AIR- AND BONE-CONDUCTIVE INTEGRATED MICROPHONES FOR ROBUST SPEECH DETECTION AND ENHANCEMENT

Yanli Zheng, Zicheng Liu, Zhengyou Zhang, Mike Sinclair, Jasha Droppo, Li Deng, Alex Acero, Xuedong Huang*†

Microsoft Research
Redmond, WA 98052

ABSTRACT

We present a novel hardware device that combines a regular microphone with a bone-conductive microphone. The device looks like a regular headset and it can be plugged into any machine with a USB port. The bone-conductive microphone has an interesting property: it is insensitive to ambient noise and captures the low frequency portion of the speech signals. Thanks to the signals from the bone-conductive microphone, we are able to detect very robustly whether the speaker is talking, eliminating more than 90% of background speech. Furthermore, by combining both channels, we are able to significantly remove background speech even when the background speaker speaks at the same time as the speaker wearing the headset.

1. INTRODUCTION

One of the most difficult problems for an automatic speech recognition system is in dealing with noises. When there are multiple people speaking, it is difficult to determine whether the captured audio signal is from the speaker or from other people. In addition, the recognition error is much larger when the speech is overlapped with other people's speech. Because the speech is non-stationary, it is extremely hard to remove the background speech from just one channel of audio signals.

In this paper, we propose a hardware device that combines regular microphone (air-conductive microphone) with a bone-conductive microphone with the purpose of handling noisy environment. The device is designed in such a way that people wear it just like a regular headset, and it can be plugged into any machine with a USB port. Compared to the regular microphone, the bone-conductive microphone is insensitive to ambient noise but it only captures the low frequency portion of the speech signals. Because it is insensitive to noise, we use it to determine whether the speaker is talking or not. And we are able to eliminate more than 90% of the background speech. Since the bone-conductive signals only contain low frequency information, it is not good to directly feed the bone-conductive signals to an existing speech recognition system. We instead use the bone-conductive signals for speech enhancement. By combining the two channels from the air- and bone- conductive microphone, we are able to significantly remove background speech even when the background speaker speaks at the same time as the speaker wearing the headset.

2. RELATED WORK

There has been a lot of work on using cameras to help with speech detection and recognition. Researchers have used both visual and audio information to determine whether the user is speaking or not [1, 2]. DeCuetors et al [3] used both video and audio signals for speaker intent detection. Chen et al [4] and Basu et al [5] used visual information to improve speech recognition in noisy environments.

Graciarena et al [6] combined the standard and throat microphones in the noisy environment. They used a probablistic optimum filter mapping algorithm to estimate the clean speech features from the speech features of both microphones. There are three main differences between their work and ours. One difference is that our hardware is different. Our hardware has the look and feel of regular headset while their hardware requires wearing two separate devices: one on the neck and a regular microphone on the face. The second difference is that we have developed an algorithm to detect speech and modulate the regular microphone signals based on the speech detection results. As a result, our headset can be used with any existing speech recognition products and it removes the noise between speeches. The third difference is in the speech enhancement algorithm. Their algorithm requires a database of simultaneous clean and noisy recordings. It achieves its best performance only when the noise condition of the test data matches the noise condition of the training data. It didn't report any results on simultaneous speech environment. In comparison, our algorithm only requires clean training data. We rely more on the bone sensor, which is insensitive to noise, to reconstruct the clean

*zheng3@ifp.uiuc.edu. Current address: University of Illinois at Urbana-Champaign

†{zliu,zhang,sinclair,jdroppo,deng,alexac,xdh}@microsoft.com

speech signals. Our algorithm is targeted at the simultaneous speech environment.

3. AIR- AND BONE-CONDUCTIVE INTEGRATED MICROPHONES

Fig. 1. The Air- and Bone-Conductive Integrated Microphone.

When we speak, there is vibration on the bones of the head. The bone-conductive sensors, when pressed again the bones, can capture the bone vibrations. The bone sensors are in general insensitive to noise. Figure 1 shows a prototype of our Air- and Bone-Conductive microphone. It is the same as a regular headset except that we added a bone-conductive sensor. Since the regular microphone input on the audio cards do not take stereo data, we use a USB HUB to combine the two channels into a stereo data. We can then plug this device into any machine which has a USB port.

Figure 2 shows an audio stream recorded by the integrated microphone. There are two people sitting about 3 feet apart. One person wears the microphone. The other person acts as a noise generator. The top row of Figure 2 is the signal from the regular microphone. The bottom is the signal from the bone sensor. We can see that it is much easier to differentiate the speech and noise from the bone sensor than from the regular microphone. Therefore, we can use the bone sensor for speech detection.

Figure 3 shows the spectral view of each channel ranging from 0 to 8KHz. Whereas the regular microphone contains wideband speech suitable for recognition, the bone sensor only contains narrowband speech. Therefore if we simply feed the bone sensor signals to an existing speech recognition system, the results would not be good.

If enough transcribed training data were available, one could build a speech recognition engine specifically targeted for bone signals. The problem is that there is no such data available.

We are taking a more practical approach: using the bone sensor to enhance the wideband noisy speech for use with an existing speech recognition system. Since the bone sensor signals contain very little noise, we can combine the bone sensor signals with the close talk microphone signals to obtain a better estimate of the clean speech.

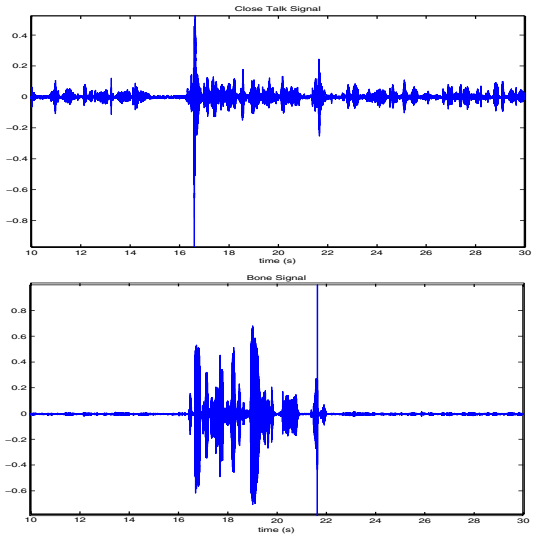

Fig. 2. An audio stream recorded by the bone microphone. Top: the regular microphone. Bottom: the bone sensor

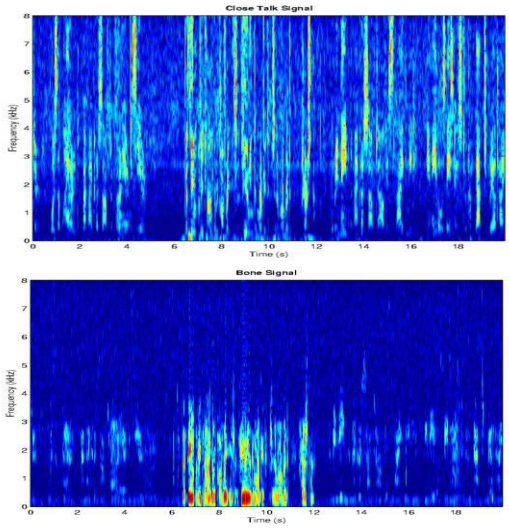

Fig. 3. The spectral view (0–8KHz). Top: the regular microphone. Bottom: the bone sensor.

4. SPEECH DETECTION

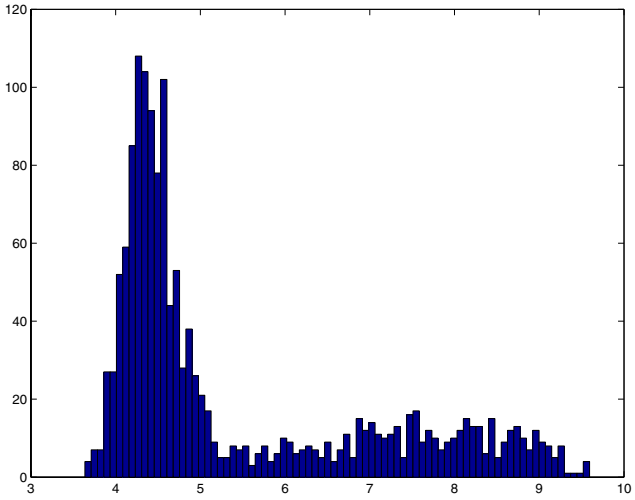

Fig. 4. Histogram of the bone sensor log energy

Figure 4 shows a histogram of the log energy of the bone sensor signals in Figure 2. We can see that the speech signals and the non-speech signals are separated very well. So we choose to use a moving-window histogram approach for speech detection. For each frame, we first compute the histogram of the audio data prior to this frame with a fixed window size (notice that we compute the histogram of the energy instead of the log energy). In our implementation, we use the window size of 1 minute. To reduce computation time, the histogram is updated incrementally for each frame. The second step is to find the valley after the first peak and use the valley as the separator.

Denote d to be the energy separator. Given any frame, let e denote its energy. Set $r = e/d$. The speech confidence for this frame is set to be

$$p = \begin{cases} 0 & : \quad r < 1 \\ \frac{r-1}{\alpha-1} & : \quad 1 \leq r \leq \alpha \\ 1 & : \quad r > \alpha \end{cases} \tag{1}$$

where α is a user specified parameter which defines the grace transition period between the two states. We choose $\alpha = 2$ in our implementation.

Finally we smooth out p by taking the average of the current frame with the previous 4 frames.

4.1. Removal of non-overlapping noises

We use the speech confidence p to modulate the audio signals from the regular microphone to remove the noises between speeches. The modulation factor f is set to be $f = h(p)$ where $h(x)$ is a Hermite polynomial with the property that $h(0) = 0$, $h(1) = 1$ and its derivatives at 0 and 1 are both zeros. For each sample x of the regular microphone signal in this frame, its modulated value is $\tilde{x} = f * x$.

	H	D	S	I	N
x	227	8	57	213	292
$\tilde{x}$	227	5	60	21	292

Table 1. The performance of the noise removal with the integrated microphone.

In this way, our integrated microphone can be directly used with any existing speech recognition system. To measure the performance of the noise removal algorithm, we used our new microphone with Microsoft's speech recognition system. The setup is as follows. We had two people each reading an article from a newspaper. One person wore the integrated microphone while the other person acted as a noise generator. The two people spoke alternatively. We recorded 5 minutes of data. Figure 2 shows a portion of the recorded data.

Table 1 shows the recognition results. The top row x is the result obtained by feeding the regular microphone signals directly to the speech recognition system. The bottom row $\tilde{x}$ is the result by feeding the modulated signal to the speech recognition system. N is total number of words and H is the number of correctly recognized words. D, S, and I are deletion, substitution, and insertion errors, respectively. We can see that the insertion error is reduced by 90% while it does not increase the deletion or substitution errors.

5. SPEECH ENHANCEMENT

In this section, we describe how to use the bone sensor for speech enhancement in an environment with highly non-stationary noises such as when there are people talking in the background.

Figure 5 shows the graphical model of the integrated microphone. Here we make the approximation that bone sensor is not affected by the noise at all. b and y are observations. y is corrupted by the noise. There is a channel distortion from from x to b. Basically b only contains frequency up to 4KHz. The speech enhancement problem becomes recovering x given b and y.

Our idea is to first predict x from b, and then combine both b and y to obtain the final estimate of x. To predict x from b, we use a technique which has some similarity to SPLICE [7], which is a frame-based noise removal algorithm for cepstral enhancement in the presence of additive noise. Instead of learning the mapping from the corrupted speech y to clean speech x as in the original SPLICE algorithm, we learn the mapping from bone sensor signals b to clean speech x.

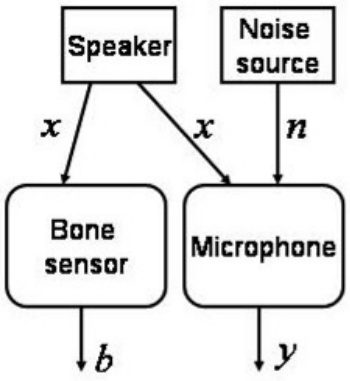

Fig. 5. The graphical model

5.1. SPLICE training

We use a piecewise linear representation to model the mapping from b to x in the cepstral domain.

$$p(x,b) = \sum_s p(x|b,s)p(b|s)p(s) \tag{2}$$

$$p(x|b,s) = N(x; b + r_s, \Gamma_s) \tag{3}$$

$$p(b|s) = N(b; \mu_b, \Gamma_b) \tag{4}$$

The bone sensor model $p(b)$ contains 128 diagonal Gaussian mixture components, and is trained using standard EM techniques, the parameters r_s of the conditional pdf $p(x|b,s)$ can be trained using the maximum likelihood criterion.

$$r_s = \frac{\sum_n p(s|b_n)(x_n - b_n)}{\sum_n p(s|b_n)} \tag{5}$$

$$\Gamma_s = \frac{\sum_n p(s|b_n)(x_n - b_n - r_s)(x_n - b_n - r_s)^T}{\sum_n p(s|b_n)} \tag{6}$$

$$p(s|b_n) = \frac{p(b_n|s)p(s)}{\sum_s p(b_n|s)p(s)} \tag{7}$$

5.2. Clean Speech Estimation

Assuming the noise is additive, in the cepstral domain we have $y = x + C\ln(1 + e^{C^{-1}(n-x)})$, where C is the DCT matrix and n is the noisy cepstral [8]. Since the noise estimation in the cepstral domain involves highly nonlinear process, we handle it in the power spectral feature domain. Assuming that the noise level in the b is negligible and the additive noise is uncorrelated with the speech signal, the problem can then be formulated as follows:

$$S_y(\omega) = S_x(\omega) + S_n(\omega) \tag{8}$$

$$S_x(\omega) = f(S_y(\omega), S_b(\omega)) \tag{9}$$

$$S_x(\omega) = H(\omega)S_y(\omega) \tag{10}$$

where S_y, S_x, S_b and S_n are the power spectrum for noisy speech , clean speech, bone signal, and noise, respectively, and f(z) is a nonlinear mapping function. Our goal is to find the optimal H (the Wiener filter). In the following, we omit ω for convenience.

Given two observations S_y and S_b under Gaussian approximation, we use MMSE estimator to find $\hat{S}_x$:

$$\hat{S}_x = (\Sigma_n^{-1} + \Sigma_{x|b}^{-1})^{-1}[\Sigma_n^{-1}(S_y - \mu_n) + \Sigma_{x|b}^{-1}\hat{S}_{x|b}] \tag{11}$$

where $\hat{S}_{x|b} = e^{C^{-1}\hat{x}}$

$$\hat{x} = b + \sum_s p(s|b)r_s$$

$$\Sigma_{x|b} = Var(S_{x|b})$$

$$\mu_n = E[S_n], \quad \Sigma_n = Var[S_n]$$

μ_n and Σ_n are estimated during non-speech section, which is detected by our speech detection algorithm.

Given S_y and $\hat{S}_x$, the estimated Wiener filter $\hat{H}$ can be calculated as:

$$\hat{H} = \frac{\hat{S}_x}{S_y} \tag{12}$$

6. EXPERIMENT RESULTS

The speech enhancement is an on-going work. Here we show some preliminary results. We collected about 150 words of clean speech of a male wearing the integrated microphone. We used these data for the SPLICE training. We then used a set of 24 words which are not in the training set as the test data. The test data is corrupted with another person's speech. We then apply our speech enhancement algorithm to estimate the clean speech.

To measure the quality of the reconstruction result, we conducted mean opinion score (MOS) [9] comparative evaluations. Table 2 shows the score criteria. To ensure a fair comparison, the non-overlapping noises (the noises between each two words) in the corrupted data are removed prior to MOS evaluation. The 24 words are divided into 4 groups each consisting of 6 words. The corrupted audio files and enhanced audio files are mixed randomly, and then played to the evaluators (the people who gave the scores) with desktop speakers. There are 4 evaluators and they do not know which files are corrupted and which ones are enhanced results. Table 3 shows the MOS results for the 0dB and 10dB cases. We observe that the enhanced data consistently gets better scores for all the evaluators. In the 0dB case, the improvement is more significant. Figure 6 shows the spectral view of the enhancement results. The top row is the corrupted data. The middle row is the result after enhancement. The bottom row is the clean speech (the ground truth). Clearly some of

the background speech are removed by our enhancement algorithm. For example, background speeches around 1KHz between 0.8s and 1.0s and around 5KHz and 7KHz between 0.2s and 0.4s are significantly reduced by our enhancement algorithm.

Score	Impairment
5 (Excellent)	Imperceptible
4 (Good)	(Just) Perceptible but not Annoying
3 (Fair)	(Perceptible and) Slightly Annoying
2 (Poor)	Annoying (but not Objectionable)
1 (Bad)	Very Annoying (Objectionable)

Table 2. MOS score criteria

	before enhancement	after enhancement
$0dB$	1.8	2.4
$10dB$	3.5	3.8

Table 3. MOS result. Each score in the table is averaged over 4 people and 4 different groups of test data.

7. CONCLUSION

We have presented an Air- and Bone-Conductive Integrated Microphone. This new hardware device has the look and feel of a regular headset. We have developed algorithms to use this new device for robust speech detection and speech enhancement in highly non-stationary noisy environment. We have shown that this new device can reduce most of the insertion errors between speeches without adding deletion or substitution errors. We also showed that this device can be used effectively for speech enhancement with highly non-stationary noises such as when people talking in the background.

8. FUTURE WORK

This is an on-going project. The reported results, although very encouraging, are preliminary. In the future, we would like to further improve our speech enhancement algorithm. We are planning on collecting more training data to improve the mapping from b to x. We would like to better estimate the noise by using a more sophisticated noise estimation algorithm.

9. REFERENCES

[1] T. Choudbury, James M. Rehg, Vladimir Pavlovic, and Alex Pentland, "Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection," in *ICPR*, 2002.

[2] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Coference on Multimedia Expo (ICME)*, 2000.

[3] P. deCuetos, C. Neti, and A. Senior, "Audio-visual intent-to-speak detection for human-computer interaction," in *ICASSP*, June 2000.

[4] Tsuhan Chen and Ram R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of IEEE*, vol. 86, no. 5, May 1998.

[5] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, and A. Verma, "Audio-visual large vocabulary continous speech recognition in the broad-cast domain," in *Workshop on Multimedia Signal Processing*, September 1999.

[6] Martin Graciarena, Horacio Franco, Kemal Sonmez, and Harry Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, March 2003.

[7] Jasha Droppo, Li Deng, and Alex Acero, "Evaluation of SPLICE on the aurora 2 and 3 tasks," in *International Conference on Spoken Language Processing*, September 2002, pp. 29–32.

[8] Xudong Huang, Alex Acero, and Xsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001.

[9] John R. Deller, John G. Proakis, and John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.

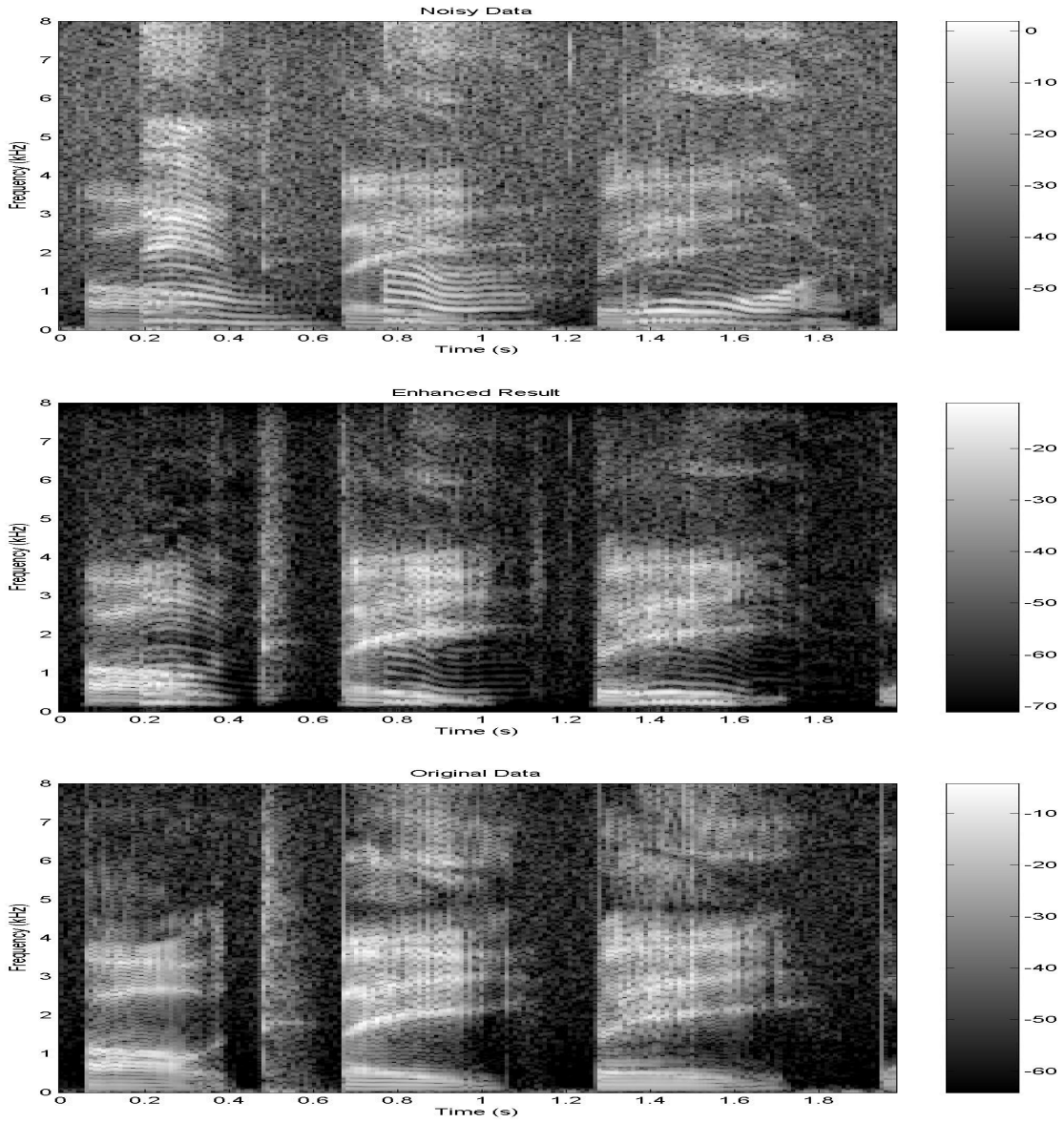

Fig. 6. The spectral view of the enhancement results. Top: the corrupted data. Middle: the enhancement result. Bottom: the clean speech.