# DATA DRIVEN BEAMFORMER DESIGN FOR BINAURAL HEADSET

*Ivan Tashev and Michael L. Seltzer*

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{ivantash, mseltzer}@microsoft.com

## ABSTRACT

Mobile devices are being used in more and more adverse noise environments. This increases the requirements for designing headsets for these devices. Regardless of the limitations for larger battery life some designs use multiple microphones to achieve appropriate noise suppression. Unfortunately the most common approaches for designing the beamformer do not provide good results due to the complex way that the sound generated by the mouth travels around the head to reach the area around the ear, where the headset is usually positioned. In this paper we propose a data driven approach for designing a time invariant beamformer using set of calibration files. The approach is illustrated with the evaluation of a beamformer for binaural headset. As evaluation criteria are used the improvements in output Signal to Noise Ratio (SNR) and objective evaluation of the perceptual sound quality. The proposed design approach delivers 6.8 dBC improvement in SNR and 0.46 MOS points improvement in sound quality, compared to a single microphone.

**Index Terms** — Sound Capture, Headsets, Microphone Array, Beamforming

## 1. INTRODUCTION

Mobile phones and other mobile and small form factor devices are increasingly being used in environments with high noise levels. As a result, mobile device users are using headsets with their telephones. Users have the option of either a wire or Bluetooth wireless technology to connect their headset to the device. For reasons of comfort, convenience, and style, most users prefer headset designs that are compact and lightweight. Typically these designs require the microphone to be located at some distance from the user's mouth. 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. Many mobile phones already contain enough memory and are used as portable media players as well. To provide good listening quality users wear dual earphone headsets, in most of the cases placed at the beginning of the ear canal. Such unobtrusive design practically excludes using even a short boom for the microphone, which places it further away from the mouth. Placing the microphone near the cheek as in short boom headsets leads to 6-7 dB decrease of the SNR, while placing it in or around the ear (integrated with the headphone) leads to a drop of 10 to 14 dB.

One way to improve the sound capture performance of the headset is to use multiple microphones configured as an array. Microphone array processing improves the SNR by spatially filtering the sound field, in essence pointing the array toward the signal of interest, which improves overall directivity [1].

Incorporating a microphone array into a headset presents a unique set of challenges. For example, conventional methods of far-field beamforming, e.g. [1][2], cannot be directly applied because the user's head is located in the path between the sound source (the mouth) and the array [3]. In addition, size, power, and cost requirements limit the number of used microphone elements, typically to two, in rare cases three.

Placing the microphone array on the head means that it will move and rotate together with the user's head. Even if the noise sources do not move, their position in the microphone array coordinate system will change due to the head movements. This, together with the restrictions in the available CPU power due to limitations for the battery life, excludes using of advanced adaptive algorithms. They place a null towards the noise source, improving the SNR, but need time to adapt. Frequent movements will reduce the gain from the adaptive algorithm. Two microphones limit the ability to place null while maintaining the constraint for unit gain and zero phase shift for the direction towards the mouth.

The design of a time invariant beamformer assumes isotropic ambient noise – usually the worst case scenario. Most of the designs assume channel matching as well. In the headset case we have additional factors, which complicate the design [4]:

- The mouth has frequency dependent directivity pattern – it is less directional in the lower part of the frequency band, has higher directivity for frequencies above 1000 Hz.
- The sound, emitted by the mouth, warps around the head in different way for different frequencies.
- The frequency distortions are position dependent, which requires different corrections for each position.

In [5] authors mitigate the effect of the position dependent frequency distortions by applying correction filters in front of each of the three microphones of the small three element microphone array for headset. They used the algorithm from [6] to design a conventional time invariant beamformer. The correction filters are measured using the actual microphone array and a close talk microphone as reference channel.

In this paper we take this approach one step further and design the entire beamformer based on the reference channel. The second chapter of this paper reviews the time invariant beamformer. The third describes the design procedure. The experimental results are presented in fourth; they are discussed in the fifth chapter.

## 2. TIME INVARIANT BEAMFORMING

Consider an array of M microphones with known positions. The sensors sample the sound field at locations $p_m = (x_m, y_m, z_m)$ where $n = \{1,\ldots,M\}$ is the microphone index. Each of the $m$ sensors has known directivity pattern $U_m(f,c)$, where $f$ is the frequency and $c$ represents the location of the sound source in either radial or rectangular coordinate system. The microphone directivity pattern is a complex function, providing the spatio-temporal transfer function of the channel. For an ideal omnidirectional microphone, $U_m(f,c)$ is constant for all frequencies and source locations. A microphone array can have microphones of different types, so $U_m(f,c)$ can vary as a function of $m$.

In this study, we process the frequency bins independently. Accordingly, for a sound source $S_{c_T}(f)$ at a location $c_T$, the signal captured by each microphone can be represented as:

$$X_m(f, p_m) = D_m(f, c_T) S_{c_T}(f) \tag{1}$$

where $D_m(f,c)$ represents the delay and the decay due to the distance between the source and the microphone. This is expressed as

$$D_m(f, c_T) = F_m(f, c_T) \frac{e^{-j\frac{2\pi f}{v}\|c_T - p_m\|}}{\|c_T - p_m\|} U_m(f, c_T) A_m(f), \tag{2}$$

where $v$ is the speed of sound and $F_m(f,c)$ represents the spectral changes in the sound due to the directivity of the human mouth and the diffraction caused by the user's head. In headset case, the signal decay due to energy losses in the air can be ignored. The term $A_m(f)$ in equation (1) is the frequency response of the system preamplifier and analog-to-digital conversion (ADC). In most cases we can use the approximation $A_m(f) \equiv 1$ for the work band. Assuming that the audio signal is processed in frames longer than twice the period of the lowest frequency in the frequency band of interest, the signals from all sensors are combined using a filter-and-sum beamformer as:

$$Y(f) = \sum_{m=1}^{M} W_m(f) X_m(f), \tag{3}$$

where $W_m(f)$ are the weights for each sensor $m$ and frequency $f$, and $Y(f)$ is the beamformer output. Throughout this paper the frame index will be omitted for simplicity. The set of all coefficients $W_m(f)$ is stored as an $N$x$M$ complex matrix $\mathbf{W}$, where $N$ is the number of frequency bins in a discrete-time filter bank, and $M$ is the number of microphones.

The beampattern $B(f,c)$ in the free field is given by

$$B(f, c) = \sum_{m=1}^{M} W_m(f) D_m(f, c) \tag{4}$$

which leads to isotropic ambient (correlated) noise suppression:

$$G_{AN}(f) = \frac{|B(f, c_T)|^2}{\frac{1}{4\pi}\int_V |B(f, c)|^2 \, dc}. \tag{5}$$

Here $V$ is the set of coordinates on a sphere around the microphone array with radius the average distance to the noise sources. The instrumental (uncorrelated) noise suppression is given by:

$$G_{IN}(f) = \sqrt{\sum_{m=1}^{M} |W_m(f)|^2}. \tag{6}$$

For given ambient and uncorrelated noise variances $\lambda_A(f)$ and $\lambda_I(f)$ respectively, the output noise variance is:

$$\lambda = G_{AN}^2 \lambda_A + G_{IN}^2 \lambda_I, \tag{7}$$

Frequency indices are omitted for simplicity.

## 3. DATA DRIVEN BEAMFORMER DESIGN

Time invariant beamformer design is the process of finding a matrix $\mathbf{W}$ that is optimal in one or another way. The design process happens before using the microphone array and the weights remain unchanged during the actual signal processing. In real conditions ambient and instrumental noises should be added to the equation (1). They usually are modeled as zero mean Gaussian random variables:

$$X_m(f, p_m) = D_m(f, c_T) S_{c_T}(f) + \mathbb{N}(0, \lambda_A) + \mathbb{N}(0, \lambda_I). \tag{8}$$

Then the design goal for classic minimum variance distortionless response (MVDR) beamformer can be presented as minimizing the noise in the output under the constraint of unit gain and zero phase shift towards the listening direction:

$$\begin{gathered} \mathbf{W}_{c_T}(f) = \arg\min_{\mathbf{W}_{c_T}(f)} \lambda(\mathbf{W}_{c_T}, f) \\ \text{subject to } \mathbf{W}_{c_T}(f)\mathbf{D}(f, c_T) = 1 \end{gathered} \tag{9}$$

The minimization of (9) can be done analytically if all elements of (2) are known in analytical form, or numerically if some of them are measured and known only as set of values.

**Figure 1.** Microphone over the earbud.

**Figure 2.** HATS with microphone array and headset.

If the source signal $S_{c_T}(f)$ is known and wideband, i.e. covers all frequencies, then we can find beamformer weights optimal in minimum mean square (MMSE) sense:

$$\mathbf{W}_{c_T}(k) = \arg\min_{\mathbf{w}_{c_T}(f)} \frac{1}{N}\sum_{n=1}^{N}\left|Y_k^{(n)} - S_k^{(n)}(c_T)\right|^2, \qquad (10)$$

or in log-MMSE sense:

$$\mathbf{W}_{c_T}(k) = \arg\min_{\mathbf{w}_{c_T}(f)} \frac{1}{N}\sum_{n=1}^{N}\left|\log\left(\left|Y_k^{(n)}\right|\right) - \log\left(\left|S_k^{(n)}(c_T)\right|\right)\right|^2. \qquad (11)$$

In previous two equations we switched to discrete frequency domain, where $k$ is the frequency bin index, $n$ is the frame number and $N$ is the total number of frames. In our case we do not have an analytical expression to minimize, we even do not know the frequency distortion $F_m(f,c)$ in (2).

From this perspective (10) or (11) can be solved numerically, using some of the methods for mathematical optimization (gradient descent, for example). As the coefficients in $\mathbf{W}$ are complex, the number of parameters for optimization doubles (real and imaginary parts) and is equal to $2M$ for each frequency bin. Strictly speaking we do not have to impose any constraints during the optimization above, but with the multidimensional optimization we can have more freedom by adding additional requirements. This helps the optimization process in cases when the gradient in one or more dimensions is very small. In our case the combined optimization criterion $Q$ can be:

$$Q = \sum_{l=1}^{L} p_l Q_l \qquad (12)$$

where $Q_l$ are partial criteria with their weights $p_l$. As partial criteria we can use:

- $Q_1 = \frac{1}{N}\sum_{n=1}^{N}\left|Y_k^{(n)} - S_k^{(n)}(c_T)\right|^2$, or the log-MMSE case from (11), which is the main optimization criterion;
- $Q_2 = G_{AN}$, $Q_3 = G_{IN}$, which are theoretically estimated from (5) and (6);
- $Q_4 = |\mathbf{WD} - 1|$, which is the unit gain and zero phase, theoretically estimated.

## 4. EXPERIMENTAL RESULTS

Using the methodology above we designed a beamformer for two element microphone array. Microphones are unidirectional cardioid and point towards the mouth. The microphone array is unobtrusive and integrated with the headphone, i.e. the microphones are placed on the earbuds at the beginning of the ear canal – see figure 1.

There are several ways to have the source signal for beamformer design. One of them is a human speaker can wear an additional headset with a close-talking microphone. The problems here are that even the close-talking microphone picks up ambient noise, and that it is difficult to ensure that a human talker covers the whole frequency band with enough energy. For these reasons, we used a Head and Torso Simulator (HATS) for our design, as in Figure 2. Proper measures were taken to remove the dependency on the HATS and the manufacturing tolerances of the microphones. In an anechoic chamber we played a wideband chirp signal (linear frequency, 100 – 7000 Hz, 0.25 sec, 10 repetitions) through the HATS's mouth simulator and recorded the sound with a measurement grade microphone, placed in front of the mouth. Then a proper inverse filter $h_{HATS}(k)$ was designed in order to preprocess all sound files before playing them through the mouth of the simulator. In the same chamber a high quality loudspeaker, positioned in front of the HATS played the same chirp signal and it was recorded by the measurement microphone and the two cardiod microphones of the microphone array. Correction filters $h_L(k)$ and $h_R(k)$, for the left and right channels correspondingly, were designed to compensate for manufacturing tolerances in sensitivity and frequency response. These filters were used to preprocess all further records.

As a training data set we used the same sequence of 10 chirp signals. The development set consisted of male and

**Table 1.** Results for SNR and MOS.

| Channel | MMSE | LSD | SNR | Impr. | MOS | Impr. |
|---|---|---|---|---|---|---|
| Left microphone | 0.05340 | 1.13440 | 13.75 | | 2.339 | |
| Delay and sum | 0.05662 | 1.06010 | 17.99 | 4.24 | 2.548 | 0.209 |
| MMSE optimized, 1.00,1.209,1.244,0.034 | 0.04070 | 1.04801 | 20.63 | 6.88 | 2.876 | 0.459 |
| log-MMSE optimized, 1.00,0.058,1.115,0.924 | 0.04640 | 1.00480 | 18.16 | 4.41 | 2.567 | 0.328 |

female voices saying ten short utterances each. The test set was another set of male and female voices saying a different set of ten utterances. All sequences were played trough the HATS mouth in normal reverberant conditions (office, $RT_{60}$=310 ms) with three types of simulated noise: office noise at 55 dB SPL; café noise at 65 dB SPL; train station noise at 75 dB SPL. The noises were played through four speakers to simulate an ambient noise environment. The HATS was wearing the microphone array in its ears. A multichannel recorder was used to record the signals from the two microphones, the signal from a close-talking microphone, and the undistorted source signal, sent to the artificial mouth. As a result we had synchronously recorded all signals necessary for the design. After correction of the two microphone signals with $h_L(k)$ and $h_R(k)$ all signals were converted to frequency domain using MCLT [8].

Evaluation of the design was done based on the output SNR and the perceptual audio quality. The SNR was measured as the proportion of the average energy of the speech and noise frames. To increase the precision the frame classification was done using the clean signal. The perceptual sound quality was measured using PESQ algorithm [9].

The final goal of the beamformer design is not to suppress noise, or to minimize the mean square error, or the log-MMSE. It is actually to make the output sound better, i.e. to maximize MOS. This is why we did the optimization in two stages. The first is the optimization problem, described in the third section, using the training set and either MMSE or log-MMSE as the main criterion. Then the solution was used to process and evaluate the development set. The average MOS result was used for tuning the parameters weights $p_1 \div p_4$. This procedure was completely automated using another gradient descent procedure, which ensured that both MMSE and log-MMSE algorithms are at their best. The final evaluation of each solution was done using the test set. The results are shown in table 1. All SNRs are C-weighted and are in dBC. The MOS results are provided by PESQ algorithm. Optimal weights for MMSE and log-MMSE and computed MMSE and LCD are shown as well.

## 5. DISCUSSION

The proposed data driven beamformer design avoids the direct measurement of the frequency responses in the points of the microphones and their directivity patterns. The correction filters will be incorporated into the beamformer anyway. This design allows indirect optimization of what actually matters – the perceptual sound quality on the beamformer output. The log-MMSE design shows minimal improvements compared to delay and sum. To counter the trend of this algorithm to reduce the noise even by suppressing the signal the secondary optimization increased the weight of the instrumental noise gain and the unit gain requirement. The MMSE design provides best perceptual sound quality and noise suppression. As in this algorithm the noise has lower importance the secondary optimization increased the weight of the noise gains. The improvement in SNR is in the range of 4.4 to 6.8 dBC. While this improvement is not substantial, the improvement of 0.46 in perceptual sound quality is well audible.

The question remains how robust the designed beamformer is to the manufacturing tolerances of the microphones. In all cases an autocalibration procedure [10] should be used. There was no formal study how robust it is to the variations of the human head size and shape, but ad hoc experiments with human subjects confirmed the improvements in SNR and MOS.

## 6. REFERENCES


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

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

[3] S. Laugesen, K. Rasmussen, T. Christiansen, "Design of a Microphone Array for Headsets," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 2003, New Paltz, NY.

[4] T. Halkosaari, M. Vaalgamaa, "Directivity of Human and Artificial Speech". Joint Baltic-Nordic Acoustics Meeting 2004, 8-10 June 2004, Mariehamn, Åland.

[5] I. Tashev, M. Seltzer, A. Acero, "Microphone Array for Headset with Spatial Noise Suppressor". Proceedings of Ninth International Workshop on Acoustic, Echo and Noise Control IWAENC 2005, Eindhoven, The Netherlands, September 2005.

[6] I. Tashev, H. Malvar, "A New Beamformer Design Algorithm for Microphone Arrays," ICASSP 2005, Philadelphia, March 2005.

[7] T. Shoup, *A practical guide to computer methods for engineers*. Prentice Hall Inc., 1979.

[8] H. S. Malvar, "A modulated complex lapped transform and its applications to audio processing," ICASSP 99, Phoenix, pp. 1421–1424, March 1999.

[9] ITU-T Recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Geneva, Switzerland, 2001.

[10] I. Tashev, "Gain calibration procedure for microphone arrays," ICME 2004, Taipei, June 2004.