Direct Filtering
for Air- and Bone-Conductive Microphones

Zicheng Liu, Zhengyou Zhang, Alex Acero, Jasha Droppo, Xuedong Huang

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{zliu, zhang, alexac, jdroppo, xdh}@microsoft.com

***Abstract*—Air- and bone-conductive integrated microphones have been introduced by the authors [5,4] for speech enhancement in noisy environments. In this paper, we present a novel technique, called *direct filtering*, to combine the two channels from the air- and bone-conductive microphone for speech enhancement. Compared to the previous technique, the advantage of the direct filtering is that it does not require any training, and it is speaker independent. Experiments show that this technique effectively removes noises and significantly improves speech recognition accuracy even in highly non-stationary noisy environments.**

Keywords—speech enhancement; speech recognition; bone sensor

I. INTRODUCTION

How to handle non-stationary noises has been one of the most difficult problems in both automatic speech recognition and audio enhancement. In the previous work [5,4], we introduced air- and bone-conductive integrated microphones and showed that such devices can be used to reliably determine whether the speaker is talking or not, and furthermore, the two channels can be combined to remove overlapping noises. We use “WITTY”, which stands for “Who Is Talking To You”, as our acronym for the air- and bone-conductive microphones. A WITTY microphone contains two sensors: a regular close-talk microphone and a bone-conductive microphone. The close-talk microphone contains wideband speech but is noise sensitive. The bone-conductive microphone is noise resistant but is narrow band. The previous work [5,4] used a channel mapping technique for speech enhancement. It works by training a piecewise linear mapping from the bone signal to the close-talk signal. One drawback of this approach is that it requires training for each speaker. In this paper, we present a new technique which does not require any training. We call this technique direct filtering. The basic idea is to directly design a filter which performs distortion correction on the bone signal and optimally combines the bone signal and the close-talk signal to remove the noises.

II. RELATED WORK

Graciarena et. al. [1] combined the standard and throat microphones in the noisy environment. They trained a mapping from the concatenated features of both microphone signals in a noisy environment to the clean speech. Compared to their system, our algorithm does not need any training, and it is not environment dependent. In addition, our algorithm produces audible speech signals so that the output can be used for perception as well as speech recognition.

Strand et. al. [3] designed an ear plug to capture the vibrations in the ear canal, and used the signals for speech recognition with MLLR adaptation. Heracleous et. al. [2] used a stethoscope device to capture the bone vibrations of the head and use that for non-audible murmur recognition. Like Strand et. al. [3], they only used the bone signals for speech recognition with MLLR adaptation.

III. AIR- AND BONE-CONDUCTIVE INTEGRATED MICROPHONES

For a detailed description of the Air- and bone-conductive integrated microphones (WITTY microphones), the reader is referred to [5]. Figure 1 shows a prototype of this device. It contains two output channels: close-talk microphone and bone sensor. The bone sensor has the interesting property that it is insensitive to ambient noise but it only captures the low frequency portion of the speech signals. We would like to combine the bone signals with the close-talk signals to remove environment noise. The previously designed channel mapping technique [5,4] has the drawback that it requires training for each speaker, and it is speaker dependent. In the following, we describe a new algorithm, called direct filtering, which does not have such limitations.

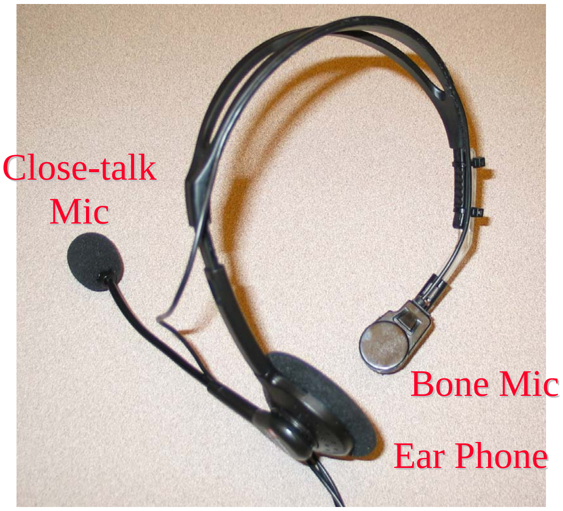

Figure 1. A WITTY microphone prototype

IV. DIRECT FILTERING

Let $y(t)$ and $b(t)$ denote the close-talk and bone signals, respectively. Let $x(t)$ denote the clean speech which is to be estimated. The following is the mathematical model of the direct filtering:

$$\begin{aligned} y(t) &= x(t) + v(t) \\ b(t) &= h(t) * x(t) + w(t) \end{aligned} \tag{1}$$

where v is the noise in the close talk channel which contains the environment noise such as background speech, w is the noise in the bone channel which contains the sensor noise and much attenuated environment noise, and h is the impulse response of the bone sensor. In the frequency domain, equation (1) becomes

$$\begin{aligned} Y_t(k) &= X_t(k) + V_t(k) \\ B_t(k) &= H(k)X_t(k) + W_t(k) \end{aligned} \tag{2}$$

where k is the frequency band. We assume $V_t(k)$ and $W_t(k)$ are zero-mean Gaussian random variables: $V_t(k) \sim N(0, \sigma_v^2), W_t(k) \sim N(0, \sigma_w^2)$.

To estimate H reliably, we use multiple frames of observation data. Let T be the number of frames used for estimating H. The maximum likelihood estimation is given by minimizing

$$\Re = \sum_{t=1}^{T} \left(\frac{1}{2\sigma_v^2}|Y_t - X_t|^2 + \frac{1}{2\sigma_w^2}|B_t - HX_t|^2\right) \tag{3}$$

Notice that X_t is complex variable, and $\Re$ is a real function of the real part and imaginary part of X_t. Therefore the partial derivatives of $\Re$ with respect to the real part and imaginary part of $\Re$ are zero at the optimum. It is easy to show that this leads to $\frac{\partial \Re}{\partial X_t} = 0$, where $\Re$ is regarded as a function of two variables: X_t and X_t^*, and the partial derivative is with respect to the first variable.

From (3), we have

$$\frac{\partial \Re}{\partial X_t} = \sum_{t=1}^{T} \left(\frac{-1}{2\sigma_v^2}(Y_t - X_t)^* + \frac{-H}{2\sigma_w^2}(B_t - HX_t)^*\right) \tag{4}$$

By setting $\frac{\partial \Re}{\partial X_t} = 0$, we have

$$X_t = \frac{\sigma_w^2 Y_t + \sigma_v^2 H^* B_t}{\sigma_w^2 + \sigma_v^2 |H|^2} \tag{5}$$

Substituting (5) into (3), we have

$$\Re = \sum_{t=1}^{T} \left(\frac{|B_t - HY_t|^2}{\sigma_w^2 + \sigma_v^2 |H|^2}\right) \tag{6}$$

By setting $\frac{\partial \Re}{\partial H} = 0$ (again, $\Re$ is regarded as a function of two variables: H and H^*), we obtain

$$a^* \sigma_v^2 (H^*)^2 - bH^* - a\sigma_w^2 = 0 \tag{7}$$

where

$$a = \sum B_t^* Y_t$$

$$b = \sum (\sigma_v^2 |B_t|^2 - \sigma_w^2 |Y_t|^2)$$

The roots of Equation (7) are:

$$H^* = \frac{b \pm \sqrt{b^2 + 4|a|^2 \ \sigma_v^2 \sigma_w^2}}{2a^* \sigma_v^2} \tag{8}$$

Therefore

$$H = \frac{b \pm \sqrt{b^2 + 4|a|^2 \ \sigma_v^2 \sigma_w^2}}{2a \sigma_v^2} \tag{9}$$

Notice that there are two possible solutions for H. In our implementation, we select the one resulting in the smaller value of $\Re$.

The clean speech is estimated also by minimizing (3) except that the summation is over a single frame. The solution is therefore the same as equation (5).

V. IMPLEMENTATION

Figure 2 shows the data flow of the direct filtering algorithm. Given a speech utterance ($y(t), b(t)$), we first divide $y(t)$ and $b(t)$ into frames and perform short-window Fourier transform for each frame. We estimate σ_v and σ_w from the initial frames of the utterance. H is estimated from the entire utterance by using equation 9. For each frame, the clean speech is estimated by using equation (5). Finally, we use inverse Fourier transform to convert back to the wave form.

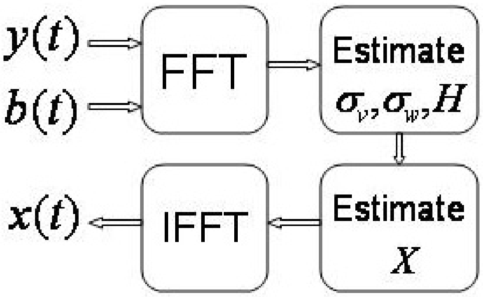

Figure 2. The direct filtering algorithm

VI. RESULTS

We have experimented with the direct filtering algorithm on both artificial data and real data. For artificial data, we recorded speech data for 4 people (2 male and 2 females) in a quiet office using the air- and bone-conductive integrated microphone (Figure 1). Each person read 42 utterances of the Walt Street Journal. We then added babble noises to the clean speech with various SNR levels. Notice that we also added certain amount of noises to the bone signal to simulate the leakage of the bone sensor (the bone sensor may still pick up a small amount of environment noises. We call it leakage). Table 1. shows the word error rates (WER) of both the close-talk signals and the enhanced signals under different SNR levels. For each SNR level, the word error rate is averaged over the 4 people.

	-5dB	0dB	5dB	10dB	15dB
Close-talk	85%	59%	34%	19%	12%
Enhanced	32%	19%	13%	10%	9%
Relative error reduction	62%	68%	62%	47%	25%

Table 1. Speech recognition results (WER) of the artificial data.

Figure 2-5 shows an example. Figure 2 is the clean speech from the close-talk microphone. Figure3 is the clean speech plus added noise (0dB). Figure 4 is bone signal with noise being added. Figure 5 is the enhanced signal. By comparing Figure 3 with Figure 5, we can see that most of the noises in the low frequency region (toward the bottom of the spectrogram) have been removed.

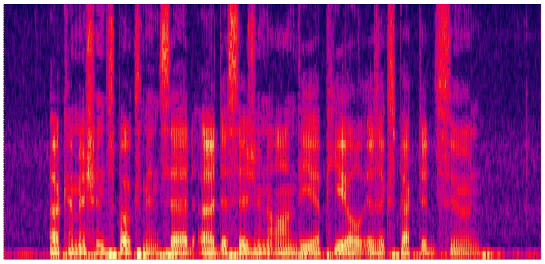

Figure 2. Clean speech

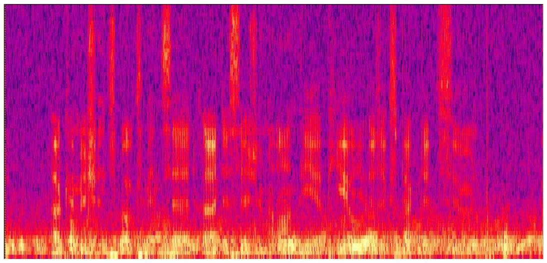

Figure 3. After adding noise

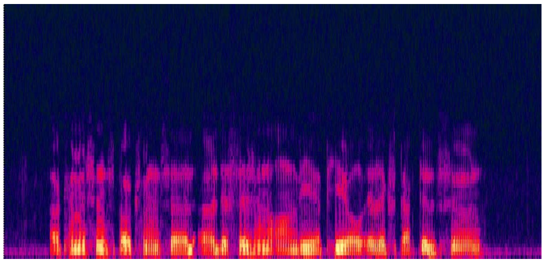

Figure 4. Bone signal with added noise

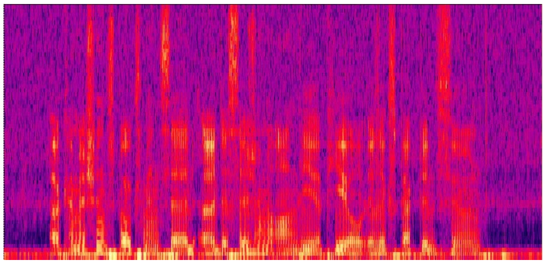

Figure 5. Enhanced result.

For real data testing, we recorded speech in two different environments for three people (one male and two females). The first environment is in an office with a different person speaking in the background. The second environment is in a cafeteria. We recorded 42 utterances for each person in each environment. Figure 6 shows an example: the top is the signal from the close-talk microphone while the bottom is the signal from the bone sensor. Notice that even though the bone sensor is much more noise resistant than the close-talk microphone, a small amount of environment noise leaking through our current bone sensor. Our current remedy is to use the simple spectral subtraction technique on the bone signal before applying direct filtering. Figure 7 shows the result from the direct filtering algorithm. We can see that the noise is reduced significantly for the lower frequency portion. The reason that it reduces noises mainly at the lower frequency is because the bone signal contains only low frequency information (up to 3.5 KHz).

Table 2 shows the speech recognition results for the read data. The word error rate is averaged over the three people. We can see that in both environments, we have achieved over 30% of relative error reduction.

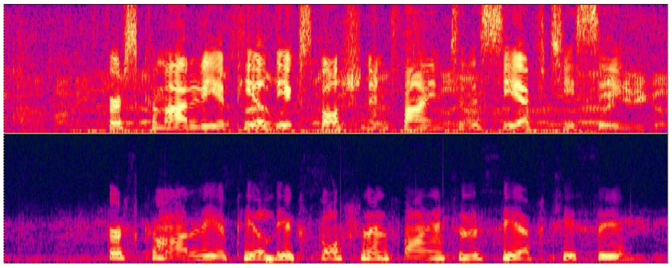

Figure 6. Spectrograms of the original data. Top: close-talk signal. Bottom: bone signal.

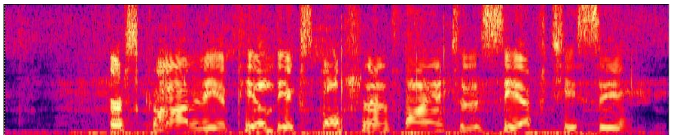

Figure 7. Result from Direct Filtering.

	Office with background speech	Cafeteria
Original	45%	35%
Enhancement	28%	24%
Relative Improvement	38%	31%

Table 1. Speech recognition results (WER) of real data.

VII. CONCLUSION AND FUTURE WORK

We have presented a new technique, called direct filtering, to combine the two channels of the air- and bone-conductive integrated microphone for speech enhancement. Our experiments show that under even highly non-stationary noisy environments, the direct filtering effectively reduces environment noises and results in cleaner speech for better perception as well as higher speech recognition accuracy.

We are currently implementing a real time system. Given that the speech detection is much easier thanks to the bone sensor [5], we will be able to estimate σ_v and σ_w dynamically, and H can then be estimated from the past frames.

VIII. REFERENCES

1. M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, Combining standard and throat microphones for robust speech recognition, IEEE Signal Processing Letters, vol. 10, no. 3, pp. 72-74, March, 2003.
2. P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano, Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation, ASRU 2003, St. Thomas, U.S. Virgin Islands, Nov 20-Dec4, 2003.
3. O. M. Strand, T. Holter, A. Egeberg, and S. Stensby, On the feasibility of ASR in extreme noise using the PARAT earplug communication terminal, ASRU 2003, St. Thomas, U.S. Virgin Islands, Nov 20-Dec4, 2003.
4. Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. D. Huang, Y. Zheng, Multi-sensory microphones for robust speech detection, enhancement, and recognition, ICASSP04, Montreal, May17-21, 2004.
5. Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero X. D. Huang, Air- and bone-conductive integrated microphones for robust speech detection and enhancement, ASRU 2003, St. Thomas, U. S. Virgin Islands, Nov. 30-Dec. 4, 2003.