MULTICHANNEL SPEECH ENHANCEMENT USING MEMS MICROPHONES

Z. I. Skordilis[1,3], *A. Tsiami*[1,3], *P. Maragos*[1,3], *G. Potamianos*[2,3], *L. Spelgatti*[4], *and R. Sannino*[4]

[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]ECE Dept., University of Thessaly, 38221 Volos, Greece
[3]Athena Research and Innovation Center, 15125 Maroussi, Greece
[4] Advanced System Technology, STMicroelectronics S.p.A., 20041 Agrate Brianza, Italy

{antsiami, maragos}@cs.ntua.gr, gpotam@ieee.org, {luca.spelgatti, roberto.sannino}@st.com

ABSTRACT

In this work, we investigate the efficacy of Micro Electro-Mechanical System (MEMS) microphones, a newly developed technology of very compact sensors, for multichannel speech enhancement. Experiments are conducted on real speech data collected using a MEMS microphone array. First, the effectiveness of the array geometry for noise suppression is explored, using a new corpus containing speech recorded in diffuse and localized noise fields with a MEMS microphone array configured in linear and hexagonal array geometries. Our results indicate superior performance of the hexagonal geometry. Then, MEMS microphones are compared to Electret Condenser Microphones (ECMs), using the ATHENA database, which contains speech recorded in realistic smart home noise conditions with hexagonal-type arrays of both microphone types. MEMS microphones exhibit performance similar to ECMs. Good performance, versatility in placement, small size, and low cost, make MEMS microphones attractive for multichannel speech processing.

Index Terms— microphone array speech processing, multichannel speech enhancement, MEMS microphone array

1. INTRODUCTION

In recent years, much effort has been devoted to designing and implementing ambient intelligence environments, such as smart homes and smart rooms, able to interact with humans through speech [1–3]. For example, among others, ongoing such research is being conducted within the EU project named "Distant-speech Interaction for Robust Home Applications" (DIRHA) [4]. Sound acquisition is a key element of such systems. It is desirable that sound sensors be embedded in the background, imperceptible to human users, so the latter can interact with the system in a seamless, natural way.

The newly developed technology of ultra-compact sensors, namely Micro Electro-Mechanical System (MEMS) microphones, facilitates the integration of sound sensing elements within ambient intelligence environments. Their very small size implies versatility in their placement, making them very appealing for use in smart homes. However, the need for far-field speech acquisition gives rise to the problem of noise suppression. Therefore, aside from the MEMS microphones advantage in terms of size, an evaluation of their effectiveness for multichannel speech enhancement is needed.

In this work, the focus is on investigating the performance of MEMS microphone arrays for speech enhancement. First, we experimentally compare the effectiveness of linear and hexagonal array geometries for this task. Using the versatile MEMS array, a new speech corpus was collected, which contains speech in both diffuse and localized noise fields, captured with linear and hexagonal array configurations. A variety of multichannel speech enhancement algorithms exist [5–8]. Here, a state-of-the-art such algorithm proposed in [9] is used on the new speech corpus, in order to explore the effectiveness of the MEMS microphones and the array geometry for suppression of various noise fields. The results indicate that the hexagonal array configuration achieves superior speech enhancement performance. Then, MEMS microphones are compared to Electret Condenser Microphones (ECMs) using the ATHENA database [10], a corpus containing speech recorded in a realistic smart home environment. This corpus contains recordings from closely positioned pentagonal MEMS and ECM arrays. The use of hexagonal-type configuration was motivated by its superior performance on the first set of experiments. The MEMS array achieves similar performance to the ECM array on the ATHENA data. Therefore, MEMS are a viable low-cost alternative to high-cost ECMs for smart home applications.

The rest of this paper is organized as follows: Section 2 provides technical details for the MEMS microphone array; Section 3 describes the speech corpora used in this study; Section 4 presents the experimental procedure and results.

2. MEMS MICROPHONE ARRAY

Microphone arrays are currently being explored in many different applications, most notably for sound source localization, beamforming, and far-field speech recognition. However, the cost and the complexity of commercially available arrays often becomes prohibitively high for routine applications. Using multiple sensors in arrays has many advantages, but it is also more challenging: as the number of signals increases, the complexity of the electronics to acquire and process the data grows as well. Such challenges can be quite formidable depending on the number of sensors, processing speed, and complexity of the target application.

The newly developed technology of ultra-compact MEMS microphones [11] facilitates the integration of sound sensing elements with ambient intelligence environments. MEMS microphones have some significant advantages over ECMs: they can be reflow soldered, have higher "performance density" and less variation in sensitivity over temperature. Recent research has demonstrated that MEMS microphones are a suitable low-cost alternative to ECMs [12]. Since their cost can be as much as three orders of magnitude lower than ECMs, they present an attractive choice.

The microphones used in this research are the STMicroelectronics MP34DT01 [13]: ultra-compact, low-power, omnidirectional, digital MEMS microphones built with a capacitive sensing element

This research was supported by the EU project DIRHA with grant FP7-ICT-2011-7-288121.

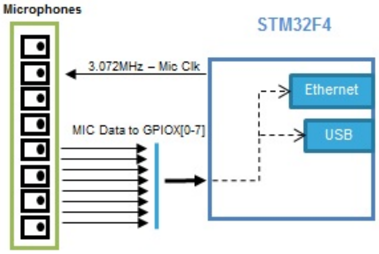


Fig. 1. The MEMS microphone array architecture developed for the DIRHA project [4] and used in this paper.

and an Integrated Circuit (IC) interface. The sensing element, capable of detecting acoustic waves, is manufactured using a specialized silicon micromachining process dedicated to the production of audio sensors. The MP34DT01 has an acoustic overload point of 120dB sound pressure level with a 63dB signal-to-noise ratio and 26dB relative to full scale sensitivity. The IC interface is manufactured using a CMOS process that allows designing a dedicated circuit able to provide a digital signal externally in pulse-density modulation (PDM) format, which is a high frequency (1 to 3.25MHz) stream of 1-bit digital samples.

Our architecture demonstrates the design of a MEMS microphone array with a special focus on low cost and ease of use. Up to 8 digital MEMS microphones are connected to an ARM© Cortex™-M4 STM32F4 microcontroller [14], which decodes the PDM of the microphones in order to obtain a pulse code modulation (PCM) and stream it using the selected interface (USB, Ethernet) (Fig. 1).

The PDM output of the 8 microphones is acquired in parallel by using the GPIO port of the STM32F4 microcontroller. The STM32F4 is based on the high-performance ARM© Cortex™-M4 32-bit RISC core operating at a frequency of up to 168MHz, it incorporates high-speed embedded memories (Flash memory up to 1Mb, up to 192Kb of SRAM), and it offers an extensive set of standard and advanced communication interfaces, like I2S full duplex, SPI, USB FS/HS, and Ethernet. The microphone's PDM output is synchronous with its input clock, therefore an STM32 timer generates a single clock signal for all 8 microphones.

The data coming from the microphones are sent to the decimation process, which first employs a decimation filter, converting 1-bit PDM to PCM data. The frequency of the PDM data output from the microphone (which is the clock input to the microphone) must be a multiple of the final audio output needed from the system. The filter is implemented with two predefined decimation factors (64 or 80), so for example, to have an output of 48kHz using the filter with 64 decimation factor, we need to provide a clock frequency of 3.072MHz to the microphone. Subsequently, the resulting digital audio signal is further processed by multiple stages in order to obtain 16-bit signed resolution in PCM format (Fig. 2). The first stage is a high pass filter designed mainly to remove the DC offset of the signal. It has been implemented via an IIR filter with configurable cutoff frequency. The second stage is a low pass filter implemented using an IIR filter with configurable cutoff frequency. Gain can be controlled by an external integer variable (from 0 to 64). The saturation stage sets the range for output audio samples to 16-bit signed.

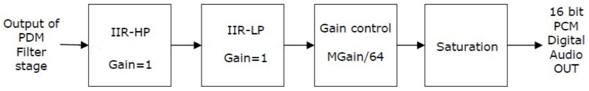


Fig. 2. Filtering pipeline used in the MEMS microphone array for converting each microphone data stream into a 16-bit PCM signal.

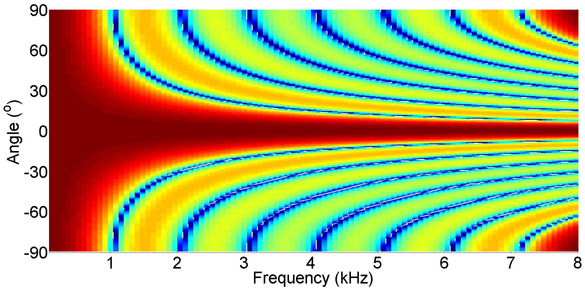


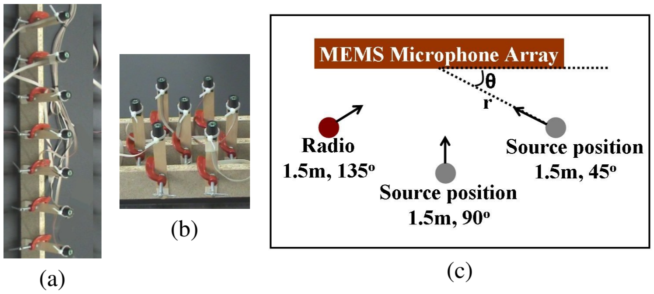
Fig. 3. Delay-and-sum beampattern of the 8-element MEMS microphone array in linear configuration with 42mm microphone spacing.

As already mentioned, the system allows data streaming via USB or Ethernet. When the USB output is selected and the device is plugged into a host, the microphone array is recognized as a standard multiple channel USB audio device. Therefore, no additional drivers need to be installed. Thus, the array can be interfaced directly with third-party PC audio acquisition software. The microphone array can be configured using a dip-switch in order to change the number of microphones (1 to 8) and the output frequency (16kHz to 48kHz). The delay-and-sum beampattern for a linear MEMS array of 8 elements with 42mm uniform spacing is shown in Fig. 3.

3. SPEECH CORPORA

3.1. MEMS microphone array corpus

To evaluate the effectiveness of MEMS microphone arrays and their geometric configuration for speech enhancement, a corpus containing multichannel recordings of real speech with various array configurations and in various noise conditions was collected. The speech data was recorded using a 7-element array of MEMS microphones resting on a flat desk. Speech was recorded for both linear and hexagonal array geometries (Fig. 4 (a) and (b), respectively). Linear array configurations are often used in practice, however hexagonal arrays possess, in theory, certain advantages [7], such as optimal spatial sampling [15–17]. Linear configurations with uniform microphone spacing of 4cm, 8cm, 12cm and hexagonal configurations with radius 8cm, 12cm, 16cm were used in the recordings. For each array configuration, speech was recorded for two frequently occurring in practice types of noise fields: diffuse and localized. The diffuse noise field arises in environments such as offices, cars, etc. [18, 19]. To generate a diffuse noise field in the recording room, computer and heater fans and air blowers were utilized. To generate a localized noise field, a single loudspeaker playing a radio program



(a)

(b)

(c)

Fig. 4. (a) Linear and (b) hexagonal MEMS array configurations. (c) Schematic of the recording setup for the MEMS array corpus: the two source positions (only one active source for each recording) and the position of the loudspeaker generating the localized noise field (not active for the diffuse noise field recordings) are shown.

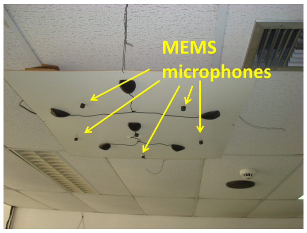


Fig. 5. ATHENA database setup: MEMS and ECM pentagons

was used. The loudspeaker was placed at a distance of $1.5\mathrm{m}$ at an angle of 135^o relative to the array center (Fig. 4 (c)). For each combination of array geometry and noise field, speech was recorded for two subject positions: angles 45^o and 90^o at a distance of $1.5\mathrm{m}$ relative to the center of the array (Fig. 4 (c)). Data was recorded for 6 subjects, 3 male and 3 female. For each combination of array geometry, noise field and subject position, each speaker, standing, uttered a total of 30 short command-like sentences, related to controlling a smart home such as the one being developed under the DIRHA project [4]. When standing, the speaker's head elevation was within $40-50\mathrm{cm}$ from the elevation of the plane where the MEMS microphones rested. Aside from the MEMS array, a close-talk microphone was used to capture a high SNR reference of the desired speech signal. All signals were recorded at a rate of $48\mathrm{kHz}$. In total, the corpus contains 4320 utterances, 720 per array configuration.

3.2. ATHENA Database

To compare MEMS microphones and ECMs, the ATHENA database was used [10]. This corpus contains 4 hours of speech from 20 speakers (10 male, 10 female) recorded in a realistic smart environment. To realistically approximate an everyday home scenario, speech (comprised of phonetically rich sentences, conversations, system activation keywords, and commands) was corrupted by both ambient noise and various background events. Data was collected from 20 ECMs distributed on the walls and the ceiling, 6 MEMS microphones, 2 close-talk microphones and a Kinect camera. The MEMS microphones formed a pentagon on the ceiling, close to a congruent ECM array (Fig. 5). More details can be found in [10]. For the experiments in the present paper, only the MEMS and ECM ceiling pentagon arrays were considered.

4. EXPERIMENTS AND RESULTS

4.1. Multichannel Speech Enhancement System

Microphone array data presents the advantage that spatial information is captured in signals recorded at different locations and can be exploited for speech enhancement through beamforming algorithms [5–8]. To further enhance the beamformer output, post-filtering is often applied. Commonly used optimization criteria for speech enhancement are the Mean Square Error (MSE), the MSE of the spectral amplitude [20] and the MSE of the log-spectral amplitude [21], leading to the Minimum MSE (MMSE), Short-Time Spectral Amplitude (STSA), and log-STSA estimators, respectively. All these estimators have been proven to factor into a Minimum Variance Distortionless Response (MVDR) beamformer followed by single-channel post-filtering [7, 22, 23].

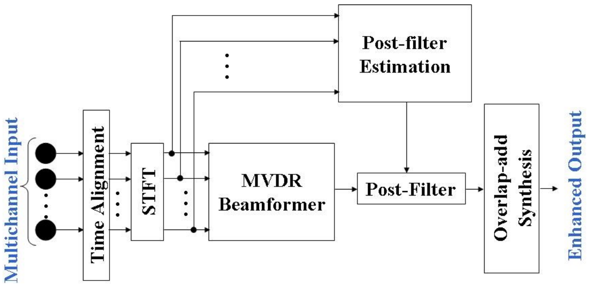


Fig. 6. The multichannel speech enhancement system reported in [9] and used in our experiments.

For our enhancement experiments, the multichannel speech enhancement system proposed in [9] is used. The system implements all aforementioned estimators using an optimal post-filtering parameter estimation scheme. Its structure is shown in Fig. 6.

The inputs to the system are the signals recorded at the various microphones, modeled as:

$$x_m(n) = d_m(n) * s(n) + v_m(n),\ m = 1,\,2,\,\ldots,\,N, \quad (1)$$

where n is the sample index, $*$ denotes convolution, $x_m(n)$ is the signal at microphone m, $s(n)$ is the desired speech signal, $d_m(n)$ is the acoustic path impulse response from the source to microphone m, and $v_m(n)$ denotes the noise. Assuming that reverberation is negligible, $d_m(n) = a_m\delta(n-\tau_m)$, where τ_m is the propagation time from the source to microphone m in samples.

In the employed algorithm, the input signals are first temporally aligned, to account for the propagation delays τ_m. Subsequently, due to the non-stationarity of speech signals, short-time analysis of the aligned input signals is employed, through a Short-Time Fourier Transform (STFT). The MVDR beamformer followed by the respective post-filter provide an implementation of one of the MMSE, STSA, or log-STSA estimators. Finally, the output signal is synthesized using the overlap and add method [24].

To estimate the MVDR weights and the post-filter parameters, the algorithm used requires prior knowledge of a model for the noise field. The spatial characteristics of noise fields are captured in the degree of correlation of the noise signals recorded by spatially separated sensors. Thus, to characterize noise fields, the complex coherence function defined as [25]:

$$C_{v_iv_j}(\omega) = \frac{\phi_{v_iv_j}(\omega)}{\sqrt{\phi_{v_iv_i}(\omega)\phi_{v_jv_j}(\omega)}}, \quad (2)$$

is often used, where ω denotes the discrete-time radian frequency and $\phi_{g_ig_j}$ the crosspower-spectral density between signals g_i and g_j. For the ideally diffuse and localized noise fields, the analytical form of the complex coherence function is known. For diffuse noise [25]:

$$C^{\mathrm{dif}}_{v_iv_j}(\omega) = \frac{\sin(\omega f_s r_{ij}/c)}{\omega f_s r_{ij}/c}, \quad (3)$$

where f_s is the sampling frequency, r_{ij} the distance between sensors i, j, and c sound speed. For localized noise [26]:

$$C^{\mathrm{loc}}_{v_iv_j}(\omega) = e^{-j\omega(\tau_{v_i}-\tau_{v_j})}, \quad (4)$$

where τ_{v_i} denotes the propagation time of the localized noise signal to microphone i. The algorithm further assumes that the noise field is homogeneous (the noise signal has equal power across sensors).

4.2. Experimental Results and Discussion

1) MEMS array corpus: The multichannel speech enhancement system described in Section 4.1 was used on the MEMS array corpus. For time alignment, to calculate propagation delays, ground truth was used for source and microphone positions. Having no dependency on the accuracy of a localization module renders the results

MEMS array corpus

			Linear geometry			**Hexagonal geometry**		
Noise field	Speaker Position (r, θ) in $(\mathrm{m}, ^o)$	MVDR post-filter	Sensor spacing			Sensor spacing		
			4cm	8cm	12cm	8cm	12cm	16cm
Diffuse	$(1.5, 90^o)$	MMSE	4.90	4.52	4.19	**7.49**	6.99	5.66
		STSA	4.85	4.46	4.12	**7.20**	6.73	5.41
		log-STSA	4.90	4.51	4.17	**7.36**	6.87	5.56
	$(1.5, 45^o)$	MMSE	3.48	2.87	1.99	**4.13**	3.75	3.12
		STSA	3.44	2.80	1.95	**4.00**	3.63	3.01
		log-STSA	3.48	2.85	1.98	**4.07**	3.70	3.07
Localized	$(1.5, 90^o)$	MMSE	1.20	1.00	0.88	**3.36**	3.18	2.73
		STSA	1.18	0.98	0.83	**2.83**	2.82	2.35
		log-STSA	1.19	0.99	0.86	**2.99**	2.96	2.50
	$(1.5, 45^o)$	MMSE	**6.61**	3.86	2.85	4.58	3.77	3.94
		STSA	**6.28**	3.44	2.47	3.44	3.14	3.04
		log-STSA	**6.45**	3.61	2.62	3.75	3.36	3.29

ATHENA database

Sensor Type	SSNRE (dB)
ECM	2.09
MEMS	2.05

Table 1. Speech enhancement on MEMS array (left) and ATHENA (right) corpora. All results are reported in SSNR enhancement in dB.

comparable across array geometries in terms of speech enhancement performance alone. To calculate the STFT, 1200-sample (25ms) Hamming windows with 900-sample overlap (18.75ms) were used. The noise field generated by fans was modeled as diffuse (Eq. (3)), while the noise field generated by the loudspeaker was modeled as ideally localized (Eq. (4)). Ground truth parameter values were used to calculate the complex coherence function in each case.

To evaluate the quality of the enhanced output of the system, the Segmental Signal to Noise Ratio (SSNR) [27] was used, which has been shown to have better correlation with the human perceptual evaluation of speech quality than global SNR. Frame SNRs were restricted to $(-15\mathrm{dB}, 35\mathrm{dB})$ before calculating the SSNR [27].

The results, in terms of average SSNR Enhancement (SSNRE) across utterances recorded under the same conditions, are presented in Table 1. For each utterance, the SSNRE is calculated as the dB difference between the output and the mean of the input SSNRs.

Overall, significant improvements in speech quality are obtained using the MEMS microphone array. The hexagonal geometry with $8\mathrm{cm}$ radius achieves about $7.5\mathrm{dB}$ average SSNRE for the diffuse noise field, while about $6.5\mathrm{dB}$ average SSNRE is observed for the linear geometry with $4\mathrm{cm}$ in the case of the localized noise field, with the desired speech source at 45^o.

In general, the hexagonal array geometry performs better than the linear one. In detail, for the diffuse noise field, the best result for a hexagonal geometry ($7.49\mathrm{dB}$ with $8\mathrm{cm}$ radius) is approximately $2.5\mathrm{dB}$ higher than the best result achieved by a linear geometry ($4.90\mathrm{dB}$ with $4\mathrm{cm}$ sensor spacing). This can be attributed to the linear array configuration having axial symmetry, which renders it impossible to differentiate among signals traveling from the far-field to the array along the same cone. Such signals have the same propagation delays τ_m and are indiscriminable. In a diffuse noise field, signals of equal power propagate from all spatial directions, so the linear array is at a disadvantage. For the localized noise field, the best performance of $6.61\mathrm{dB}$ is achieved by the linear array with $4\mathrm{cm}$ spacing for speaker position at 45^o. However, the hexagonal geometries with radii $12\mathrm{cm}$ and $16\mathrm{cm}$ produce superior results compared to the linear ones with $8\mathrm{cm}$ and $12\mathrm{cm}$ spacing, respectively. Namely, with sparser sampling of the acoustic field, the hexagonal geometries still outperform the linear. Also, for talker positioned at 90^o the hexagonal geometry produces superior results overall.

Intuitively, the superior performance of the hexagonal array geometry can be explained by considering the advantages of sampling the spatial field with a hexagonal grid. It has been shown that hexagonal array sampling requires the least amount of samples to completely characterize a space-time field [7, 15–17]. Therefore, given a number of sensors, it is expected that the hexagonal array can capture more spatial information than the linear one. Also, the hexagonal array can capture the same amount of spatial information with sparser sampling of the spatial field (larger sensor spacing).

For a given geometry, performance deteriorates as spatial sampling becomes sparser. By the spatial sampling theorem, larger sensor spacing decreases the maximum frequency that the array can spatially resolve [7], yielding worse performance.

2) ATHENA database: To compare MEMS and ECM arrays, the speech enhancement system was used on the ATHENA ceiling pentagonal arrays data. The use of a pentagon array was motivated by the superior performance of hexagonal-type arrays observed in the MEMS array corpus experiments. Ground truth was used for source and microphone locations for the same reason as in the MEMS array corpus case. For STFT calculation, window length and overlap was the same as for the MEMS corpus. Noise was modeled as diffuse, as a multitude of background noises occur in each session [10]. Results in terms of average SSNRE across the database for each microphone type are presented in Table 1. For each utterance, the SSNRE is calculated as the dB difference between the output and the central microphone SSNR. The performance of the low-cost MEMS array is comparable to the expensive ECM array with a very small decrease of $0.04\mathrm{dB}$ in average SSNRE. Therefore, MEMS arrays are a viable low-cost alternative to ECM arrays.

5. CONCLUSIONS AND FUTURE WORK

Using MEMS microphones, very satisfactory speech enhancement performance was observed ($7.49\mathrm{dB}$ best SSNRE on the MEMS corpus). The comparison of array geometries revealed superior performance of the hexagonal array, which can be attributed to optimality of hexagonal grid sampling. The comparison of pentagonal MEMS and ECM arrays in a realistic smart home environment revealed no significant difference in performance. MEMS microphones are low-cost, compact, portable, and easy to configure in any geometry. Combined with good speech enhancement performance in challenging conditions, comparable to that of bulky and expensive ECMs, these attributes make them attractive for smart home applications.

In future work, we plan to investigate MEMS microphone performance for other multichannel processing problems, such as time-delay of arrival estimation and source localization, for which robust methods are known in the literature [28–31].

6. REFERENCES

[1] M. Chan, E. Campo, D. Estève, and J.-Y. Fourniols, "Smart homes – current features and future perspectives," *Maturitas*, vol. 64, no. 2, pp. 90–97, 2009.

[2] A. Waibel, R. Stiefelhagen, et al., "Computers in the human interaction loop," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J.C. Augusto, Eds., pp. 1071–1116. Springer, 2010.

[3] "AMI: Augmented Multi-party Interaction," [Online]. Available: `http://www.amiproject.org`.

[4] "DIRHA: Distant-speech Interaction for Robust Home Applications," [Online]. Available: `http://dirha.fbk.eu/`.

[5] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech and Signal Process. Mag.*, vol. 5, pp. 4–24, 1988.

[6] M.S. Brandstein and D.B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[7] H.L. Van Trees, *Optimum Array Processing*, Wiley, 2002.

[8] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1, Springer, 2008.

[9] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7, pp. 657–666, 2007.

[10] A. Tsiami, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis, G. Potamianos, and P. Maragos, "ATHENA: A Greek multisensory database for home automation control," in *Proc. Interspeech*, 2014.

[11] J.J. Neumann Jr. and K.J. Gabriel, "A fully-integrated CMOS-MEMS audio microphone," in *Proc. Int. Conf. on Transducers, Solid-State Sensors, Actuators and Microsystems*, 2003.

[12] E. Zwyssig, M. Lincoln, and S. Renals, "A digital microphone array for distant speech recognition," in *Proc. ICASSP*, 2010.

[13] STMicroelectronics, *MP34DT01 MEMS audio sensor omnidirectional digital microphone datasheet*, 2013, [Online]. Available: `http://www.st.com/web/en/resource/technical/document/datasheet/DM00039779.pdf`.

[14] STMicroelectronics, *DS8626 - STM32F407xx datasheet*, 2013, [Online]. Available: `http://www.st.com/web/en/resource/technical/document/datasheet/DM00037051.pdf`.

[15] D.P. Petersen and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in n-dimensional Euclidean spaces," *Information and Control*, vol. 5, no. 4, pp. 279–323, 1962.

[16] R.M. Mersereau, "The processing of hexagonally sampled two-dimensional signals," *Proc. of the IEEE*, vol. 67, no. 6, pp. 930–949, 1979.

[17] D.E. Dudgeon and R.M. Mersereau, *Multidimensional Digital Signal Processing*, Prentice-Hall, 1984.

[18] J. Meyer and K.U. Simmer, "Multichannel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. ICASSP*, 1997.

[19] I.A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[20] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[21] Y. Ephraim and D. Mallah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[22] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Eds., pp. 39–60. Springer, 2001.

[23] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2002.

[24] L.R. Rabiner and R.W. Schafer, *Digital Signal Processing of Speech Signals*, Prentice Hall, 1978.

[25] G.W. Elko, "Spatial coherence function for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Eds., pp. 61–85. Springer, 2001.

[26] S. Doclo, *Multi-microphone noise reduction and dereverberation techniques for speech applications*, Ph.D. thesis, Katholieke Universiteit Leuven, 2003.

[27] J.H.L. Hansen and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998.

[28] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[29] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP*, 1994.

[30] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Proc. Interspeech*, 2005.

[31] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Eds., pp. 157–180. Springer, 2001.