Enhanced Sound Capture System for Small Devices

Slavy G. Mihov[1] Tyler Gleghorn[2], Ivan Tashev[2]

***Abstract* – With mass propagation of the cellular phones and other small form factor devices as PDAs and other handhelds their usage in noise adverse environment is substantially increased. With adoption of the 3G and 4G wireless technologies the transition to videophone mode of communication is imminent. In addition most of the modern mobile phones have integrated cameras and are able to record short videos. In all these cases we have worse quality of the captured sound due to the old paradigm of a single microphone, which is supposed to be positioned close to the mouth of the human speaker. In this paper we propose enhanced sound capturing system for mobile devices. It consists of two directional microphones, pointing in opposite directions. Using a novel approach for the beamformer design we achieve satisfactory sound quality levels. The system output improves the perceptual sound quality with 0.29 MOS points and the SNR with 15 dB.**

***Keywords* – sound capture, mobile devices, microphone array, beamforming**

I. INTRODUCTION

Mobile devices are increasingly being used in situations that require hands-free communication. As a result, mobile phone users are now using headsets with their telephones. Despite the option of using either wired or Bluetooth wireless headsets, for reasons of comfort, convenience and style, most users prefer to use their handhelds without any headsets. In camcorder mode the device is supposed to capture sound sources from distances of 1-3 meters. In these usage modes the sound source is located at some distance from the microphone. This positioning is suboptimal, and when compared to a well-placed close-talking microphone, yields a significant decrease in the Signal-to-Noise Ratio (SNR) of the captured speech signal. Considering the fact that most users operate their phones in noisy environments, the decrease in SNR of the captured speech signal leads to inability to use the device.

One way to improve the quality of the sound capture system is to use multiple microphones configured as an array instead of a single one. Microphone array processing improves the SNR by spatially filtering the sound, in essence pointing the array beam toward the signal of interest, which improves the overall sound quality due to better directivity [1]. The use of multiple, spatially separated, microphones allows performing spatial filtering along with conventional temporal filtering, which can better reject the interference signals, resulting in an overall improvement of the captured sound quality. Microphone array algorithms jointly process the signals from all microphones to create a single-channel output signal with higher SNR compared to a single microphone. In practice, beamforming algorithms have their limitations, so an adaptive post-filter is typically applied to the array output in order to provide additional noise reduction [1, 2].

Incorporating a microphone array into a handheld device presents a unique set of challenges. For example, conventional methods of far-field beamforming, e.g. [1, 2], can't be directly applied because the distance between the elements of the microphone array tends to be too small. In addition, size, power, and cost requirements limit the number of used microphones. We have non-typical requirements for the desired beamshape. The device should capture well sounds coming from front (in telephone mode) and back (in video camera mode).

With small number of microphones, the performance of any beamforming algorithm will be limited. In the microphone array processing, the phase difference between the signals received from a pair of microphones gives indication for the direction of arrival (DOA) of a given sound source [8, 9]. In small microphone arrays, with very small distance between the microphones, the phase difference decreases and is masked by the ambient and instrumental noises, which results in losing DOA information for the sound source. In such arrays, where the microphone elements are too close, determining DOA depends thoroughly on microphone directivities instead of time delay of signal arrival.

This paper describes an enhanced sound capturing system for mobile devices. It is based on using directional microphones and non-trivial approach to design the beamformer weights. A time invariant beamformer with low CPU requirements provides well audible improvement in the perceptual sound quality measured and in SNR, compared to single microphone case.

II. MODELLING

The enhanced sound capture system for handheld devices consists of two unidirectional microphones, positioned back-to-back, on both sides of the device pointing in opposite directions (Fig. 1). The microphones form two-element microphone array, which can capture and process sounds from both front and rear directions.

This two element microphone array is used to improve the Signal-to-Noise Ratio (SNR) with its spatial selectivity. As considered in [3, 4], an array of M microphones has known positions of its elements, determined by vector p; the sensors sample the signal field at locations $p_m = (x_m, y_m, z_m) : m = 0, 1, \ldots, M-1$. This yields a set of signals that we denote by the vector $x(t, p)$. Each sensor m has known directivity pattern $U_m(f,$

[1] Slavy G. Mihov is with the Faculty of Electronic Engineering and Technologies in Technical University – Sofia, Bulgaria, Sofia 1000, Kliment Ohridski St. 8, e-mail: smihov@tu-sofia.bg. Work done while being intern in Microsoft Research.

[2] Tyler Gleghorn, Ivan Tashev – Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA, e-mail: {tylerg, ivantash}@microsoft.com

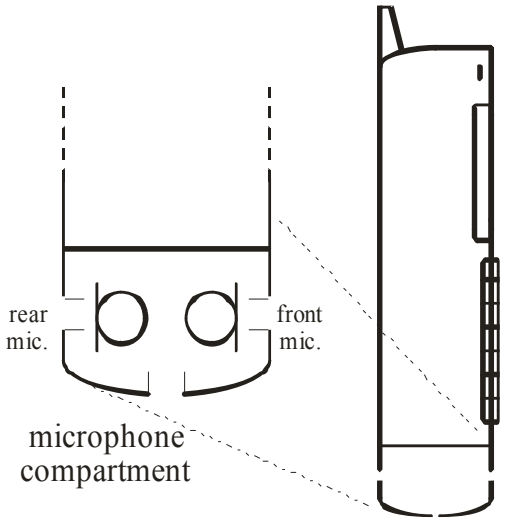

Fig. 1. Microphone configuration in a handheld device

c), where $c = \{\varphi, \theta, \rho\}$ represents the coordinates of the sound source in radial coordinate system. The coordinates can also be represented in a rectangular coordinate system, $c = \{x, y, z\}$. The microphone directivity pattern is a complex function, providing the spatio-temporal transfer function of this channel. For an ideal omni-directional microphone $U_m(c,f) = const$. The microphone array can have microphones of different types, so $U_m(c, f)$ can vary as a function of m. Even for microphones of same type, $U_m(c, f)$ can vary, due to manufacturing tolerances and constructional peculiarities of the array.

For adequate usage of the microphones in a microphone array for beamforming is essential to have precise models of their directivity patterns. The models consist of analytical or measured expression of the microphone gain as function of the frequency and the incident angle. As in most of the cases analytical form is either complex or not precise, we measured the directivity patterns of the used microphones. We recorded a chirp signal in an anechoic chamber with the device prototype to determine the directivity patterns of each of its microphones. The radial position of the speaker towards the device was changed and a record was captured every 10 degrees.

With this set of 36 experimental records we were able to determine the directivity patterns of the microphones in our array configuration, using interpolation for the transitional incident angles. The estimated directivity diagram for one of the microphones for 1000 Hz is shown in Fig. 2. The microphone directivity index is DI = 3.5 dB, and results in 3.9 dB noise suppression for signal coming along the main response axis. Microphone's magnitude response as function of the signal frequency and the incident angle is shown in Fig. 3. In general it is a subcardioid directivity pattern with a slope towards the lower part of the frequency band.

III. Beamformer Design

The use of microphone arrays has been extensively studied in the literature because of their effectiveness in enhancing the quality of the captured audio signal in scenarios where the use of a close talking microphone is undesirable, but high quality

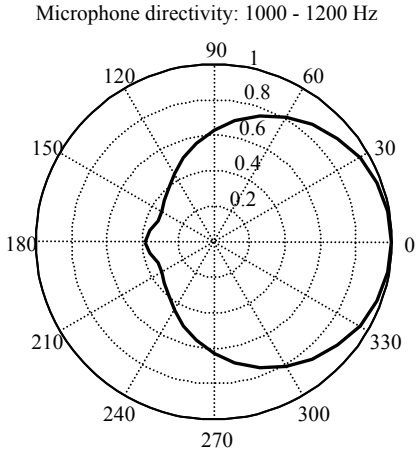

Fig. 2. Microphone directivity pattern

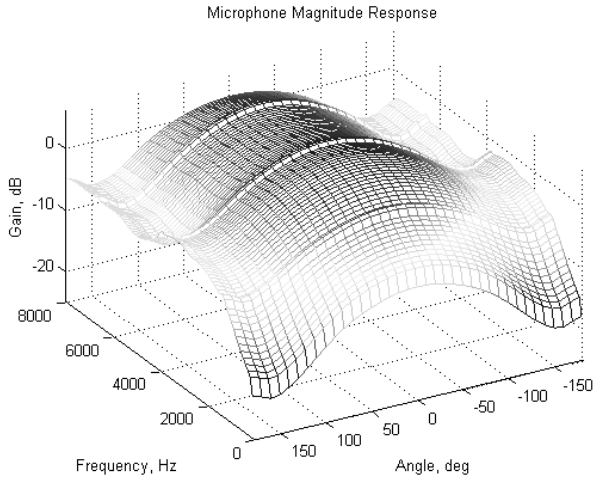

Fig. 3. Microphone magnitude response

audio is a critical component [4]. To improve the audio quality, beamforming and noise removal filtering algorithms are frequently applied. Microphone array beamforming is a technique used to "aim" the microphone array in an arbitrary direction to enhance the SNR. Such processing is linear and doesn't introduce distortions or artifacts, such as musical noise [5]. For computational efficiency and low latency (compared to adaptive filters), can be used delay-and-sum beamforming [2]. Due to its higher directivity the beamformer also reduces the reverberation in the captured audio, which improves its quality as well.

The mobile device in our case uses a fixed, time invariant beamforming approach and benefits from its advantages: the desired beam is designed off-line, and then a computationally efficient code is used to process signals in real time. This results in low CPU power requirements in run-time.

Assuming that audio signal is processed in frames longer than twice the period of the lowest work band frequency, combining the signals from all sensors is just a weighted sum:

$$Y(f) = \sum_{m=1}^{M-1} W_m(f) X_m(f) \quad (1)$$

where $W_m(f)$ are the frequency-dependent weights vector for each sensor m and $Y(f)$ is the beamformer output. In real systems the set of vectors $\mathbf{W}$ (f) is an $N \times M$ complex matrix, where N is the number of frequency bins in a discrete-time filter bank, and M is the number of microphones. For each set of weights $\mathbf{W}(f)$, there is a corresponding beam shape $B(c, f)$, which is the beamformer complex gain as function of the sound source position:

$$B(f,c) = \sum_{m=1}^{M-1} W_m(f) D_m(f,c) U_m(f,c) \tag{2}$$

were, $D_m(f, c)$ represents the delay and the decay due to the distance to the microphone [4]. The beam shape function represents the beamformer directivity.

Designing the microphone array beamformer means to calculate an optimal, in one or another way, matrix of weights $W_m(f)$ in (1). One of the criteria for optimality can be the weights to provide maximal noise suppression, i.e. they minimize the noise level in the output signal. Another is a specific shape of the beam directivity.

In our particular two-element microphone array, the beamforming combination of the input signals (1) takes the form:

$$\begin{aligned} Y_F^{(n)}(k) &= W_{FF}(k).X_F^{(n)}(k) + W_{FR}(k).X_R^{(n)}(k) \\ Y_R^{(n)}(k) &= W_{RF}(k).X_F^{(n)}(k) + W_{RR}(k).X_R^{(n)}(k) \end{aligned} \tag{3}$$

where the indexes F and R denote the two opposite microphones (Front and Rear), n is the frame number, and k is the frequency bin number. In the need of signal enhancement, regarding the typical usage scenarios of the handheld device, we want to capture from both front and rear directions, forming beamshape like figure 8. "Figure-8" criterion for the beam design maximizes signal captured from both directions (front and rear) and naturally suppresses sounds coming from the sides – usually ambient noise. The analytic expression, describing the beam design criterion is:

$$Q_{Fconst} = \max_{W_{FF}, W_{FR}} \left(\frac{\int_{\theta \in L} \left(W_{FF} X_F(\theta) + W_{FR} X_R(\theta) \right) d\theta}{\int_{\theta \in S} \left(W_{FF} X_F(\theta) + W_{FR} X_R(\theta) \right) d\theta} \right) \tag{4}$$

where θ is the incident angle of signal source and L and S denote the listening and suppression areas. For "Figure 8" beam we choose L to be $\pm\Delta\theta$ around directions 0^O and 180^O. For suppression area we choose $\pm\Delta\theta$ around 90^O and 270^O.

The problem of maximizing the criterion above should be solved as an optimization task, under constraints of unit gain and zero phase shift for sounds coming from the focus points for the working frequency band [3]. This leads to typical nonlinear constrained minimization problem. The constraints can be added as punishing functions, converting the constrained minimization problem to non-constrained. After this a well known optimization method, as steepest gradient descent, can be used to find the solution.

Designing the "Figure-8" beam is a one-time optimization process of determining the beamformer weights according to the criterion above. Once determined these weights are used as-is during normal operation of the handheld device.

IV. RESULTS

The beamforming design technique presented in the previous sections is an off-line design procedure, which produces the corresponding near-optimal beam weights for maximum signal capture from both front and rear. Those are used in the real-time processing engine in the following way: for each incoming frame containing N signal samples from each microphone, we compute the short time spectra by weighting and converting to frequency domain (e.g. using FFT or the modulated complex lapped transform [6]). We then apply the optimal weights using (3) and compute the output signal spectrum. We then use the standard overlap and add procedure to generate the time domain signal. Note that the computational complexity of (3) is low and the off-line beam design lowers the run-time CPU requirements.

Normal usage of the handheld device (for example, in a camcorder mode) suggests necessity for best signal capture of sound coming from front and rear of the device. The optimization procedure generated weights forming "Figure-8" beam shape. Its directivity pattern (Fig. 4 shows it for 1000 Hz) resembles digit 8, whence its name comes. The magnitude response (Fig. 5) shows no frequency dependence.

For evaluation of the performance of the so designed beam-

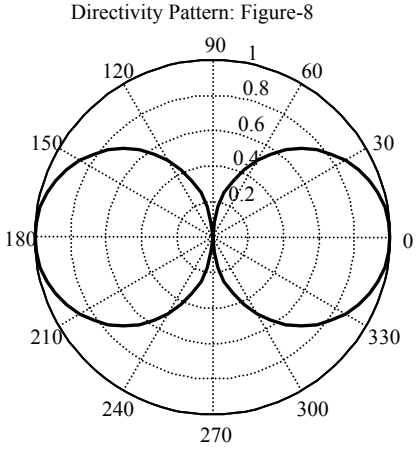

Fig. 4. Beam directivity pattern

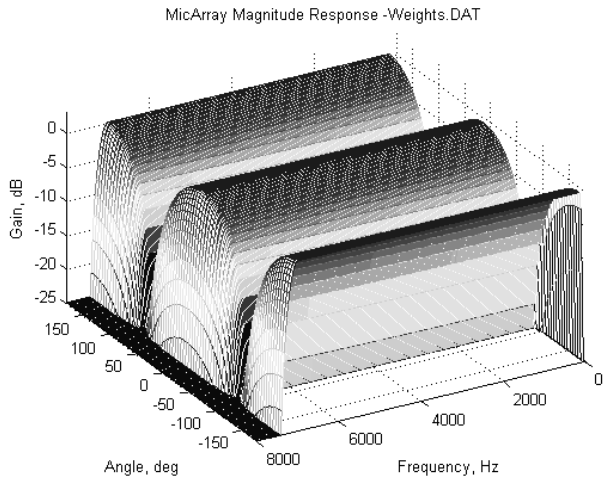

Fig. 5. Beam magnitude response

former, we used an evaluation set of test records, containing human speech in various scenarios and noise conditions. The evaluation set consists of 16 three channel files, containing the signals captured from the front and rear microphones of the device and a reference channel, captured with a close talk microphone. Each recording is 45 seconds long. These records contain either male or female voices, contaminated with two different types of noise (stationary office noise and human speech in the background). The scenarios of speech source positions and noise type mimic the most common hands-free modes for use of the mobile device without a headset:

- Speaker one meter in front of the handheld device (telephone mode);
- Speaker one meter in rear of the device (videophone mode, recording near user speaking);
- Speaker five meters in rear of the device (camcorder mode, recording far user speaking);
- Two alternating sources – five meters in front and one meter in rear of the device, respectively (camcorder mode, recording a dialog).

The performance of the designed beamformer was evaluated in comparison with single microphone mode. In this mode is used the most appropriate microphone (front or rear) without any processing.

In our evaluation of the beamformer performance, two metrics were used. The first is Signal-to-Noise Ratio (SNR), which gives the proportion of the wanted and unwanted signals. For classification of the frame as "signal" or "noise", the reference channel was used. The SNR is the proportion of the averaged energy during the "signal" and "noise" frames. This metric gives an indirect estimate of the sound quality. Mean Opinion Score (MOS) – ITU-T P.800 was used as a primary metric for the quality of the output signal after processing. This is a dimensionless quantity with values ranging from 1 to 5. It gives an estimate of human perception of sound quality.

Estimating MOS with real humans is long and expensive procedure, involving many humans listening to the records and giving their subjective opinion. For this reason, MOS is not suitable for use during the stage of algorithm development. We used objective Perceptual Evaluation of Sound Quality (PESQ) – ITU-T P.862. It produces similar results to MOS results in the same scale (from 1 to 5) to give an estimate of human perception of sound quality too. We used the MatLab implementation of PESQ algorithm [7] which requires reference channel.

Evaluating the entire test set records with our "Figure-8" beamformer in comparison with single microphone mode gives the average values shown in Table I. The particular results for each of the test case scenarios vary, having different contribution to the average improvement shown. The scenarios in which the sound source is at a distance of 1 m from the device (in front and in rear) show almost no difference between single microphone and "Figure-8" modes in MOS, but significant improvement in SNR (due to higher level of ambient noise suppression). For the rest of the scenarios, "Figure-8" shows substantial improvement in both metrics. The average improvement in SNR due to beamformer usage only is 14.95 dB. The primary MOS metric shows 0.29 MOS points improvement in human perception of sound quality.

TABLE I
AVERAGE EVALUATION RESULTS

Mode	MOS	SNR [dB]
Single Microphone	2.216	5.88
Figure-8	2.506	20.82
Improvement	0.290	14.95

V. DISCUSSION

The need to present clean sound inputs to today's real-time communication and speech recognition engines has fostered large amount of research in the areas of noise suppression, microphone array processing, acoustic echo cancellation and methods for reducing the effects of acoustic reverberation.

In this paper, we described a microphone array for handheld devices, consisting of two directional microphone elements pointing to opposite directions. The microphones are used in an array configuration, which after processing forms a "Figure-8" beam, optimized for maximum signal capture from front and rear of the device. This enhanced sound capture system was evaluated in close to real conditions and gave well audible improvement in MOS (0.29 points) and in SNR (14,95 dB). The cost of this improvement is usage of one more microphone, some CPU power and operational memory for processing the algorithmic computations.

In general, our technique achieves improvement in signal quality with reasonable resources. Further improvement in captured signal quality can be achieved by combining suitable pre-processing and post-filtering algorithms in addition to the microphone array beamformer.

REFERENCES

[1] H. Van Trees, *Detection, Estimation and Modulation Theory, Part IV: Optimum array processing*, Wiley, New York, 2002.

[2] M. Brandstein and D. Ward, *Microphone Arrays*, Springer-Verlag, Berlin, 2001.

[3] I. J. Tashev and H. S. Malvar, "A New Beamformer Design Algorithm for Microphone Arrays", *ICASSP 2005*, Philadelphia, March 2005.

[4] I. J. Tashev, M. L. Seltzer and A. Acero, "Microphone Array for Headset with Spatial Noise Suppressor", *IWAENC 2005*, Eindhoven, September 2005.

[5] I. J. Tashev, "Gain Calibration Procedure for Microphone Arrays", *ICME 2004*, Taipei, June 2004.

[6] H. S. Malvar, "A Modulated Complex Lapped Transform and its Applications to Audio Processing", *ICASSP 1999*, Phoenix, pp. 1421-1424, March 1999.

[7] P. C. Loizou, *Speech Enhancement Theory and Practice*, Taylor & Francis Ltd, ISBN-13: 978-0849350320, 1st edition (7 Jun 2007).

[8] X. Zhang and Y. Jia, "A Soft Decision Based Noise Cross Power Spectral Density Estimation for Two-Microphone Speech Enhancement Systems", *ICASSP 2005*, Philadelphia, March 2005.

[9] C. Lai and P. Aarabi, "Multiple-Microphone Time-Varying Filters for Robust Speech Recognition", *ICASSP 2004*, Montreal, May 2004.