MICROPHONE ARRAY FOR HEADSET WITH SPATIAL NOISE SUPPRESSOR

Ivan Tashev, Michael L. Seltzer, and Alex Acero

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

ABSTRACT

This paper presents a hybrid microphone array architecture used for processing the signals from a small microphone array that is used in a headset. The processing chain consists of fixed end-fire beamforming, adaptive spatial noise reduction and stationary noise suppression. The far-field design algorithm used for the fixed beamformer is adapted to the specifics of the headset by compensating for the directivity of the mouth and the sound diffraction around the head. The spatial noise reduction generalizes the suppression rule for optimal MMSE power noise reduction to multiple dimensions. The algorithm was tested with a headset that used a three element microphone array. It is computationally effective and provides up to 18 dB of ambient noise suppression.

1. INTRODUCTION

Mobile phones are increasingly being used in situations that require hands-free communication. As a result, mobile phone users are now using headsets with their telephones. Users have the option of using either a wire or Bluetooth wireless technology to connect their headset to their phone. 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. When you consider the fact that most users operate their phones in noisy environments, the SNR of the captured speech signal is even more significant.

One way to improve sound capture performance is to capture the speech signal using 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]. In practice, array processing algorithms cannot remove all noise from the sound field 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 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 dictate that a limited number of microphone elements, typically two or three, be used.

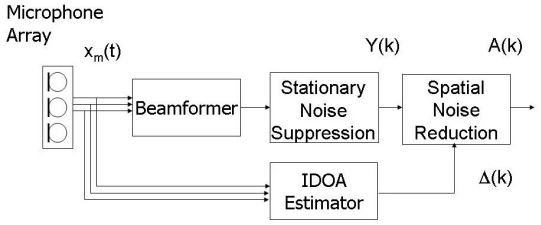


Figure 1. *Processing chain for the proposed microphone array headset.*

With a small number of microphones, the performance of any beamforming algorithm will be limited. To improve the noise suppression performance in these configurations, spatial noise reduction techniques have been proposed [4][5]. These techniques typically employ estimates of the instantaneous direction-of-arrival (DOA) in each frequency bin in order to suppress noise components that arrive from directions that differ from the main response axis (MRA).

In this paper, we present a hybrid microphone array processing algorithm that consists of a three stage processing chain: fixed linear beamforming, spatial noise reduction designed to remove directional noise sources, and single-channel adaptive noise suppression designed to remove any residual ambient or instrumental stationary noise. Figure 1 illustrates the processing chain for the microphone array headset. This beamformer design algorithm adapts our previously proposed algorithm, which was designed for far-field processing [6], to the headset scenario. This adaptation explicitly accounts for the sound diffraction around the user's head. Unlike [3], this is achieved without utilizing a physical model. Finally, a multi-channel noise reduction technique is presented that exploits the spatio-temporal distribution of the speech and the noise to further enhance the speech signal.

Because the headset uses a fixed beamforming approach, the desired beam is designed off-line, and then an efficient algorithm is used to process signals in real time. Through a series of experiments on speech recorded under varying noise conditions, we validated the performance of the proposed algorithms.

2. TIME-INVARIANT BEAMFORMER

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 $m = \{1,...,M\}$ is the microphone index. Each of the m sensors has a known directivity pattern $U_m(f,c)$, where f is the frequency band index and c

represents the location of the sound source in either a radial or a 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 $S(f)$ at a location *c*, the signal captured by each microphone can be represented as:

$$X_m(f, p_m) = D_m(f,c) A_m(f) U_m(f,c) S(f) \quad (1)$$

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

$$D_m(f,c) = F_m(f,c) \frac{e^{-j2\pi f v \|c - p_m\|}}{\|c - p_m\|} \quad (2)$$

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 a headset scenario, the signal decay due to energy losses in the air can be ignored. The term $A_m(f)$ in Eq. (1) is the frequency response of the system preamplifier and analog-to-digital conversion (ADC). In most cases we can use the approximation $A_m(f) \equiv 1$.

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:

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

where $W_m(f)$ are the weights for each sensor *m* and subband *f*, and $Y(f)$ is the beamformer output. (Note: Throughout this paper the frame index is omitted for simplicity.) The set of all coefficients $W_m(f)$ is stored as an $N \times M$ complex matrix **W**, where *N* is the number of frequency bins in a discrete-time filter bank, and *M* is the number of microphones. The matrix **W** is computed using the methodology described in [6]. In order to do so, the filter $F_m(f,c)$ in Eq. (2) must be determined. Its value can be estimated theoretically using a physical model, as in [1], or measured directly by using a close-talking microphone as reference. In this study we used direct measurement.

In any beamformer design, there is a tradeoff between ambient noise reduction and the instrumental noise gain. For the purposes of this study, we chose to obtain a more significant ambient noise reduction at the expense of increased instrumental noise gain. However, this additional noise is stationary and it can easily be removed using the stationary noise suppressor described in Section 3.

3. STATIONARY NOISE SUPPRESSOR

The stationary noise suppressor is an implementation of a gain-based noise suppression algorithm with MMSE power estimation and a suppression rule similar to that described in [8]. Besides removing the stationary part of the ambient noise remaining after the time-invariant beamformer, this noise suppressor reduces the instrumental noise from the microphones and preamplifiers.

4. SPATIAL NOISE REDUCTION

The output of the stationary noise suppressor is then processed by a novel spatial noise suppression algorithm. The proposed algorithm is a multi-dimensional generalization of the simplified Ephraim and Malah suppression rule [7] presented in [8].

For each frequency bin f the beamformer output $Y(f) \triangleq R(f).\exp(j\theta(f))$ consists of signal $S(f) \triangleq A(f).\exp(j\alpha(f))$ and noise $D(f)$. If we assume that they are uncorrelated, then $Y(f) \triangleq S(f) + D(f)$.

Given an array of microphones, we can find the instantaneous direction-of-arrival (IDOA) for a particular frequency bin based on the phase differences of non-repetitive pairs of input signals.

For *M* microphones these phase differences form an $M-1$ dimensional space, spanning all potential IDOA. If we define an IDOA vector in this space as

$$\Delta(f) \triangleq [\delta_1(f), \delta_2(f), \ldots, \delta_{M-1}(f)] \quad (4)$$

where

$$\delta_{j-1}(f) = \arg(X_1(f)) - \arg(X_j(f)) \quad j = \{2, \ldots, M\} \quad (5)$$

then we can define the signal and noise variances in this space as

$$\begin{aligned} \lambda_Y(f \mid \Delta) &\triangleq E\left[|Y(f \mid \Delta)|^2 \right] \\ \lambda_D(f \mid \Delta) &\triangleq E\left[|D(f \mid \Delta)|^2 \right] \end{aligned} \quad (6)$$

From [8], if we define

$$\frac{1}{\lambda(f \mid \Delta)} = \frac{1}{\lambda_Y(f \mid \Delta)} - \frac{1}{\lambda_D(f \mid \Delta)} \quad (7)$$

then we can define the *a priori* spatial SNR $\xi(f \mid \Delta)$ and the *a posteriori* spatial SNR $\gamma(f, \Delta)$ as follows:

$$\begin{aligned} \xi(f \mid \Delta) &\triangleq \frac{\lambda(f \mid \Delta)}{\lambda_D(f \mid \Delta)} \\ \gamma(f \mid \Delta) &\triangleq \frac{R^2(f \mid \Delta)}{\lambda_D(f \mid \Delta)} \end{aligned} \quad (8)$$

Based on these equations and the minimum-mean square error spectral power estimator, the suppression rule described in [8] can be generalized to

$$H(f\,|\,\Delta)=\sqrt{\frac{\xi(f\,|\,\Delta)}{1+\xi(f\,|\,\Delta)}\left(\frac{1+\vartheta(f\,|\,\Delta)}{\gamma(f\,|\,\Delta)}\right)} \tag{9}$$

where $\vartheta(f\,|\,\Delta)$ is defined as

$$\vartheta(f\,|\,\Delta)\triangleq\frac{\xi(f\,|\,\Delta)}{1+\xi(f\,|\,\Delta)}\gamma(f\,|\,\Delta)\,. \tag{10}$$

Thus, for each frequency bin of the beamformer output, the IDOA vector $\Delta(f)$ is estimated based on the phase differences of the microphone array input signals $\{X_1(f),...,X_M(f)\}$. The spatial noise suppressor output for this frequency bin is then computed as

$$A(f)=H(f\,|\,\Delta).|Y(f)|.\exp(j\theta(f)) \tag{11}$$

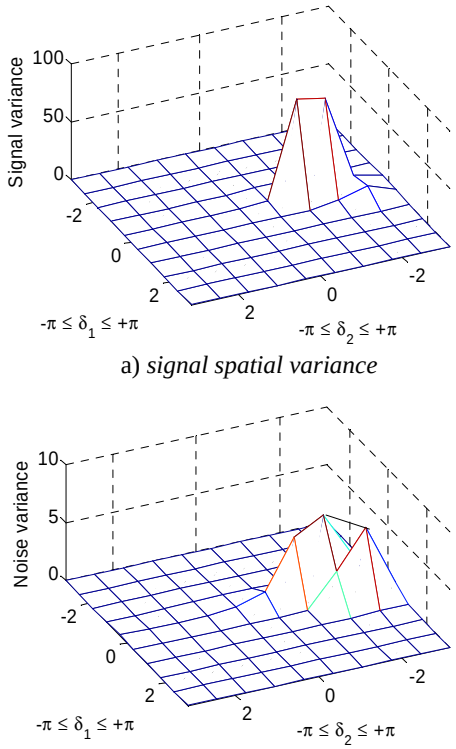
Note that this is a gain-based estimator and accordingly we directly apply the phase of the beamformer output signal.

In practical realizations of the proposed spatial noise reduction algorithm, the $(M-1)$-dimensional space of the phase differences is discretized. Empirically, we found that using 10 bins to cover the range $[-\pi,+\pi]$ provided adequate precision and results in a resolution of the differences in the phases of 36^{O}. This converts λ_Y and λ_D to square matrices for each frequency bin. The noise and input signal variance models are computed based on the decision of a speech activity detector and Eq. (6). In addition to updating the current cell in λ_Y and λ_D, the averaging operator $E[\ ]$ performs "aging" of the values in the other matrix cells.

To increase the adaptation speed of the spatial noise suppressor, the signal and noise variance matrices λ_Y and λ_D are computed for a limited number of equally spaced frequency subbands. The values for the remaining frequency bins can be computed using a linear interpolation or nearest neighbor technique. The variance matrices for the subband around 1000 Hz are shown in Figure 2. Note that the vertical axis is different in each plot. These variances were measured under 75 dB SPL ambient cocktail-party noise. The figure clearly shows that the signal from the speaker is concentrated in certain area – direction 0^{O}. The uncorrelated instrumental noise is spread evenly in the whole angular space, while the correlated ambient noise is concentrated around the DOA trace $0-\pi/2-\pi$. Due to the beamformer, the variance decreases as it goes farther from the focus point at 0^{O}.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed algorithms, we constructed a headset composed of a three element microphone array. The array consists of two directional cardioid microphones set at distances of 64 and 25 millimeters from the third microphone, which is omnidirectional. We also used a close-talking microphone on a long boom to create a recording that was used for reference and calibration. All four channels were recorded synchronously at a sampling rate of 16 kHz. The experimental headset is shown in Figure 3.



a) *signal spatial variance*

b) *noise spatial variance*

Figure 2. *Signal and noise spatial variances*

The signal decay around the head $F_m(f,c)$ was measured during calibration using the close-talk microphone as reference. The magnitudes of $F_m(f,c)$ for the first two microphones are shown on Figure 4. They are smoothed to remove speaker/speech dependencies and used in the beamformer design. The device uses a multiple subband variant of the autocalibration procedure described in [10] for compensation of the microphone manufacturing tolerances and errors in the estimation of $F_m(f,c)$.

For testing, utterances were recorded in three different environments: a quiet office, a café which had 75 dB SPL cocktail-party type ambient noise, and a car with 90 dB SPL. The following processing was used for each speech sample. The incoming waveforms from each microphone were segmented into 40 ms frames with a 20 ms frame shift. Each frame was weighted with a Hanning window, and converted to the frequency domain using an MCLT transform [9]. The microphone array signals were combined using the time-invariant beamformer described in Section 2. The output signal was then processed by the stationary noise suppressor and

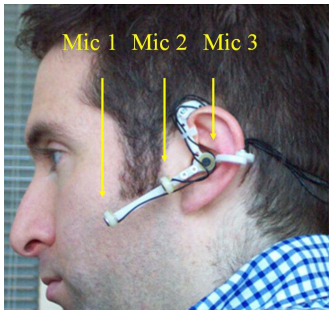


Figure 3. *Experimental headphone setup.*

the spatial noise reduction algorithm described in Sections 3 and 4, respectively.

Table 1 shows the resulting SNRs for the different processing stages in the different sampling environments. The SNR is measured as the difference in the average RMS in dB of the signal and noise frames. In the table, ***SB*** is the short-boom, the single microphone closest to the mouth (Mic 1 on Figure 3), ***BF*** is the beamformer output, ***NS*** is the output of the stationary noise suppressor and ***SR*** is the output of the spatial noise reduction.

Table 1. *SNR of the output signal after different processing stages.*

	SB	***BF***	***NS***	***SR***
Office, 55 dB	25.2	22.5	29.4	34.7
Café, 75 dB	7.2	12.3	17.5	22.8
Car, 90 dB	3.2	6.4	11.1	16.4

From a noise suppression perspective, the processing chain is well balanced, providing relatively equal SNR improvement for each block in the chain. Note that the beamformer actually causes a reduction in the SNR in low noise environments because of the instrumental noise gain. However, the SNR is improved significantly by the later processing stages. In the café noise case, the total noise suppression reaches 15.6 dB, while 12.6 dB of noise suppression is obtained in the car. This is probably due to the fact that the microphone array on the head of the passenger is oriented toward the engine, thereby picking up the engine noise.

6. RESULTS AND DISCUSSION

In this paper, we described a microphone array for headset applications. The signals captured by the array are processed in three stages: fixed linear beamforming, spatial noise reduction, and stationary noise suppression. The proposed algorithm generates output signals that have good sound quality and significant improvement in SNR. We note that while we used a fairly aggressive quantization of the phase space in the spatial noise reduction, increasing the resolution may improve the noise reduction and decrease any introduced

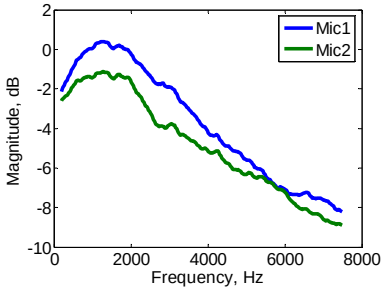


Figure 4. *Diffraction around the head*

distortions. One drawback of this approach is that it does not scale very well to larger arrays. Adding more microphones will increase the number of noise and signal spatial variance matrices, which reduces the adaptation speed and increases memory requirements.

7. 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] C. Lai, P. Aarabi, "Multiple-Microphone Time-Varying Filters For Robust Speech Recognition," ICASSP 2004, Montreal, May 2004.

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

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

[7] Y. Ephraim, D. Malah. "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-32, No. 6, December 1984.

[8] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," In Proceedings of the IEEE Workshop on Statistical Signal Processing, pages 496-499, 2001.

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

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