

A COMPACT MULTI-SENSOR HEADSET FOR HANDS-FREE COMMUNICATION

Zicheng Liu, Michael L. Seltzer, Alex Acero, Ivan Tashev, Zhengyou Zhang, and Mike Sinclair

Microsoft Research
Redmond, WA 98052 USA

{zliu, mseltzer, alexac, ivantash, zhang, sinclair}@microsoft.com

ABSTRACT

The need for hands-free communication has led to an increased popularity in the use of headsets with mobile phones. Comfort and portability concerns have led to the desire for headsets with a small form factor. Unfortunately, this size constraint typically requires that the microphone be placed farther from the user's mouth, making it highly susceptible to environmental noise. One long term goal of our work is to develop a headset that can achieve the sound capture performance of a close-talking microphone located at the user's mouth, while maintaining the desired compact size. Toward this end, we have designed a headset consisting of three air microphones and a bone-conductive sensor. The speech enhancement is performed in two stages, a fixed beamformer followed by a single-channel adaptive post-filter. Unlike other techniques, the beamformer is calibrated in a purely data-driven manner. We present preliminary experimental results using real data collected in multiple environments. The proposed approach results in significant improvements in both speech recognition accuracy and SNR.

1. INTRODUCTION

As mobile phones continue to grow in popularity, they are increasingly being used in places requiring hands-free communications, for reasons of either safety or convenience. As such, the growth of the cell phone industry has been followed by an almost equally large growth in the headset industry. For reasons of portability, convenience, comfort, and style, anecdotal evidence suggests that users strongly prefer headset designs which are compact and lightweight. These design constraints require that the microphone be placed at some distance from the user's mouth, typically on a short boom, the earpiece itself, or along the wire that connects the earpiece to the phone. As a result of this sub-optimal microphone placement, speech captured by such headsets is highly susceptible to distortion from environmental noise. In contrast, a traditional close-talking headset with a long-boom can place the microphone right at the user's mouth, resulting in a much higher signal-to-noise (SNR) ratio. The goal of this work is to design a headset that can achieve the sound capture performance of such a close-talking microphone, while maintaining a compact form factor.

There have been a small number of papers published that consider the same or a similar problem. In [1], a close-talking adaptive microphone array is proposed which aims to correct the excessive high frequency gain that results from placement more than a 1-2 cm from the user's mouth. The proposed algorithm required an endfire configuration in order to estimate of the distance from the source to the array based on level differences. In [2], the endfire constraint was relaxed by using both time-difference-of-arrival (TDOA) estimates and level differences to estimate the source loca-

tion. However, the farfield attenuation was limited if the array was not in an endfire orientation. Most recently, a method to overcome this drawback was proposed [3].

A fixed beamformer design for headset microphone array which relied on a physical model of the user was proposed in [4]. The head was modeled as a rigid sphere and a filter-and-sum beamformer was designed to maximize a variation on the conventional directivity index. Recently, there has been interest in the use of non-traditional sensors for speech processing. In [5], a traditional desktop headset was augmented with a bone conductive microphone, while in [6], the use of a variety of sensors including air microphones and throat and head accelerometers was proposed used to improve sound capture in noisy military environments.

In this paper, we explore the use of multiple sensors to create a compact, lightweight, mobile communications headset. The proposed headset consists of three air microphones and one bone conductive microphone. A microphone array is formed using two noise-canceling microphones and one omnidirectional microphone, while the omnidirectional microphone is also used in conjunction with the bone-conduction microphone to provide adaptive noise suppression.

We propose a fixed beamforming algorithm in which the array parameters are optimized during a calibration phase using a purely data-driven approach. Such an approach was successfully applied to an microphone array speech recognition task in [7]. Optimizing the array parameters in a data driven manner has several potential benefits over a more conventional array processing approach. For example, no assumptions are made about the microphone element characteristics. Manufacturing variations in the microphone characteristics will naturally be compensated for, as the gain and phase of each microphone are adjusted to optimize the desired objective function according to the data. Furthermore, a data-driven optimization criterion avoids the need for physical models such as those used in [4].

Once the beamformer is calibrated it is fixed for future processing. To gain additional noise suppression, the beamformer output is then fed into an adaptive post-filter that utilizes the accurate speech activity detection provided by the bone-conduction microphone. Unlike previous techniques, e.g. [8, 9], the proposed post-filter operates by directly modeling the noise transfer function between the omnidirectional microphone and the beamformer output.

Through a series of speech enhancement and speech recognition experiments on real data, we show that the combination of these processing stages results in an output signal that is significantly better than a short-boom headset microphone.

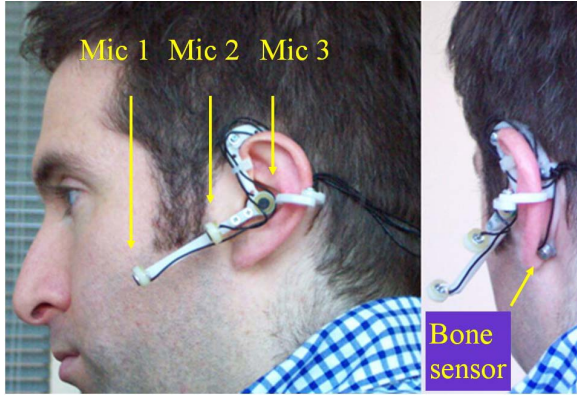


Figure 1: The proposed multi-sensor headset. Microphones 1 and 2 are directional microphones, while microphone 3 is omnidirectional. The bone sensor rests on the skull just behind the earlobe.

2. MULTI-SENSOR HEADSET DESIGN

Figure 1 shows our multi-sensor headset. It has three air microphones and a bone-conductive sensor. The three air microphones are placed along a short boom, forming a linear array. The spacing between the first and second microphones is 40 mm, and the spacing between the second and the third microphone is 25 mm. The first two air microphones are noise-canceling, while the third is omnidirectional. It is used both as part of the microphone array and for capturing the ambient noise for downstream adaptive filtering. The headset is worn with the three air microphones oriented toward the user's mouth. The omnidirectional microphone is located at the ear canal and the bone sensor rests on the skull behind the ear. The bone-conductive sensor is highly insensitive to ambient noise, and as such, provides robust speech activity detection [5].

3. ALGORITHM OVERVIEW

Figure 2 is an overview of our algorithm. The goal of this work is to create a compact headset with sound capture performance that approaches that of a high-quality close-talking microphone using the combination of a microphone array and a bone-conductive sensor. One method of doing so is to use the close-talking microphone signal itself as a reference signal to calibrate the parameters of the array. The details of the proposed method are discussed in Section 4. The linear beamformer will inevitably not eliminate all the environmental noise. To obtain additional noise reduction, the output of the calibrated linear beamformer is processed by an adaptive, non-linear noise suppressor, described in detail in Section 5.

4. HEADSET ARRAY CALIBRATION

In this section we describe in detail how the parameters of the headset microphone array are calibrated using the close-talking microphone reference signal.

Using a small sample of training recordings in which the user's speech is captured by both the microphone array and the close-talking microphone, the proposed calibration algorithm operates as follows.

We use a conventional subband filter-and-sum linear beamforming architecture. Thus, if we represent the k th subband of

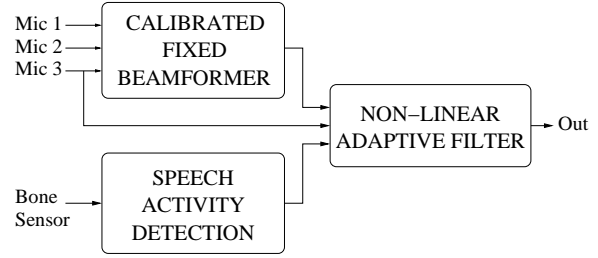


Figure 2: A block diagram of the processing stages for the multi-sensor headset.

short-time Fourier transform of the signal captured by microphone m at frame t as $Y_{m,t}[k]$, the beamformer output can be expressed as

$$Z_t[k] = \sum_{m=1}^M H_m[k] Y_{m,t}[k] \quad (1)$$

where $H_m[k]$ is the filter coefficient applied to subband k of microphone m and M is the total number of microphones in the array. If we define the reference signal from the close-talking microphone as $R_t[k]$, the goal of the proposed calibration algorithm is to find the array parameters that minimize the following objective function

$$\begin{aligned} \epsilon_k &= \sum_t |R_t[k] - Z_t[k]|^2 \\ &= \sum_t \left| R_t[k] - \sum_{m=1}^M H_m[k] Y_{m,t}[k] \right|^2 \end{aligned} \quad (2)$$

Taking the derivative of Eq. (2) with respect to $H_m^*[k]$ and setting the result to zero gives

$$\sum_t \left(R_t[k] - \sum_{n=1}^M H_n[k] Y_{n,t}[k] \right) Y_{m,t}^*[k] = 0 \quad (3)$$

By rearranging terms, we obtain

$$\sum_{n=1}^M \left(\sum_t Y_{m,t}^*[k] Y_{n,t}[k] \right) H_n[k] = \sum_t Y_{m,t}^*[k] R_t[k] \quad (4)$$

The filter coefficients $\{H_1[k], \dots, H_M[k]\}$ can then be found by solving the linear system in Eq. (4). This optimization is performed over all subbands $k = \{1 \dots N/2\}$, where N is the DFT size.

4.1. Relationship to LCMV Beamforming

For comparison purposes, we will show how the proposed beamformer calibration method is related to a more conventional array processing algorithms, Linearly Constrained Minimum Variance (LCMV) beamforming [10]. Because we process all subbands independently, we will drop the subband index k from our notation for clarity.

Let X denote the clean speech signal at the sound source. We can then model the microphone signals as follows:

$$Y_{m,t} = G_m X_t + N_{m,t} \quad m = \{1, \dots, M\} \quad (5)$$

$$R_t = G_R X_t + N_{R,t} \quad (6)$$

where G_m and G_R represent the source-to-sensor transfer functions for the microphones in the array and the reference microphone, respectively, and $N_{m,t}$ and $N_{R,t}$ represent the ambient noise at the corresponding sensors. Substituting Eq. (5) and Eq. (6) into Eq. (2) results in

$$\epsilon = \sum_t \left| \left(G_R - \sum_{m=1}^M H_m G_m \right) X_t + \left(N_{R,t} - \sum_{m=1}^M H_m N_{m,t} \right) \right|^2 \quad (7)$$

If we assume that the reference microphone represents ground truth, i.e. $G_R = 1$ and $N_{R,t} = 0$, then Eq. (7) simplifies to

$$\epsilon = \sum_t \left| \left(1 - \sum_{m=1}^M H_m G_m \right) X_t - \sum_{m=1}^M H_m N_{m,t} \right|^2 \quad (8)$$

If we had no information about X , we could simply constrain the beamformer to have unity gain in the look direction, i.e. set

$$\sum_{m=1}^M H_m G_m = 1 \quad (9)$$

In this case, the first term in Eq. (8) drops out, and the resulting optimization problem is to minimize

$$\epsilon = \sum_t \left| \sum_{m=1}^M H_m N_{m,t} \right|^2 \quad (10)$$

subject to the linear constraint defined by Eq. (9). This is the LCMV beamforming formulation.

The parameters of an LCMV beamformer are usually chosen by making some assumptions about the noise, e.g. the noise is spherically or cylindrically isotropic. In contrast, we make no assumptions about the noise and are utilizing *a priori* information about the target speech signal.

5. NON-LINEAR ADAPTIVE FILTERING

Following the calibration-based beamforming stage, we have a single-channel output signal Z . Invariably, the calibrated beamformer will not be able to remove all the ambient noise from the signal. To reflect this, we model the beamformer output Z as

$$Z_t = G_Z X_t + H_{Z,t} V_t \quad (11)$$

where G_Z is the spectral tilt induced by the array, V_t is the ambient noise, H_Z is the effective filter formed by the beamforming process.

To further enhance the output signal, we apply an adaptive filter to the output of the microphone array. This filter relies on noise information from the omnidirectional microphone and exploits the precise speech activity detection provided by the bone-conductive sensor.

If we define Y_o to be the omnidirectional microphone signal (microphone 3 in Figure 1), this signal can be modeled as

$$Y_{o,t} = G_o X_t + H_{o,t} V_t \quad (12)$$

We now define the following variables:

$$\tilde{X}_t = G_o X_t \quad (13)$$

$$\tilde{V}_t = H_{o,t} V_t \quad (14)$$

$$\tilde{G}_Z = G_Z / G_o \quad (15)$$

$$\tilde{H}_{Z,t} = H_{Z,t} / H_{o,t} \quad (16)$$

Substituting Eq. (13) – (16) into Eq. (11) and (12) gives

$$Z_t = \tilde{G}_Z \tilde{X}_t + \tilde{H}_{Z,t} \tilde{V}_t \quad (17)$$

$$Y_{o,t} = \tilde{X}_t + \tilde{V}_t \quad (18)$$

In essence, \tilde{G}_Z is the signal transfer function between the beamformer output and the omnidirectional microphone and $\tilde{H}_{Z,t}$ is the corresponding noise transfer function.

Notice that $\tilde{H}_{Z,t}$ in Eq. (17) is a function of time. However, if we assume that this variation over time is strictly a function of its phase, while its magnitude is relatively constant, we can rewrite $\tilde{H}_{Z,t}$ as

$$\tilde{H}_{Z,t} = |\tilde{H}_{Z,t}| e^{j\phi_t} \quad (19)$$

If we assume that the speech X and the noise V are uncorrelated, we can combine Eq. (17) – (19) to obtain

$$|Z_t|^2 = |\tilde{G}_Z|^2 |\tilde{X}_t|^2 + |\tilde{H}_{Z,t}|^2 |\tilde{V}_t|^2 \quad (20)$$

$$|Y_{o,t}|^2 = |\tilde{X}_t|^2 + |\tilde{V}_t|^2 \quad (21)$$

Solving for $|\tilde{X}_t|^2$ using these two expressions leads to

$$|\tilde{X}_t|^2 = \frac{|Z_t|^2 - |\tilde{H}_{Z,t}|^2 |Y_{o,t}|^2}{|\tilde{G}_Z|^2 - |\tilde{H}_{Z,t}|^2} \quad (22)$$

Because the denominator of Eq. (22) is constant over time, it acts simply as a gain factor. Therefore, we estimate $|\tilde{X}_t|^2$ simply as

$$|\hat{X}_t|^2 = |Z_t|^2 - |\tilde{H}_{Z,t}|^2 |Y_{o,t}|^2 \quad (23)$$

This leads to the following estimate of the magnitude of \tilde{X}_t

$$|\hat{X}_t| = |Z_t| \sqrt{\max \left(1 - \frac{|\tilde{H}_{Z,t}|^2 |Y_{o,t}|^2}{|Z_t|^2}, \epsilon \right)} \quad (24)$$

where ϵ is a small constant. As in other magnitude-domain noise suppression algorithms, e.g. spectral subtraction, we use the phase of the array output signal Z for the filter output as well. Thus, the final estimate of \tilde{X} is

$$\hat{X}_t = |\hat{X}_t| e^{j\angle Z_t} \quad (25)$$

where $\angle Z_t$ represents the phase of Z_t .

We estimate $|\tilde{H}_{Z,t}|$ using non-speech frames. In these frames, Eq. (20) and Eq. (21) simplify to

$$|Z_t|^2 = |\tilde{H}_{Z,t}|^2 |\tilde{V}_t|^2 \quad (26)$$

$$|Y_{o,t}|^2 = |\tilde{V}_t|^2 \quad (27)$$

Using these expressions, the least-squares solution for $|\tilde{H}_{Z,t}|$ is

$$|\hat{H}_{Z,t}| = \frac{\sum_t |Z_t| |Y_{o,t}|}{\sum_t |Y_{o,t}|^2} \quad (28)$$

	close-talk	short boom	calib array	+ adapt filter
Cafeteria	25.2	6.8	12.0	14.3
Car	26.8	11.6	12.6	16.4
Average	26.0	9.2	12.3	15.3

Table 1: The SNR (dB) obtained using a close-talking microphone, a short-boom microphone, the calibrated beamforming technique, and the beamformer combined with a nonlinear adaptive filter.

6. EXPERIMENTAL RESULTS

To test the performance of our proposed multi-sensor headset processing, we recorded speech data from a user wearing the headset shown in Figure 1 in two different environments, a cafeteria and a car. In addition to the multi-sensor headset, the user also wore a high-quality close-talking microphone with a long boom. The audio streams from all five sensors were captured simultaneously. In order to calibrate the headset beamformer, three utterances from each of the environments were used as training data. The utterances ranged were approximately 6-10 seconds long. The signals recorded by the three air microphones and the close-talking microphone were used to optimize the array coefficients using the technique described in Section 4. Once these parameters were computed, they were fixed for future processing.

The test data consisted of the user reading 42 utterances from the Wall Street Journal (WSJ0) corpus [11] in both the car and cafeteria environments. For each test utterance, the speech captured by the headset was processed by the calibrated fixed beamformer. The signal generated at the output of the beamformer was then processed by the adaptive filter described in Section 5. For each utterance, the bone sensor on the headset was used as a speech activity detector to locate non-speech frames in order to estimate $|\hat{H}_Z|$.

To evaluate the proposed algorithms, we measured both the SNR of the resulting output signal and the speech recognition accuracy. The SNR was measured by segmenting each utterance into speech and non-speech segments, and then computing the average signal energy in the speech frames and the noise frames. The results are shown in Table 1 for the close-talking microphone, a short-boom microphone (Mic 1 in Figure 1), the output of the calibrated beamformer, and the output after the adaptive filtering is applied. The performance is shown for the car and cafeteria environments, as well as the average.

The speech recognition accuracy was measured using a commercially available speech recognition engine, with the dictionary and language model constrained to the 5000-word WSJ0 task. Table 2 shows the Word Error Rate (WER) obtained by the system in the different environments.

As the tables indicate, the calibrated beamformer obtains a 27% relative WER improvement over performance of the single microphone even though the improvement in SNR is only 3 dB. The nonlinear adaptive filtering phase achieves an additional 3 dB SNR improvement. The fact that it gains a relative 3% improvement in WER is a good indication that although the filter is non-linear, it does not introduce significant distortion in the signal. By combining the headset beamformer with the nonlinear adaptive filter, we are able to achieve a 30% relative WER reduction, and 6dB increase in SNR compared to a single short-boom microphone.

	close-talk	short boom	calib array	+ adapt filter
Cafeteria	11.0	31.1	22.7	21.9
Car	12.4	26.4	19.2	18.6
Average	11.7	28.7	20.9	20.2

Table 2: WER (%) obtained using a close-talking microphone, a short-boom microphone, the calibrated beamforming technique, and the beamformer combined with a nonlinear adaptive filter.

7. CONCLUSIONS AND FUTURE WORK

We presented a multi-sensor headset designed to improve the sound capture performance over a conventional single-channel compact headset. The speech enhancement was performed by the headset using a fixed beamformer calibrated in a purely data-driven manner, and a non-linear adaptive post-filter which obtained robust speech activity detection from the bone sensor. Preliminary results showed significant improvement in both speech recognition accuracy and SNR.

The need for the close-talking reference signal in calibration is a drawback to the beamforming algorithm presenting in this paper. However, because of the variation in head sizes and shapes of different users, it is unclear whether a set of array coefficients calibrated for one user will generalize to other users. We would like to explore the construction of a codebook of beamformers in which each codeword corresponds to a physical user profile. Calibration would then consist of searching for the codeword or mixture of codewords that is the best match for the user. Finally, we would also like to evaluate other multi-channel signal processing methods. We note that the proposed nonlinear adaptive filtering technique will be still applicable with any other microphone array processing methods.

8. REFERENCES

- [1] G. W. Elko, J. E. West, and R. A. Kubli, "An adaptive close-talking microphone," in *Proc. 32nd Asilomar Conf. on Sig., Sys., and Comp.*, Pacific Grove, CA, Nov. 1998.
- [2] H. Teutsch and G. W. Elko, "An adaptive close-talking microphone," in *Proc. WASPAA*, New Paltz, NY, Oct. 2001.
- [3] J. Meyer and G. W. Elko, "Adaptive close-talking microphone based on sound-field eigenmode decomposition," in *Proc. HSCMA*, Piscataway, NJ, Mar. 2005.
- [4] S. Laugesen, K. B. Rasmussen, and T. Christiansen, "Design of a microphone array headset," in *Proc. WASPAA*, New Paltz, NY, Oct. 2003.
- [5] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. D. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, St. Thomas, USVI, Nov. 2003.
- [6] C. Farncourt, "Blind source separation with multi-modal speech sensors," in *Proc. HSCMA*, Piscataway, NJ, Mar. 2005.
- [7] M. L. Seltzer and B. Raj, "Speech-recognizer-based filter optimization for microphone array processing," *IEEE Signal Processing Lett.*, vol. 10, no. 3, pp. 69–71, Mar. 2003.
- [8] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. ICASSP*, Orlando, FL, May 2002.

- [9] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *Proc. ICASSP*, Munich, 1997.
- [10] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. NJ: Prentice Hall, 1993.
- [11] D. B. Paul and J. M. Baker, "The design of the Wall Street Journal-based CSR corpus," in *Proc. ARPA Speech and Nat. Lang. Workshop*, Harriman, NY, Feb. 1992, pp. 357–362.