome and hinders its use for everyday end-users. Recent advances in MEMS (Micro Electro Mechanical System) microphones have led to the development of low-cost digital microphones that do not require amplifiers and AD converters. Using MEMS, digital microphone arrays can be constructed with a significant reduction in cost compared to their analog counterpart. Digital arrays can be highly portable and compact due to their high level of integration that does not require an external sound card.

Recent works on digital microphone arrays discuss their implementation and compare their performance to analog arrays for various signal processing tasks, such as distant speech recognition [9], recognition of overlapping speech [10], speech enhancement [11], and speaker diarisation [12]. The work in [13] presents a digital array design for aeroacoustic measurements, while the implementation of a system for sound acquisition with MEMS microphones is presented in [14]. However, this system comprises a very large number of microphones (300 elements), which greatly exceeds the number of microphones found in typical microphone array systems.

As mentioned above, arrays with digital MEMS microphones have been examined for various signal processing tasks. However, to the best of our knowledge, the use of such arrays for spatial audio acquisition and reproduction has not been considered so far. In this paper, we investigate the potential of a digital microphone array for DOA estimation and spatial audio capturing using our previously proposed methods of [7], [8], [16]. To do that, we built our own digital array comprised of MEMS microphones. Through listening tests with various types of recordings, we compare the performance of our digital array to an analog one. Our results in Section IV reveal that the performance of our digital array is very similar to that of the analog array, albeit it benefits from significant cost savings and is much more easily deployable.

II. Digital microphone array

In order to investigate the capabilities of digital MEMS microphones we designed, developed, and manufactured our own digital array. Instead of incorporating expensive, bulky, analog microphones, which require the support of an analog

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

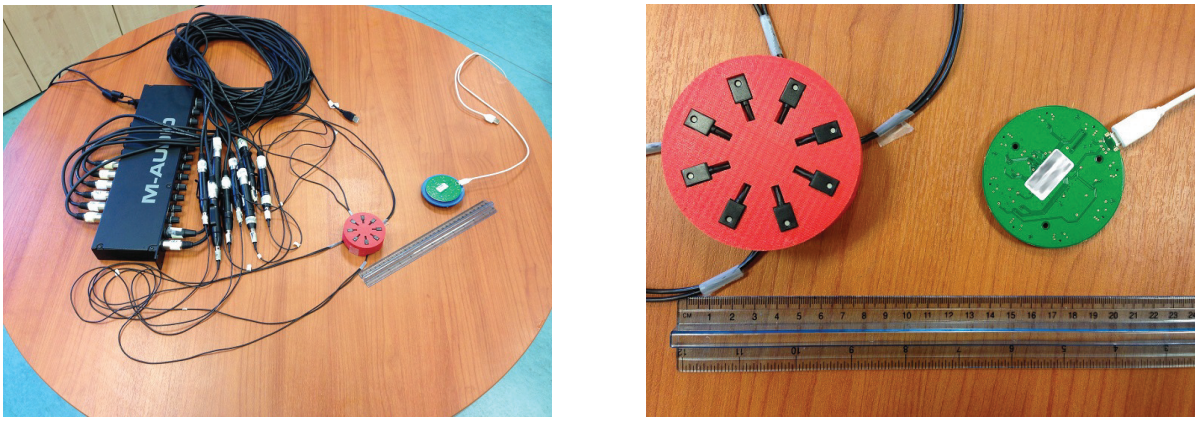


Fig. 1. The analog microphone array with all the necessary cables and the external multichannel sound card along with its digital MEMS microphones counterpart. All the bulky and expensive equipment can be replaced by our digital array, which requires only one USB cable.

amplification stage before any conversion to digital signal, the digital microphone arrays utilize microphones that directly provide a digital representation of the signal. The MEMS microphones' digital bitstream output should be processed accordingly by a specialized digital component, e.g., ASIC, FPGA, micro-processor, which is then formatted and multiplexed in order to be forwarded to the host PC by the use of an interface of adequate throughput, e.g., USB, Firewire, Ethernet.

We opted to use the InvenSense ICS-43432 [15] digital MEMS bottom port microphones which provide state of the art performance with 50 to 20000 Hz frequency response and 65 dB SNR, presenting only 29 dBA equivalent input noise. These microphones, although they have a relative low sensitivity of -26 dBFS, tested with a 1 kHz signal at 94 dB SPL, they can handle quite loud signals of over 110 dB SPL with less than 1% total harmonic distortion (THD). Another significant advantage of the selected microphones is the low variance of their sensitivity, which the manufacturer sets at ± 1 dB, opposed to the analog array ones that typically require calibration, due to the variability of the amplification stage among others. Our digital array is equipped with eight of the aforementioned microphones in a uniform circular arrangement. Its diameter is 60 mm, measured from the microphone ports, while the whole board's diameter is 67 mm. We designed the array with a USB 2.0 interface, which also provides the power to the device, therefore a single cable is needed in order for the device to operate. The host PC operating system recognizes the digital array as a typical 8 input-channels USB sound card, enabling maximum compatibility with any audio recording software. Thus, a circular board of less than 7 cm diameter and 1 cm height is able to substitute multiple cables, bulky connectors and expensive multichannel sound cards at a fraction of the cost (see Fig. 1).

III. Spatial audio capturing and reproduction method

In this section, we discuss our previously proposed method which we use to evaluate our digital MEMS microphone array for real-time spatial audio acquisition and reproduction. The method is divided into two stages: the capturing stage, where the microphone array signals are processed to extract the spatial attributes of the soundscape and the reproduction stage, where the encoded soundscape is reproduced using multiple loudspeakers.

We consider an environment where P sound sources are active. A uniform circular microphone array with M microphones is used to capture the soundscape. The microphone array signals are transformed into the Short-Time Fourier Transform (STFT) domain. To estimate the number of active sources and their DOAs, we utilize the method of [16] which is capable of estimating the DOAs in real-time and with high accuracy in reverberant environments for multiple simultaneously active sources. The method outputs the estimated number of sources, $\hat{P}_k$, and a vector with the estimated DOAs for each source (with 1° resolution), $\boldsymbol{\theta}_k = \left[\theta_1 \cdots \theta_{\hat{P}_k}\right]$, per time frame k. The interested reader is referred to [16] for more details on the DOA estimation and counting which are omitted here due to space limitations.

Source separation is then performed by applying a beamforming and post-filtering operation on the microphone array signals. First, $\hat{P}_k$ fixed superdirective beamformers are employed, each of them steering its beam to one of the directions in $\boldsymbol{\theta}_k$. The filter coefficients for the beamformer are calculated by maximizing the array gain, while maintaining a minimum constraint on the white noise gain [17]. Since the beamformer is fixed, i.e., signal independent, the filter coefficients are estimated offline, facilitating its use for real-time applications.

The beamforming operation results in the signals $B_s(k,\omega)$, $s = 1, \cdots, \hat{P}_k$, which are given by:

$$B_s(k,\omega) = \mathbf{w}(\omega,\theta_s)^H \mathbf{X}(k,\omega), \tag{1}$$

where $\mathbf{X}(k,\omega)$ is the $M \times 1$ vector containing the microphone array signals for time frame k and frequency index ω, $\mathbf{w}(\omega,\theta_s)$ is the $M \times 1$ vector of complex beamformer weights for frequency index ω and steering direction θ_s, and $(\cdot)^H$ denotes the Hermitian transpose operation.

A post-filter is then applied to the beamformer output to isolate the sources' signals. The post-filter constructs $\hat{P}_k$ binary masks. The mask for the sth source is defined as [18]:

$$U_s(k,\omega) = \begin{cases} 1, & \text{if } s = \arg\max_p |B_p(k,\omega)|^2,\ p = 1,\cdots,\hat{P}_k \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The final separated source signals are:

$$\hat{S}_s(k,\omega) = U_s(k,\omega)B_s(k,\omega), \quad s = 1,\cdots,\hat{P}_k. \tag{3}$$

During the post-filtering operation, in each time-frequency point only the source with the highest energy is kept; the others are set to zero. The fact that each time-frequency point is assigned to only one source allows us to implement a simple, yet efficient downmixing scheme by summing the individual separated signals $\hat{S}_s(k,\omega), s = 1,\cdots,\hat{P}_k$ in order to create one full spectrum signal. This monophonic downmix signal together with side-information, namely the DOA of the source that each frequency belongs to, are used to encode the soundscape and reproduce it at the reproduction stage. The downmix signal can be further encoded with an MP3 encoder to reduce the bitrate without noticeable degradation in the spatial impression or quality of the reproduced soundscape [7]. A lossless compression scheme for the side-information channel has also been proposed in [7].

At the reproduction stage, multiple loudspeakers are used to re-create the encoded soundscape. The downmix signal is first transformed into the STFT domain and Vector-Base Amplitude Panning (VBAP) [19] is applied at each frequency bin in order to reproduce each frequency of the downmix signal from the direction indicated by the side-information channel.

Our method also offers the potential to include a diffuse part in the soundscape by setting a cut-off frequency ω_C, which defines the frequency up to which directional information—through beamforming and post-filtering—is extracted. For the frequencies above ω_C the spectrum from an arbitrary microphone of the array is included to the downmix signal, keeping no side-information for these frequencies. For reproduction, the diffuse part is played back from all loudspeakers after appropriate scaling with the reciprocal of the square root of the number of loudspeakers for energy preservation. The effect of ω_C on the spatial impression and sound quality of the reproduced soundfield has been investigated through listening tests for various types of sounds and the interested reader is referred to [7], [8]. Since the focus of this paper is the comparison of the different microphone array technologies (analog and digital) for spatial audio applications, in the remainder we consider $\omega_C = f_s/2$ which corresponds to the case of no diffuse part in the encoded soundscape.

IV. Experimental evaluation

To evaluate the performance of our digital microphone array, we conducted a listening test using real microphone array recordings obtained from our digital array and a typical analog array of the same geometry.

A. Test data

For the listening test we used three recordings[1]: a 10-second rock music recording with one male singer at $0°$ and 4 instruments at $45°$, $90°$, $270°$, and $315°$, which is publicly available from the band "Nine Inch Nails"; a 15-second classical music recording with 4 sources at $0°$, $45°$, $90°$, and $270°$, which is available from [20]; and a 10-second speech recording with three speakers, where two speakers (one male at $225°$, one female at $45°$) are continuously and simultaneously active from the beginning and a third speaker (female at $135°$) starts speaking at the third second and remains active thereafter, resulting in three simultaneously active sound sources. The recordings were multi-track (each source on a separate track) and included both impulsive and non-impulsive sounds. Each source signal (track) was reproduced by a loudspeaker (Genelec 8050) located at the aforementioned directions at 2.10 m distance from the center of the array. The separate tracks were reproduced simultaneously and captured from the microphone array. Two arrays were used for recording: our digital microphone array, described in Section II, and an analog array of the same geometry (8-microphones circular array with 3 cm radius) comprised of Shure SM93 omnidirectional microphones—which have very good specifications [21]—and an M-Audio M-Track Eight USB sound card with 8 channels. The sampling frequency for both arrays was set to 48 kHz. The recordings took place in the reference listening room, located at FORTH-ICS, which follows the ITU-R BS.1116 recommendation [22]. The reverberation time of the room was measured to be $T_{60} = 0.27$ seconds.

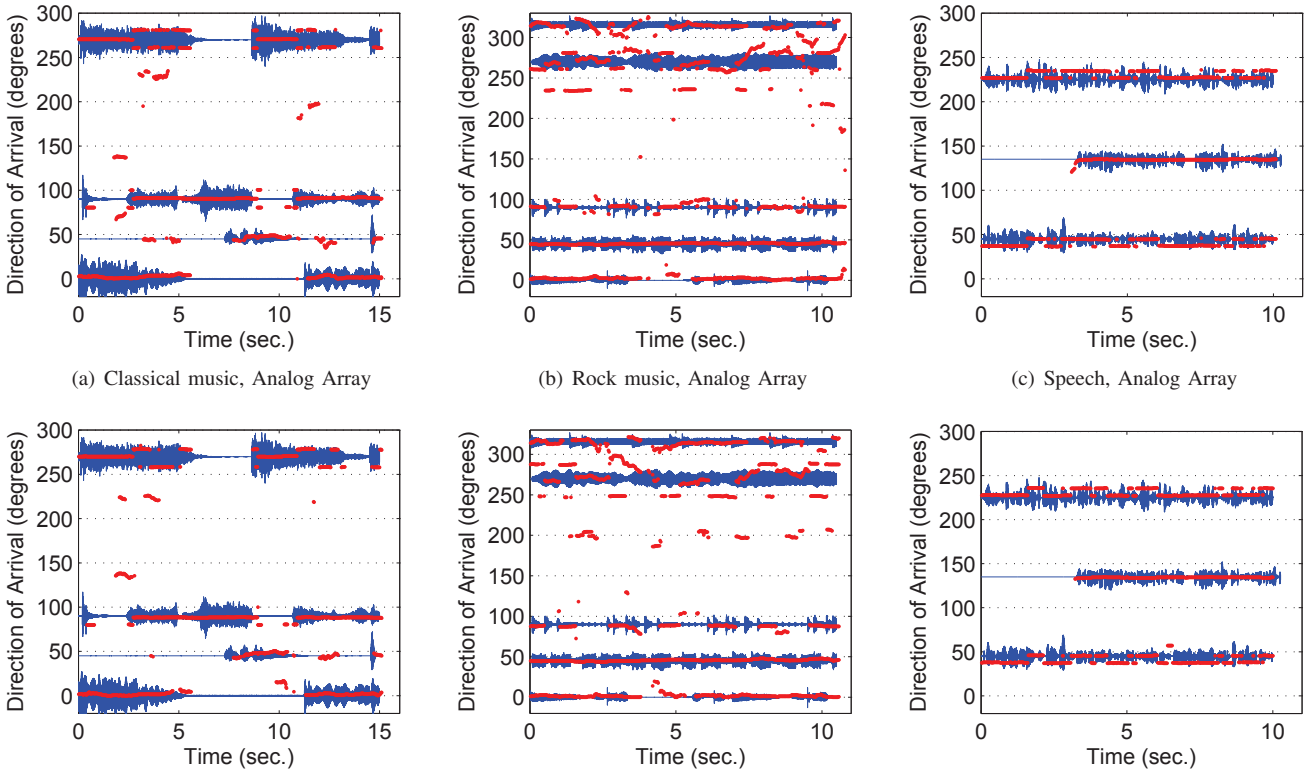
B. DOA estimation results

First, we evaluate the performance of the DOA estimation and counting method used during the spatial audio capturing, which was proposed in [16]. Fig. 2 shows the DOA estimates obtained for each time frame, for all recordings, using the analog (top row) and the digital (bottom row) array. The signal of each source is plotted at its corresponding direction and the DOA estimates are overlayed on top. We observe consistent and smooth DOA estimates, especially for the speech and classical music recording. Some spurious estimates are evident in the classical recording which occur due to the overestimation of the number of sources at these frames. A performance degradation can be observed for the rock music recording, where the sources at $90°$, $270°$, and $315°$ failed to be estimated for short periods of time due to the challenging setup in terms of the number of sources and their angular separation. Comparing the DOA estimation results between the two arrays, we can observe that the performance of the digital array is very similar to the analog array for all three recordings.

C. Listening test results

The listening test was based on the ITU-R BS.1116 methodology [22] and took place at the reference listening room,

[1]The sound files used in the experiment are available at http://www.ics.forth.gr/~mouchtar/eusipco2016.html



(a) Classical music, Analog Array (b) Rock music, Analog Array (c) Speech, Analog Array

(d) Classical music, Digital Array (e) Rock music, Digital Array (f) Speech, Digital Array

Fig. 2. DOA estimates using the analog array signals (top row) and the digital array signals (bottom row).

located at FORTH-ICS. Thirteen volunteers participated in the test (authors not included). For reproduction we employed a circular loudspeaker configuration with 8 loudspeakers (Genelec 8050) and a radius of 2.10 m. The coordinate system used for reproduction places the 0° in front of the listener, increasing clockwise. The microphone array signals from each array were processed using our spatial audio capturing and reproduction method (Section III), using frames of 2048 samples with 50% overlap, windowed with a Hanning window. The FFT size was 4096.

To create the reference signals, each track was positioned at its corresponding direction using VBAP [19]. The low-pass filtered (3.5 kHz cutoff frequency) reference recording served as quality anchor, while the signal from an arbitrary microphone, played back from all loudspeakers, was used as a spatial anchor. The subjects—sitting at the "sweet spot"—were asked to compare sample recordings against the reference using a 5-scale grading, with 1 denoting "very annoying" difference compared to the reference and 5 denoting "not perceived" difference with respect to the reference. The test was conducted in two separate sessions: in the first session the subjects were asked to grade the recordings in terms of spatial impression, while in the second session the grading was based on sound quality.

The mean scores and 95% confidence intervals for the spatial impression and quality sessions are depicted in Figs. 3 and 4. As expected, the ratings for the reference and anchor signals are at the opposite ends of the scale. All recordings exhibit satisfactory results. A performance degradation of the rock music recording compared to the others can be observed, which can be explained from the DOA estimation results in Fig. 2. As mentioned in Section IV-B, some sources could not be detected for short periods of time, which caused degradation in the reproduction for the rock music recording.

More importantly, the scores achieved using the analog and digital array signals are very close to each other for all recordings and for both spatial impression and sound quality. An Analysis of Variance (ANOVA) indicates that no statistical difference between the two arrays exists in the spatial impression and quality ratings, with p-values significantly greater than 0.01. Multiple comparison tests using Tukey's least significant difference at 95% confidence were performed on the ANOVA results. The comparison tests again indicate that no significant differences between the two arrays exist for both spatial impression and sound quality, validating that our digital array can achieve very similar performance to its analog counterpart.

V. Conclusions

In this work we presented our own digital microphone array with MEMS microphones and investigated its potential application for capturing and reproducing spatial audio. We conducted an experimental study where we compared the

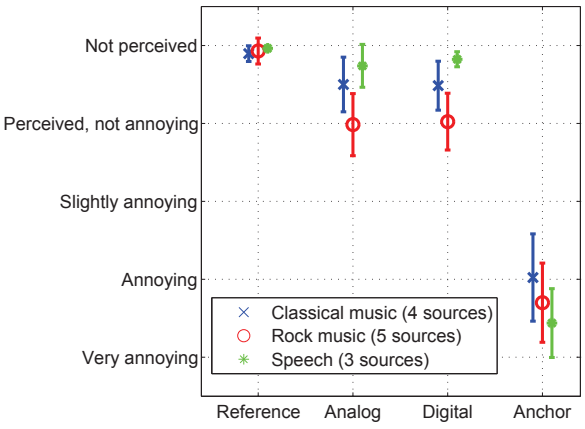


Fig. 3. Listening test results for spatial impression.

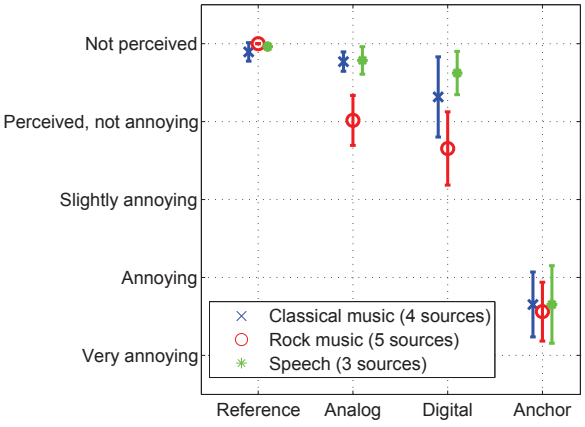


Fig. 4. Listening test results for sound quality.

performance of the digital array to a typical analog array. We applied our previously proposed method for spatial audio capturing and reproduction to the signals acquired from both—analog and digital—arrays and through listening tests we compared the efficiency of the two arrays. Our listening test results revealed that both arrays achieve the same reproduction quality both in terms of spatial impression and sound quality. Thus, our digital microphone array is a low-cost, portable, and efficient alternative to analog microphone arrays with the potential to emerge as an essential device in consumers' everyday life. In the future, we intend to investigate the performance of our digital array in other signal processing tasks, such as source separation and distant speech recognition.

REFERENCES

[1] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, March 2015.

[2] M. Cobos, S. Spors, J. Ahrens, and J. J. Lopez, "On the use of small microphone arrays for wave field synthesis auralization," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, 2012.

[3] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 2:1–2:13, 2010.

[4] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA), 2008.*, May 2008, pp. 37–40.

[5] O. Thiergart, M. Kallinger, G. D. Galdo, and F. Kuech, "Parametric spatial sound processing using linear microphone arrays," in *Microelectronic Systems*, A. Heuberger, G. Elst, and R. Hanke, Eds. Springer Berlin Heidelberg, 2011, pp. 321–329.

[6] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, Aug 2015.

[7] A. Alexandridis, A. Griffin, and A. Mouchtaris, "Directional coding of audio using a circular microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 296–300.

[8] ——, "Capturing and reproducing spatial audio based on a circular microphone array," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[9] E. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010, pp. 5106–5109.

[10] E. Zwyssig, F. Faubel, S. Renals, and M. Lincoln, "Recognition of overlapping speech using digital MEMS microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7068–7072.

[11] Z. Skordilis, A. Tsiami, P. Maragos, G. Potamianos, L. Spelgatti, and R. Sannino, "Multichannel speech enhancement using MEMS microphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2729–2733.

[12] E. Zwyssig, S. Renals, and M. Lincoln, "On the effect of SNR and superdirective beamforming in speaker diarisation in meetings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4177–4180.

[13] D. P. Arnold, T. Nishida, L. N. Cattafesta, and M. Sheplak, "A directional acoustic array using silicon micromachined piezoresistive microphones," *The Journal of the Acoustical Society of America*, vol. 113, no. 1, pp. 289–298, 2003.

[14] I. Hafizovic, C. C. Nilsen, M. Kjolerbakken, and V. Jahr, "Design and implementation of a MEMS microphone array system for real-time speech acquisition," *Applied Acoustics*, vol. 73, no. 2, pp. 132 – 143, 2012.

[15] "InvenSense ICS-43432 Digital MEMS Microphone," http://www.invensense.com/products/digital/ics-43432.

[16] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[17] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[18] H. K. Maganti, D. Gatica-perez, and I. A. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, 2007.

[19] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[20] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, Dec. 2008.

[21] "Shure SM93 Lavalier Microphone," http://www.shure.com/americas/products/microphones/sm/sm93-lavalier-microphone.

[22] ITU-R, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.