# A Graphical Model for Multi-Sensory Speech Processing in Air-and-Bone Conductive Microphones

*Amarnag Subramanya*[†]*, Zhengyou Zhang, Zicheng Liu, Jasha Droppo, Alex Acero*[‡]

[†]SSLI Lab, University of Washington, Seattle, WA - 98195
[‡]Microsoft Research, One Microsoft Way, Redmond, WA - 98052.
asubram@ee.washington.edu,{zhang,zliu,jdroppo,alexac}@microsoft.com

## Abstract

In continuation of our previous work on using an air-and-bone-conductive microphone for speech enhancement, in this paper we propose a graphical model based approach to estimating the clean speech signal given the noisy observations in the air sensor. We also show how the same model can be used as a speech/non-speech classifier. With the aid of MOS (mean opinion score) tests we show, that the performance of the proposed model is better in comparison to our previously proposed direct filtering algorithm.

## 1. Introduction

Speech Enhancement is one of the oldest disciplines of signal processing. Though many techniques have been proposed to enhance speech in the presence of stationary background noise, enhancement in the presence of non-stationary background noise is still an open problem. In our previous work, [1, 2], we have developed a novel hardware solution to combat against highly non stationary acoustic noise such as background interfering speech. The device makes use of an inexpensive bone-conductive microphone in addition to the regular air-conductive microphone. The signal captured by the latter is corrupted by environmental conditions, whereas the signal in the former is relatively noise-free. The bone sensor captures the sounds uttered by the speaker but transmitted via the bone and tissues in the speaker's head. High frequency components ($>$ 3Khz) are absent in the bone sensor signal.

As explained above, the information from an air-and-bone conductive microphone (ABCM) consists of two channels, one which is corrupted by the ambient noise, and another which is the relatively noise free, but distorted. Thus, the challenge here is to enhance the signal in the air-channel by fusing the two streams of information. In [1], we proposed an algorithm based on the SPLICE technique to learn the mapping between the two streams and the clean speech signal. One drawback of this approach is that it requires prior training and therefore can lead to generalization problems. In the same work, we also proposed a speech detector based on a histogram of the energy in the bone channel. In [3], we proposed an algorithm called direct filtering (DF) that does not require any prior training in order to estimate the clean speech signal, i.e. the transfer function from the close-talking channel to the bone-channel is learned from the given utterance and the clean signal is estimated in a maximum likelihood framework. It was also shown that the performance of the DF algorithm is better in comparison to the SPLICE technique for speech enhancement. However, one drawback with the DF algorithm is the absence of a speech model, which can lead to distortion in the enhanced signal. One problem associated with the bone sensor signal in noisy environments is that a small amount of the environmental noise leaks into the sensor, which causes a significant drop in performance. In [4], we proposed an algorithm to remove this leakage by estimating the transfer function between the two sensors during non-speech frames. It was also found that in certain circumstances an artifact known as *teethclack* appears in the bone-sensor signal. Teethclacks are caused when the users' upper and lower jaws come in contact with each other during the process of articulation. For detailed discussion of how teeth clacks effect the estimation of clean speech signal and an algorithm for their removal, the reader is referred to [4].

In model-based speech enhancement algorithms, it is important to have accurate speech and noise models. The speech model captures the variability in the users speech, whereas the variability in environmental conditions is captured by the noise model. Owing to the large variability of speech, the speech model is usual trained offline whereas the noise model is computed online. Although algorithms have been proposed to estimate the noise model when both signal and noise are present [5], they have been successful only to a limited extent; therefore, invariably the noise model is estimated when the signal in absent. This requires accurate speech/voice activity detection. The approach proposed in [1], makes use of a function of the energy in the bone sensor. This approach has two problems associated with it: A) some classes of phones (e.g., fricatives) have low energy in the bone sensor causing false negatives; and B) leakage in the bone sensor can lead to false positives. Also, by using just the bone sensor for speech detection, we are not leveraging the two channels of information provided by the ABCM. In this paper, we propose an algorithm that takes into account the correlation between the two channels for speech detection. Further, as mentioned previously, one of the drawbacks of the DF algorithm is the absence of a speech model. The proposed graphical model based approach incorporates a speech model within its framework thereby reducing the amount of distortion in the enhanced signal.

## 2. Related Work

Graciarena et al. [6] 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, is not environment dependent and produces an audible speech signal so that the output can be used for perception as well as speech recognition.

This work was done at Microsoft Research

Strand et. al. [7] 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. [8] used a stethoscope device to capture the bone vibrations of the head and use that for non-audible murmur recognition. Like [7], they only used the bone signals for speech recognition with MLLR adaptation, while we use both bone and air signals.

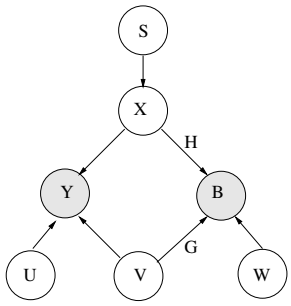


Figure 1: Proposed Model

## 3. Model Description

The proposed model is shown in figure 1. $S$ is a discrete random variable representing the state ($S = \{\text{speech, silence}\}$), $X$ represents the clean speech signal that is to be estimated, $Y$ is the signal captured by the air microphone, $B$ is the signal captured by the bone microphone, $V$ is the background noise, $U$ is the sensor noise in the air microphone channel, $W$ is the sensor noise in the bone microphone channel, $H$ is the optimum mapping from clean speech signal to bone sensor signal and $G$ models the background noise that leaks into the bone channel. The variables $X, Y, V, B$ are all in the complex frequency domain. All variables in the model are a function of time. Hence the figure shows the model for a given time $t$. For mathematical tractability, we assume that given $S_t$, the variables in the model are independent across both time and frequency. We assume the following distributions in the model, i.e.,

$$p(X_t|S_t) \sim N(X_t; 0, \sigma_s^2), p(V_t) \sim N(V_t; 0, \sigma_v^2) \quad (1)$$

$$p(U_t) \sim N(V_t; 0, \sigma_u^2) \;\&\; p(W_t) \sim N(W_t; 0, \sigma_w^2) \quad (2)$$

where $p(X_t|S_t)$ and $p(V_t)$ are the speech and noise models respectively. We also assume that the transfer functions $H$ and $G$ are known. As it can be seen, except for $Y_t$ and $B_t$ all variables in the model are hidden.

Note that in our current implementation, only two states (speech and silence) are considered, but the subsequent analysis is valid if more states (e.g., fricative, voiced, nasal) are used.

## 4. Estimating the Clean Speech Signal

In this section we provide a detailed description of how the clean signal may be inferred given the observations and show how a speech detector is a by-product of this inference engine. Our goal is to estimate the clean speech signal $X_t$ from the noisy observations $Y_t$ and $B_t$ using

$$p(X_t|Y_t, B_t) = \sum_{s \in \{S\}} p(X_t|Y_t, B_t, S_t = s) p(S_t = s|Y_t, B_t) \quad (3)$$

where $S = \{\text{speech, silence}\}$, $p(X_t|Y_t, B_t, S_t = s)$ is the likelihood of $X_t$ given the current observations and the state $s$, and $p(S_t = s|Y_t, B_t)$ is the likelihood of the state $s$ given the observations. Note that we assume that the state variable $S_t$ is not a function of frequency although theoretically $S_t$ can be frequency dependent.

### 4.1. Computing the posteriors

We have

$$p(X_t|Y_t, B_t, S_t = s) \propto p(X_t, Y_t, B_t, S_t = s) \quad (4)$$

In order to compute the joint probability $p(X_t, Y_t, B_t, S_t)$ we use

$$p(X_t, Y_t, B_t, S_t) = \int_V p(X_t, V_t, S_t, Y_t, B_t) dV \quad (5)$$

But the joint distribution over all the variables in the model factorizes as

$$p(X_t, V_t, S_t, Y_t, B_t) = p(Y_t|X_t, V_t) p(B_t|X_t, V_t) \\ p(X_t|S_t) p(V_t) p(S_t) \quad (6)$$

After some algebra we obtain,

$$p(X_t, S_t, Y_t, B_t) \sim N\big(Y_t; X_t, \sigma_u^2 + g^2\sigma_v^2\big) p(X_t|S_t) p(S_t) \\ N\left(G\frac{g^2\sigma_v^2(Y_t - X_t)}{\sigma_u^2 + g^2\sigma_v^2}; B_t - HX_t, \sigma_w^2 + |G|^2 \frac{g^2\sigma_v^2\sigma_u^2}{\sigma_u^2 + g^2\sigma_v^2}\right) \quad (7)$$

and according to our assumptions $p(X_t|S_t = s) \sim N(X_t; 0, \sigma_s^2)$.

### 4.2. Computing the Likelihood of the State

In this section we derive the expression to estimate the most likely state given the observations. We have

$$p(S_t = s|Y_t, B_t) = \int_X \frac{p(X_t, Y_t, B_t, S_t = s)}{p(Y_t, B_t)} dX \\ \propto \int_X p(X_t, Y_t, B_t, S_t = s) dX \quad (8)$$

We obtain the likelihood of the state given the observations using

$$p(S_t|Y_t, B_t) \propto N\left(B_t; \frac{(\sigma_s^2 H + g^2\sigma_v^2 G) Y_t}{\sigma_s^2 + g^2\sigma_v^2 + \sigma_u^2}, C\right) \\ N(Y_t; 0, \sigma_s^2 + \sigma_u^2 + g^2\sigma_v^2)\, p(S_t) \quad (9)$$

where

$$C = \sigma_w^2 + |G|^2 \frac{g^2\sigma_v^2\sigma_u^2}{\sigma_u^2 + g^2\sigma_v^2} + H_{mod} \frac{\sigma_s^2(\sigma_u^2 + g^2\sigma_v^2)}{\sigma_s^2 + \sigma_u^2 + g^2\sigma_v^2} \quad (10)$$

$$H_{mod} = \left|H - G\frac{g^2\sigma_v^2}{\sigma_u^2 + g^2\sigma_v^2}\right|^2 \quad (11)$$

The expression for $p(S_t|Y_t, B_t)$ is intuitively appealing: the first distribution models the correlation between the air and bone microphone channels whereas the second term makes use of the prior (along with variance and sensor noise in the air microphone channel) to explain the observation in the air microphone channel. The second term is important because we cannot rely on the correlation for classes of phones that are weak in the bone sensor (e.g. fricatives).

As it may be recalled, each of the variables in equation (9) is defined for a particular frequency bin in the complex spectral domain and hence may be used to compute likelihood of the state for each frequency independently. However, since we define a single state variable for each frame, we compute the likelihood of the state for a frame by aggregating the likelihood across the frequency bins as follows

$$L(S_t) = \prod_{\text{all } f} L(S_t(f)) \tag{12}$$

where $L(S_t(f)) = p(S_t(f)|Y_t(f), B_t(f))$ is the likelihood for frequency bin $f$ as defined in equation 9. If the likelihood computation is carried out in the log-likelihood domain, the product in the above equation is be replaced by a summation.

It can be easily seen that the likelihood computed above may be used build a speech/non-speech classifier based on a likelihood ratio test, i.e., if

$$g = \log \frac{L(S_t = \text{speech}|Y_t, B_t)}{L(S_t = \text{silence}|Y_t, B_t)} \tag{13}$$

then the frame at time $t$ is classified as speech if $g > 0$, and as a silence frame otherwise.

### 4.3. MMSE estimate of the clean speech signal

Coming back to the problem of estimating the clean speech signal. Recall that

$$p(X_t|Y_t, B_t) = \sum_{s\in\{S\}} p(X_t|Y_t, B_t, S_t = s)p(S_t = s|Y_t, B_t) \tag{14}$$

We use minimum mean square error estimator (mean of the posterior distribution) to obtain an estimate of the clean speech signal $X_t$, which gives

$$\begin{aligned}\hat{X}_t &= E(X_t|Y_t, B_t) \\ &= \sum_{s\in\{S\}} p(S_t = s|Y_t, B_t)E(X_t|Y_t, B_t, S_t = s)\end{aligned} \tag{15}$$

Using results from the last two sections we get

$$\begin{aligned}&E(X_t|Y_t, B_t, S_t = s) = \\ &\sigma_s^2\Big(\frac{\sigma_p^2 Y_t + M^*\big((\sigma_u^2 + g^2\sigma_v^2)B_t - g^2\sigma_v^2 G Y_t\big)}{\sigma_p^2\big(\sigma_u^2 + g^2\sigma_v^2 + \sigma_s^2\big) + |M|^2\sigma_s^2\big(\sigma_u^2 + g^2\sigma_v^2\big)}\Big)\end{aligned} \tag{16}$$

where

$$\sigma_p^2 = \sigma_w^2 + \frac{g^2\sigma_v^2\sigma_u^2}{\sigma_u^2 + g^2\sigma_v^2|G|^2} \tag{17}$$

$$M = H - \frac{g^2\sigma_v^2}{\sigma_u^2 + g^2\sigma_v^2}G \tag{18}$$

Thus, the MMSE estimate of the clean speech signal $X$ is given by

$$\hat{X}_t = \sum_{s\in\{S\}} \pi_s^* E(X_t|Y_t, B_t, S_t = s) \tag{19}$$

where $\pi_s^*$ is the posterior on the state and is given by

$$\pi_s^* = \frac{L(S_t = s)}{\sum_{s\in\{S\}} L(S_t = s)} \tag{20}$$

where $L(S_t = s)$ is given by equation (12).

| 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 1: MOS Evaluation Criteria

## 5. Parameter Estimation

As stated in section 3, we assume that $H$ and $G$ are known. Details about the estimation of these transfer functions may be obtained in [3, 4]. The parameters $\sigma_s^2, \sigma_u^2$, and $\sigma_v^2$ are obtained using the following formulation: we run an energy based speech detector ([1]) to obtain an initial estimate of the variances. These estimates are then used as input to the model and subsequently used to estimate the state of each frame in an utterance. Then

$$\sigma_v^2 = \sum_{t\in N_v} |Y_t|^2, \sigma_w^2 = \sum_{t\in N_v} |B_t|^2 \tag{21}$$

where $\{N_v\}$ is the set of frames classified as non-speech. Also we set $\sigma_u^2 = 10^{-4}\sigma_w^2$. This is based on empirical studies and the observation that close-talk sensor technology is more advanced than bone-sensor technology. In order to estimate $\sigma_s^2$, we use

$$\sigma_{s,t}^2 = \beta|X_{t-1}|^2 + (1-\beta)\max(|Y_t|^2 - \sigma_v^2, \alpha) \tag{22}$$

where a small value for $\alpha$ results in a large amount of noise reduction at the cost of more distortion whereas a larger value of $\alpha$ leads to lesser noise removal. In our experiments we have found that setting $\alpha$ to $0.1$ yields good results. To ensure smoothness of $\sigma_s^2$ over time we use $\beta|X_{t-1}|^2$ in the above computation. In our current implementation we set $\beta = 0.1$. Thus, it can be seen that we do not use any prior models in this approach. All parameters are estimated from the given utterance.

## 6. Experimental Setup

To measure the quality of the enhanced utterances, we conducted mean opinion score (MOS) [9] comparative evaluations. Table 1 shows the score criteria. We selected 16 utterances recorded in real-world environments (such as cafeteria, office with background speakers, car with radio/stereo in operation, etc.) with an equal proportion of male and female speakers. The ambient noise in these recordings varied from $75dbc$ to $85dbc$. Each speaker wears a head-set (as depicted in [3]) that consists of a close-talking and a bone-microphone. Each utterance was processed using three algorithms, (a) the classical spectral subtraction, (b) the DF algorithm [4, 3] and (c) the proposed algorithm. In the case of spectral subtraction, the bone signal was only used to manually segment the utterance into regions of speech and non-speech. The non-speech frames were used to obtain a noise profile which was input to the spectral subtraction algorithm. The DF algorithm used is detailed in [3]. In the case of the proposed and DF algorithms all processing was done without any manual intervention.

For every given noisy utterance, there were 3 processed utterances resulting in a set of 4 utterances and 16 such sets (one for each utterance). There were a total of 17 participants in MOS test. The evaluators were presented with a random ordering of the sets of utterances and random ordering within a

| Original | SS | DF | GM |
|---|---|---|---|
| 2.2650 | 2.2063 | 2.9062 | 3.8313 |

Table 2: Mean Opinion Score (MOS) Results: SS-Spectral Subtraction, DF - Direct Filtering, GM - Proposed Graphical Model

set. The participants were blind to the relationship between the utterances and the processing algorithm.

## 7. Results

The results of the MOS tests are shown in table 2. It can be seen the proposed algorithm out performs the other enhancement algorithms. It is interesting to note that the utterances processed by spectral subtraction were less favorable to the listeners when compared to the original utterance. One reason for this could be the non-stationary nature of the corrupting noise in most of the utterances leading to distortion. Figure 2 shows the signal captured by the bone sensor, spectrograms of the signal in the air microphone, results of the DF and the proposed algorithms for a noisy utterance. The utterance was recorded when a user was speaking in the presence of a group of people acting as background speakers with an ambient noise level of 75 dbc. The first figure also shows the actual probability of speech (dotted-line) obtained by manual segmentation. Note that signal and noise both overlap in the utterance in question, hence in the case of manual segmentation, a frame was classified as speech if signal from the user was present. A comparison of the results of the DF and proposed algorithms shows that: a) the proposed algorithm results in improved speech detection, for example, the initial non-speech frames are eliminated by the proposed algorithm but the DF algorithm only attenuates them, and b) signal is enhanced with lesser amount of distortion.

## 8. Conclusions & Future Work

In this paper we have proposed a graphical model based approach to enhancing the corrupted signal in a ABCM that does not require any prior-trained models. The proposed model is different and better than our previously proposed algorithms because it uses a better (two state) model for speech and a better speech detector based on the correlation between the two channels resulting in a better estimate of the noise model. The MOS tests show that the proposed algorithm is able to remove significant amounts of noise in the utterance while maintaining the distortion levels at a minimum.

We are currently working on a system where the noise can be estimated recursively in an EM framework. Another possible area is the expansion of the state space. One short-coming of the proposed algorithm (and [3]) is the frequency independence assumption. Therefore, yet another possible area of future work would be towards relaxing the independence assumption while keeping the models mathematically tractable.

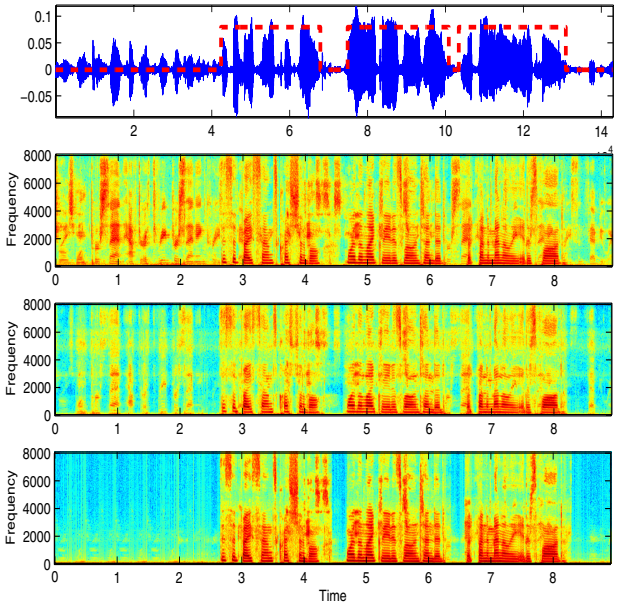


Figure 2: Enhancement Results (sequentially from top to bottom): a) the original bone sensor data; b) spectrogram of noisy utterance; c) result of the DF algorithm; c) result of the proposed model.

## 9. References

[1] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X Huang. "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," *Proc. IEEE ASRU Workshop*, Dec. 2003, St. Thomas, US Virgin Islands.

[2] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo. X. Huang, Y. Zheng. "Multisensory microphones for robust speech detection, enhancement, and recognition," *Proc. ICASSP*, Montreal, Canada, May 2004.

[3] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang. "Direct filtering for air- and bone-conductive microphones." *Proc. MMSP*, Siena, Italy, Sept. 2004.

[4] Z. Liu, A. Subramanya, Z. Zhang, J Droppo, and A. Acero, "Leakage Model and Teeth Clack removal for Air-and-Bone conductive microphones", *Proc. of ICASSP, Philadelphia*, 2005.

[5] Chen J., Huang Y., and Benesty J. "Filtering techniques for noise reduction and speech enhancement" Adaptive Signal Processing: Applications to Real-World Problems, J. Benesty and Y. Huang, Eds., pp. 129154, Berlin, Germany: Springer, 2003.

[6] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," IEEE Signal Processing Letters, 2003, vol. 10, pp. 7274.

[7] 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, 2003.

[8] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Accurate hidden markov models for non-audible murmur (nam) recognition based on iterative supervised adaptation," in ASRU, St. Thomas, U.S. Virgin Islands, 2003.

[9] X. Huang,A. Acero, and X-H. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice Hall PTR, 2001.