MULTI-SENSORY MICROPHONES
FOR ROBUST SPEECH DETECTION, ENHANCEMENT AND RECOGNITION

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

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{zhang,zliu,sinclair,alexac,deng,jdroppo,xdh}@microsoft.com; [†]zheng3@ifp.uiuc.edu

ABSTRACT

In this paper, we present new hardware prototypes that integrate several heterogeneous sensors into a single headset and describe the underlying DSP techniques for robust speech detection, enhancement and recognition in highly non-stationary noisy environments. We also speculate other business uses with this type of devices.

1. INTRODUCTION

One of the most difficult problems for an automatic speech recognition system resides 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 speech is non-stationary, it is extremely hard to remove the background speech from just one channel of audio signals.

We at Microsoft Research recently started a project called WITTY (Who Is Talking To You) to deal with this and related problems. In this paper, we present a few hardware prototypes we have recently developed that integrate several heterogeneous sensors into a single headset and describe the underlying DSP techniques for robust speech detection and enhancement. Besides, we believe this type of devices has many other business applications, which are described in Section 5.

2. RELATED WORK

There has been a lot of work on using cameras to help with speech detection and recognition [1,2,3,4,5]. The work closely related to ours is that of Graciarena et al. [6]. They combined the standard and throat microphones for speech recognition in a noisy environment, and very good results were reported. There are three main differences between their work and ours. The first difference is in hardware. 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 is in the algorithms.

Their algorithm requires three-channel simultaneous recordings (clean close-talk, noisy close-talk and noisy throat microphone signals) to learn a piecewise linear mapping from the combined noisy feature vector of both microphones to the standard clean-speech feature vectors. It thus achieves its best performance only when the noise condition of the test data matches that of the training data. It is not clear how well their technique will work in simultaneous speech environments because of non-stationarity of the background speech, and they did not report any results on that. In comparison, our algorithm only requires two-channel simultaneous recordings (clean close-talk and clean bone or other sensor signals) to learn the mapping from the bone sensor to the close-talk. The predicted speech from the bone sensor is then fused with the noisy close-talk speech in order to reconstruct the clean speech signals. Our work aims at simultaneous speech environments. Finally, we also develop techniques for speech detection and speech enhancement. As a result, our headset can be used to feed cleaned signals into any existing speech recognition systems.

3. PROTOTYPES

We have developed several prototypes of multi-sensory headsets.

The first one, as shown in Figure 1 is a headset that combines regular close-talk microphone (air-conductive microphone) with a bone-conductive microphone. The device is designed in such a way that people wear and feel 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 (less than 3 KHz) 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 95% 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. Some preliminary results with this device have been reported in our ASRU paper [8].


Figure 1: Air- and bone-conductive integrated microphone headset
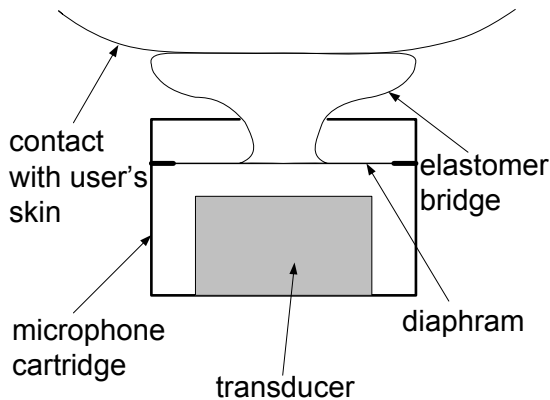
Figure 2: Diagram of a bone-conductive microphone

One design of a bone-conductive sensor is illustrated in Figure 2. In this sensor, a soft elastomer bridge is adhered to the diaphragm of a normal air-conductive microphone. This soft bridge conducts vibrations from skin contact of the user directly to the microphone. When the user talks, the user's skin vibrates, and the microphone converts the skin vibration into analog electrical signal. This signal is finally sampled by an analog-to-digital converter. Together with the sampled signal from the close-talk microphone, the air- and bone-conductive integrated microphone headset outputs two channels of audio signals.
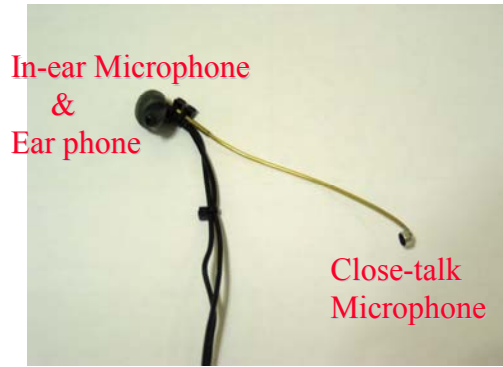
Figure 3: In-ear and air-conductive microphone

Figure 4: Throat and air-conductive microphone

Figure 5: Infrared augmented air-conductive microphone

The prototype shown in Figure 3 is designed for people who feel comfortable to wear an earphone. There is an in-ear microphone in side the earphone. The in-ear microphone picks up voice vibrations from the ear canal and/or surrounding bone. The signals from the in-ear microphone are, however, low quality, and are in the low-frequency range. To obtain better audio quality, a close-talk microphone is integrated. This system works in a similar fashion as the headset with bone-conductive microphone.

The prototype shown in Figure 4 is designed to be worn around the neck. The throat microphone is best positioned at either side of the user's "Adam's Apple" over the user's Voice Box. The audio quality from the throat microphone is not very high, but does not degrade significantly in noisy environment. We integrate an air-conductive microphone with the throat microphone in order to take advantages of both types of microphones.

The prototype shown in Figure 5 has an infra-red device attached to the close-talk microphone. The infrared device has a pair of near infrared LED and receiver. When the user adjusts the boom to point the microphone towards the mouth, the infrared device also points to the mouth. The LED emits (invisible) infrared signal and shines on the mouth. The receiver measures the amount of infrared lights reflected from the surface. When the mouth moves, the amount of light being reflected will vary due to both skin deformation and the fact that some lights that go inside the mouth will not be reflected. The infrared-enhanced microphone can thus robustly detect whether the speaker is talking or not even in very noisy environment.

4. ROBUST SPEECH DETECTION AND ENHANCEMENT WITH THE AIR- AND BONE-CONDUCTIVE INTEGRATED MICROPHONE

In this section, we provide some details on how the air- and bone-conductive integrated microphone is used for robust speech detection, enhancement, and recognition.

4.1. Speech Detection

In Figure 6, we show the speech signals from this integrated microphone when two people take turn to talk.
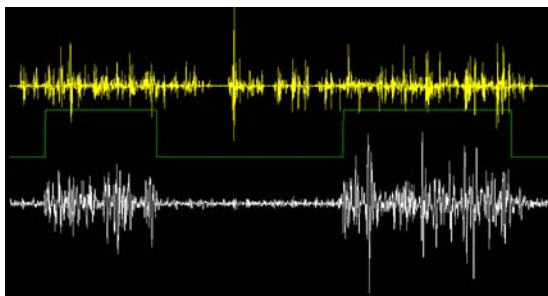
Figure 6: An example of speech detection

The yellow curve shows the signal from the close-talk microphone. The white curve shows the signals from the bone sensor. The local speaker (who wears the headset) and a background speaker were talking in an alternative way. The background speaker was talking loudly and clapped in the middle. The green curve shows our detection result. Notice that if we only look at the yellow curve, it is almost impossible to tell which section belongs

to the background speaker. But it is so obvious from the white curve. Details can be found in [8].

4.2. Speech Enhancement

In addition to speech detection, the integrated microphone can be used for speech enhancement when the speech signal is corrupted by highly non-stationary noise. Our idea is to train a mapping from the bone signals to the clean speech signals. Let b and y denote the observations from the bone and close-talk microphones, respectively. Let x denote the clean speech. To predict x from b, we use a technique that has some similarity to SPLICE [7]. 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. We use a piecewise linear representation to model the mapping from b to x in the cepstral domain.

$$p(x,b) = \sum_s p(x \mid b,s) p(b \mid s) p(s)$$

$$p(x \mid b,s) = N(x; b + r_s, \Gamma_s)$$

$$p(b \mid s) = N(b; \mu_s, \Sigma_s)$$

$p(b)$ contains 128 diagonal Gaussian mixture components trained by using a standard EM technique. r_s and Σ_s are trained using the maximum likelihood criterion.

$$r_s = \frac{\sum_n p(s \mid b_n)(x_n - b_n)}{\sum_n p(s \mid b_n)}$$

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

$$p(s \mid b_n) = \frac{p(b_n \mid s)p(s)}{\sum_s p(b_n \mid s)p(s)}$$

Finally, the estimated clean speech is

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

To obtain the waveform, we construct a Wiener filter by using the estimated cepstral features:

$$H = M^{-1} e^{C^{-1}(\bar{x} - y)}$$

where M is the matrix for Mel-frequency mapping, C is the DCT matrix. We then apply this filter to the waveform of the close-talk microphone signals to obtain the waveform of the estimated clean speech.

4.3. Speech Recognition

To measure the performance of speech enhancement, we used our integrate microphone to record 42 sentences for one person in an office with a different person talking at the same time. We then feed the waveform of the

estimated clean speech to a commercially available speech recognition system. For comparison, we also feed the signals from the close-talk microphone to the same speech recognition system. Table 1 shows the result. The top row is the result by using the single channel from the close-talk microphone. The bottom row is the result of using the enhanced signal. We see large reductions in both the insertion errors and the substitution errors. The accuracy is improved from the original 36% to 70%, or the word error rate is reduced by more than a half.

Table 1: Speech recognition result
(H: # correct words; D: # deleted words; S: # substitutions;
I: # insertions; N: total number of words)

	H	D	S	I	N
Without Enhancement	422	15	200	193	637
With Enhancement	504	9	124	59	637

5. APPLICATIONS AND BUSINESS IMPACTS

The multi-sensory headsets described above can find quite a number of applications that can considerably impact end-user experiences.

• Removal of the push-to-talk button. Recent Microsoft Office Usability study showed that users did not grasp the concept of using Press & Hold (Push to talk) to interact with Speech modes and that users would begin to speak concurrently with pressing the hardware buttons, leading to the clipping at the beginning of an utterance.

• Reduction/removal of background speeches. Again, according to a recent study, "People talking in the background" is identified as the most common noise source by 63% of respondents, followed by "Phones ringing" (59%) and "Air conditioning" (53%).

• Improving speech recognition accuracy in noisy environment. About half of speech users identify themselves in somewhat noisy environment, and the speech recognition accuracy is too low in noisy environment. By combining multiple channels of audio signals (such as bone microphone with normal microphone), we expect to improve the speech quality even in very noisy environment.

• Variable-rate speech coding. Since we know whether the person is talking or not, a much more efficient speech coding scheme can be developed to reduce the bandwidth requirement in audio conferencing.

• Floor control in Real-time communication. One important aspect that is missing in audio conferencing is lack of a natural mechanism to inform others that you want to talk; this may lead to the situation that one participant monopolizes the meeting. With the new headset, we can establish a convention to inform other participants that you

want to talk, e.g., by opening and closing your mouth a few times.

• Power management of a PDA/Tablet. Battery life is a major concern in portable devices. Through knowing whether the user is talking or not, we can allocate adequately the resources devoted to the DSP and speech recognition.

6. CONCLUSION

We have presented new hardware prototypes that integrate several heterogeneous sensors into a single headset and described techniques of using them for robust speech detection, enhancement and recognition in highly non-stationary noisy environment. Other possible applications were also discussed.

7. REFERENCES

1. Choudhury, T., Rehg, J. M., Pavlovic, V. and Pentland, A., "Boosting and structure learning in dynamic Bayesian networks for audio-visual speaker detection", *ICPR*, vol. III, pages 789-794, 2002.

2. Cutler, R., and Davis, L., "Look who's talking: Speaker detection using video and audio correlation", *IEEE International Conference on Multimedia Expo* (ICME), pages 1589-1592, 2000.

3. deCuetos, P., Neti, C., and Senior, A., "Audio-visual intent-to-speak detection for human-computer interaction", *ICASSP*, pages 1325-1328, June 2000.

4. Chen, T., and Rao, R. R., "Audio-visual integration in multimodal communication", *Proceedings of IEEE*, vol. 86, no. 5, pages 837-852, May 1998.

5. Basu, S., Neti, C., Rajput, N., Senior, A., Subramaniam, L., and Verma, A., "Audio-visual large vocabulary continuous speech recognition in the broad-cast domain", *Workshop on Multimedia Signal Processing*, pages 475—482, September 1999.

6. Graciarena, M., Franco H., Sonmez, K., and Bratt, H., "Combining standard and throat microphones for robust speech recognition", *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72-74, March, 2003.

7. Droppo, J., Deng, L., and Acero, A., "Evaluation of SPLICE on the aurora 2 and 3 tasks", *International Conference on Spoken Language Processing*, pages 29-32, September 2002.

8. Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A., and Huang, X, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement", *IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU03), Nov. 30 – Dec. 4, 2003.