Model-Based Fusion of Bone and Air Sensors for Speech Enhancement and Robust Speech Recognition

John Hershey, Trausti Kristjansson, Zhengyou Zhang

Microsoft Research
One Microsoft Way
Redmond, WA, USA
{hershey,traustik,zhang}@microsoft.com

Abstract

We present a probabilistic framework that uses a bone sensor and air microphone to perform speech enhancement for robust speech recognition. The system exploits advantages of both sensors: the noise resistance of the bone sensor, and the linearity of the air microphone. In this paper we describe the general properties of the bone sensor relative to conventional air sensors. We propose a model capable of adapting to the noise conditions, and evaluate performance using a commercial speech recognition system. We demonstrate considerable improvements in recognition – from a baseline of 57% up to nearly 80% word accuracy – for four subjects on a difficult condition with background speaker interference.

1. Introduction

Automatic speech recognition systems are notoriously susceptible to error in the context of interfering noise[1]. This is especially true when the noise is a background speaker, and it may be difficult to determine which voice is intended for the speech recognizer. The speech research group at Microsoft Research has recently been exploring methods to address this problem using multiple sensors, as part of an ongoing project called WITTY (Who Is Talking To You).

One promising technology involves the use of a bone sensor along with a conventional microphone. A bone sensor is a microphone that directly touches the side of a person's face directly in front of the ear. The bone sensor can easily be incorporated into the standard head-mounted headset with a close-talk air microphone. A prototype of such a device is illustrated in Figure 1, and has been described previously in [1].

The advantage of the bone sensor is that it is much more immune to external interfering sounds than a regular microphone. However, the response to higher frequencies, as well as to aspects of speech dynamics is poor. (You can get a good impression of what it sounds like by talking with your ears plugged.) Thus existing speech recognition systems perform poorly when given the bone signal as input. Unfortunately there is not enough data recorded with such devices to create a recognizer especially tuned to bone signals.

Instead of direct recognition, we focus on enhancing speech prior to recognition. Due to variability in the relationship between air and bone signals, it is a challenging problem to map from one to the other. In this paper we discuss the combination of an air microphone and a bone sensor in a probabilistic framework to do speech recognition and speech enhancement.

Related work has been presented in [2], and [1], in which different types of models were used. One important difference between the current work and these works is the principled inclusion of an adaptive noise model.

2. The Relationship Between Air and Bone Signals

The air and bone microphones are sensitive to different aspects of the speech signal, and their spectra differ as a function of the speaker, the placement of the sensors, and as a function of the articulation of speech itself. The air sensor receives acoustic signals coupled primarily through the mouth aperture (and somewhat less through the nose), whereas the bone sensor receives vibrations that are conducted from the rear of the vocal tract through the facial anatomy. We can think about the relationship between signals in the air and bone sensors in terms of differences in the log amplitude spectra of the two signals.

A major difference between the two sensors is in the overall frequency response of the signals they receive. Due to its placement, the bone microphone picks up signals that have propagated through flesh and bone, which absorbs high-frequency components of speech. Thus the bone sensor receives predominantly low-frequency components of speech, peaking sharply at around 400Hz, with the response dropping by about 20 dB to a plateau stretching between 1KHz and 2.5KHz. For higher frequencies,

[1] We use the term *noise* in the sense of an interfering signal, without implying anything about its statistical properties.

Figure 1: Air and bone sensors mounted on a headset

attenuation increases to about 40dB at 4kHz, and remains essentially at the noise floor thereafter (see Figure 2 and 3). The air microphone is sensitive to the full range of speech frequencies.

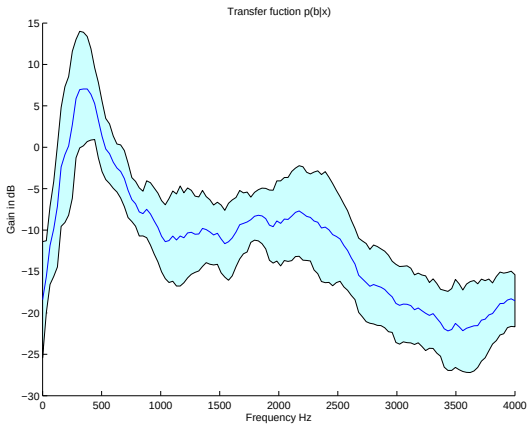


Figure 2: The difference in log magnitude between the signal received at the air and the bone sensors. Error bars indicate plus or minus one standard deviation.

The broad characteristics of the relationship between air and bone sensor signals largely reflects the bone structure and facial tissue through which the signal propagates to reach the bone sensor. These physical characteristics can vary from speaker to speaker, altering the quality of the signals. This effect can be termed *speaker dependence*. In addition, different placement of the air microphone and bone sensor introduces further differences in the signals. This effect may be termed *placement dependence*. In our experiments we assume both speaker and placement dependence effects are stationary and control them by training speaker-dependent models, and ensuring that the sensor placement in training and test sets is the same.

A particularly strong and non-stationary effect can be termed *articulation dependence*. Different speech phones entail different distributions of source energy, and different patterns of its coupling to the two sensors. These physical differences result in pronounced differences in the log spectra between the two sensors. Figure 3 illustrates the smoothed log spectrum of examples of the three phones /A/, /n/, and /t/, taken from a single sentence, for the air and bone sensors.

One effect of articulation derives from the closing or *stricture* of the oral tract during speech. When the mouth is open as in a vowel sound such as /A/, the acoustic energy is well coupled to both the air and bone sensors. However when the mouth is restricted, as in a nasal stop such as /n/, the acoustic coupling to the air sensor is greatly attenuated relative to the bone sensor.

Another effect of articulation has to do with the location and manner of the generation of acoustic energy. Voiced speech sounds originate in the throat with the vocal cords and are well transmitted to the nearby bone sensors. In contrast, fricatives, such as /t/, are generated at the place of articulation, as turbulent air passes through an aperture in the mouth and are thus typically much more readily transmitted to the air sensor than to the bone sensor.

A third articulation effect is an artifact of the bone sensor's frequency response. Because the bone sensor is effectively at the noise floor for frequencies above 2kHz, differences in the log spectrum between air and bone sensors will simply reflect the energy in the air sensor for those frequencies. This energy varies greatly with the phone, with fricatives such as /t/ having a great deal of high-frequency energy, in contrast to nasal stops such as /n/.

3. Models

We seek to capture the statistics of the relationship between air and bone signals in a model, in a way that allows us to efficiently infer the air signal in the presence of acoustic interference, or noise. The complex relationship between the two sensors motivates a model that supports a flexible mapping between the two.

A simple and tractable model that accomplishes this is a Gaussian mixture model on the high-resolution log spectra of each sensor, with the frequency components conditionally independent given the state.

In previous work, a Mel-frequency scale was used for compatibility with speech recognition systems designed for clean speech. However it was subsequently discovered that the greater frequency resolution of the linear frequency scale (for a given window size) allows us to take advantage of the harmonic structure of voiced speech [3]. Because the bone sensor preserves pitch well, but with a

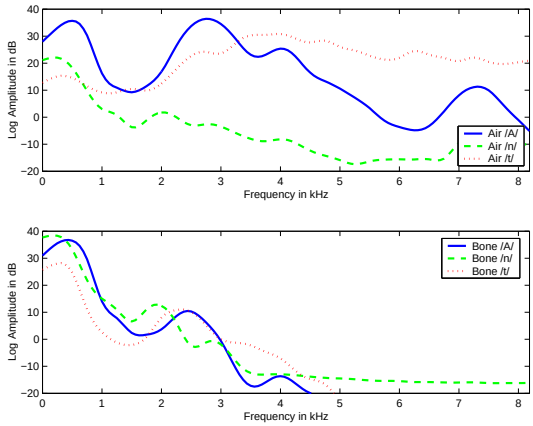


Figure 3: Smoothed log spectra received at the air (top) and bone (bottom) sensors for three phones /A/ (as in "ape"), /n/ and /t/, showing the differences in the relative patterns of the spectra in each sensor. In particular, notice how the /n/ spectrum is dramatically attenuated in the air sensor relative to the other phones, but not in the bone sensor.

limited bandwidth, we suppose that such *high-resolution* models will allow us to extrapolate harmonics from the bone sensor to the air sensor, and filter out noise between the harmonics.

Whereas high-resolution models have the advantage that they can focus on high signal-to-noise ratio peaks at the harmonics, they also have the drawback that they require many states to represent the detail in the spectrum. To avoid wasting representational resources on the attenuated components of the bone signal, which can have negligible impact on the results, we simply discard bone sensor frequency components above 4kHz from consideration in the model.

3.1. Speech model

The model we propose is illustrated in Figure 4, and can be defined as follows. Denote the log amplitude of a windowed short time fourier transform of the clean air signal and bone signal respectively, for frequency f as x_f^a and x_f^b. We form a mixture model on x_f^a and x_f^b with discrete speech state as s_x. The conditional independence assumptions in such a model are given by the following factorization:

$$p(x^a, x^b, s^x) = p(s^x) \prod_f p(x_f^a|s^x) p(x_f^b|s^x) \qquad (1)$$

It is convenient to use a Gaussian for $p(x_f^a|s^x)$ and $p(x_f^b|s^x)$. Let $N(x;\mu,\sigma)$ denote the univariate normal distribution defined on x with mean mean μ and variance σ. The model used in our experiments can then be formulated as follows.

$$\begin{aligned} p(s^x) &= \pi_{s^x} \\ p(x_f^a|s^x) &= N(x_f^a; \mu^a_{s^x,f}, \sigma^a_{s^x,f}) \\ p(x_f^b|s^x) &= N(x_f^b; \mu^b_{s^x,f}, \sigma^b_{s^x,f}) \end{aligned}$$

Note that this model is equivalent to a simple Gaussian mixture model formulated on the concatenated air and bone spectrum vectors. The parameters pi_s, $\mu^a_{s^x,f}$, $\mu^n_{s^x,f}$, $\sigma^a_{s^x,f}$ and $\sigma^b_{s^x,f}$ can thus be estimated from bone and air microphone recordings in quiet conditions, using the standard expectation maximization (EM) algorithm [4].

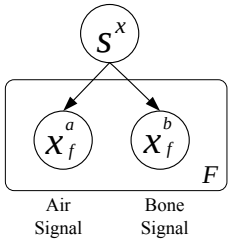


Figure 4: Generative model for air and bone sensors. s^x is the discrete state, x_f^a and x_f^b are the log spectra at frequency f of air and bone signals respectively. The plate labeled F indicates that the series of frequency components indexed by f are all all conditionally independent given the state s^x.

3.2. Noise Model

A similar model is posited for the noise.

$$\begin{aligned} p(s^n) &= \pi_{s^n} \\ p(n_f^a|s^n) &= N(n_f^a; \mu^a_{s^n,f}, \sigma^a_{s^n,f}) \\ p(n_f^b|s^n) &= N(n_f^b; \mu^b_{s^n,f}, \sigma^b_{s^n,f}) \end{aligned}$$

Because the noise is usually unknown, we are interested in adapting the parameters to the noise at test time. We therefore use a small number of states to avoid overfitting. Section 5 describes the noise adaptation.

3.3. Sensor Model

To complete the model we have to specify how the speech and noise combine in the air and bone sensors. Because the speech and noise models are defined in the log spectrum, but the speech and noise signals are combined in the linear spectrum with unknown phases, their combination is nonlinear and results in analytically intractable distributions for the observed sensor signals, and for the posterior distribution of the hidden speech and noise components. For simplicity we shall describe the model in terms of a generic sensor signal model, y_f, x_f and n_f, since the same model applies to both air and bone sensors.

The model for a given frame of noisy speech in the frequency domain is

$$Y_f = X_f + N_f \tag{2}$$

where X_f, N_f, and Y_f denote the complex Fourier transform at frequency f of the the clean signal, the noise, and the noisy sensor signal respectively. This can also be written in terms of the magnitude and the phase of each component:

$$|Y_f|\angle Y_f = |X_f|\angle X_f + |N_f|\angle N_f \tag{3}$$

where $|Y_f|$ is the magnitude of Y_f and $\angle Y_f$ is the phase.

We model only the magnitude components and do not explicitly model the phase components. The relationship between the magnitudes is

$$|Y_f|^2 = |X_f|^2 + |N_f|^2 + 2|X_f||N_f|\cos(\theta) \tag{4}$$

where θ is the phase angle between X and N. Next we take the logarithm, defining $x_f \triangleq \ln|X_f|^2$, and likewise for y_f and n_f. We arrive at the relationship in the high resolution log-power-spectrum domain.

$$y_f = \ln\left[\exp(x_f) + \exp(n_f)\right] + \varepsilon \tag{5}$$

where

$$\varepsilon = \ln\left[1 + 2\cos(\theta)\frac{\sqrt{\exp(x_f + n_f)}}{\exp(x_f) + \exp(n_f)}\right] \tag{6}$$

The formulation in terms of x_f plus a correction term will be convenient for taking derivatives later. We approximate ε as Gaussian noise, as in [5], and write this relationship in terms of a distribution over the noisy speech features y_f as

$$p(y_f|x_f, n_f) = N(y_f; \ln\left[\exp(x_f) + \exp(n_f)\right], \psi) \tag{7}$$

where ψ is the variance of ε. Duplicating this model for both the air and bone sensors, and we can combine it with the speech and noise models introduced above as illustrated in Figure 5.

4. Inference

For the purpose of signal reconstruction, we are interested in the expected value of the clean speech given the noisy speech in both the air and bone sensors, or $E(x^a|y^a, y^b)$, where we write the components in vector form with $x^a \triangleq [x_1^a ... x_F^a]^T$, and similarly with x^b, n^a, and n^b. To do so we have to estimate the posterior distribution $p(x^a|y^a, y^b)$. The true posterior is a mixture of individual posteriors:

$$p(x^a|y^a, y^b) = \sum_{s^x, s^n} p(s^x, s^n|y^a, y^b)\, p(x^a|y^a, y^b, s^x, s^n) \tag{8}$$

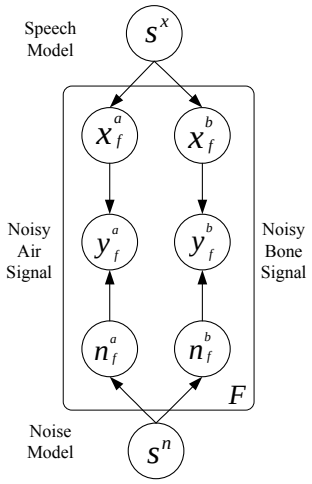


Figure 5: Model of air and bone speech signals (x_f^a,x_f^b), noise (n_f^a,n_f^b), and their interaction in bone and air sensors (y_f^a,y_f^b). Note that as indicated by the plate labeled F, all the frequencies in both sensors are conditionally independent given a combination of speech state s^x and noise state s^n.

The individual mixture components decouple due to conditional independence given the state, so $p(x^a|y^a, y^b, s^x, s^n) = p(x^a|y^a, s^x, s^n)$, and we have:

$$p(x^a|y^a, s^x, s^n) = \frac{1}{z} p(x^a|s^x) \int p(y^a|x^a, n^a)\, p(n^a|s^n) dn^a, \tag{9}$$

where z is a normalizing constant. This posterior is non-Gaussian and analytically intractable, due to the non-linearity of $p(y^a|x^a, n^a)$ in Equation 7. The state posteriors, $p(s^x, s^n|y^a, y^b)$ in Equation 8 are analyticaly intractable for the same reason.

To solve this problem we iterate a Laplace method to approximate the posteriors in a framework known as Algonquin [5].

For notational convenience, we define

$$x \triangleq \begin{bmatrix} x^a \\ x^b \end{bmatrix}, n \triangleq \begin{bmatrix} n^a \\ n^b \end{bmatrix}, \text{ and } z \triangleq \begin{bmatrix} x \\ n \end{bmatrix},$$

so that we can write the mean of Equation. (7) in vector form:

$$g(z) \triangleq \ln\left[\exp(x_f) + \exp(n_f)\right]. \tag{10}$$

If we linearize this function using a first order Taylor series expansion at the point z_0, we can write the linearized version of the likelihood,

$$p_l(y|x, n) = p_l(y|z) = N(y; g(z_0) + G(z_0)(z - z_0), \Psi) \tag{11}$$

where z_0 is the linearization point and

$$G(z_0) = \left[\text{diag}\left(\frac{\partial g(z)}{\partial x}\right), \text{diag}\left(\frac{\partial g(z)}{\partial n}\right)\right]_{z_0} \tag{12}$$

is a matrix of the derivatives of $g(z)$, evaluated at z_0. We can now write a Gaussian approximation to the posterior for a particular speech and noise combination as

$$p_l(x,n,y|s^x,s^n) = p_l(y|x,n)p(x|s^x)p(n|s^n) \quad (13)$$

For notational convenience we abbreviate $s \triangleq \{s^x, s^n\}$, and define $\sigma^a_{s^x} \triangleq [\sigma^a_{s^x,1},...,\sigma^a_{s^x,F^a}]^T$ and similarly with $\sigma^b_{s^x}$,$\sigma^a_{s^n}$,and $\sigma^b_{s^n}$. We combine air and bone sensor parameters thus:

$$\sigma_{s^x} \triangleq \begin{bmatrix} \sigma^a_{s^x} \\ \sigma^b_{s^x} \end{bmatrix}, \sigma_{s^n} \triangleq \begin{bmatrix} \sigma^a_{s^n} \\ \sigma^b_{s^n} \end{bmatrix}, \text{and } \Sigma_s \triangleq \text{diag}^{-1} \begin{bmatrix} \sigma_{s^x} \\ \sigma_{s^n}, \end{bmatrix}$$

and similarly with $\mu^a_{s^x}$,$\mu^b_{s^x}$,$\mu^a_{s^n}$,$\mu^b_{s^n}$. Ψ is a diagonal matrix of the variances of Equation 7, which we set to a constant, (i.e, $\Psi = (.01)I_{F^a+F^b}$). It can then be shown[5] that the $p(x,n|y,s)$ is jointly Gaussian with mean

$$\eta_s = \Phi_s \left[\Sigma_s^{-1}\mu_s + G(z_0)^T\Psi^{-1}(y - g(z_0) - G(z_0)z_0)\right] \quad (14)$$

where

$$\eta_s \triangleq \begin{bmatrix} \eta^x_{s^x s^n} \\ \eta^n_{s^n s^n} \end{bmatrix}$$

and covariance matrix

$$\Phi_s = \left[\Sigma_s^{-1} + G(z_0)^T\Psi^{-1}G(z_0)\right]^{-1}, \quad (15)$$

where

$$\Phi_s \triangleq \begin{bmatrix} \Phi^{xx}_{s^x s^n} & \Phi^{xn}_{s^x s^n} \\ \Phi^{nx}_{s^x s^n} & \Phi^{nn}_{s^x s^n} \end{bmatrix}$$

The posterior state probability $p(s|y)$ can be shown to be

$$\gamma_s = |\Sigma_s|^{-1/2}|\Psi|^{-1/2}|\Phi_s|^{1/2} \cdot \exp\Bigg[-\frac{1}{2}(\mu_s^T\Sigma_s^{-1}\mu_s + (y - g(z_0) + G(z_0)z_0)^T\Psi^{-1}(y - g(z_0) + G(z_0)z_0) - \eta_s^T\Phi_s^{-1}\eta_s)\Bigg]. \quad (16)$$

In the Algonquin algorithm, we attempt to iteratively move the linearization points towards the mode of the true posterior. In each iteration the mode of the approximate posterior in the previous iteration is used as a linearization point of the likelihood. The algorithm converges in three to four iterations.

It is then a simple matter to find the marginal expected value of the speech given the noise:

$$\hat{x^a} = \frac{\sum_s \gamma_s \eta^a_s}{\sum_s \gamma_s}. \quad (17)$$

Once the log spectral energies are inferred, we compute magnitudes and combine them with the phases from the air sensor, to resynthesize the enhanced waveform for a given frame. The waveforms are overlapped and added together across frames using a synthesis window derived from the analysis window, such that the product of the two windows overlapped and added across frames sums to unity everywhere.

5. Adaptation

Noise conditions vary from one environment to the next. We therefore wish to adapt the noise model to the current noise conditions. Since the algorithm above provides posteriors over the noise as well as the signal, we can perform an extra adaptation step in which we adjust the parameters of the noise model, as in [6]. This adaptation comprises an M-step to in a generalized expectation-maximization framework, to maximize the expected log-likelihood of the data with respect to the posteriors.

The resulting update equations are:

$$\widehat{\pi_{s^n}} \leftarrow \left\langle \sum_{s^x} \gamma_{s^x s^n t} \right\rangle_t \quad (18)$$

$$\widehat{\mu_{s^n}} \leftarrow \left\langle \sum_{s^x} \frac{\gamma_{s^x s^n t}}{\widehat{\pi_{s^n}}} \eta^n_{s^x s^n t} \right\rangle_t \quad (19)$$

$$\widehat{\sigma^n_s} \leftarrow \left\langle \sum_{s^x} \frac{\gamma_{s^x s^n t}}{\widehat{\pi_{s^n}}} \text{diag}\left[\Phi^{nn}_{s^x s^n t} + (\eta^n_{s^x s^n t} - \mu_{s^n})(\eta^n_{s^x s^n t} - \mu_{s^n})^T\right] \right\rangle_t, \quad (20)$$

where $\langle \cdot \rangle_t$ denotes averaging over time. This can be done in batch mode, as in the experiments presented here. Alternately it can be handled via running averages for online adaptation.

6. Results

We trained speaker-dependent speech models on a database of four subjects (two males, two females) reading 41 sentences from the Wall Street Journal. In the training set, speakers were recorded in a quiet office environment with the air and bone sensors, and in the test set they were recorded in the same environment with an interfering male speaker talking loudly a short distance away. The sound was 16-bit 16 kHz, and was processed in 50ms windows in 256 frequency bands, and with a frame shift of 20ms. We trained speech models using 512 states, although 256 states worked nearly as well. Noise models had 2 states.

Prior to processing we smoothed the log spectrograms temporally by applying a smoothing kernel, $([1,2,1]\frac{1}{4})$, across time frames in each frequency bin. This reduces variance in the spectrum and stabilizes the inference. In order to speed up testing we eliminated extremely unlikely states by approximating the likelihood of y_a and y_b given the states by matching moments to the log-normal sum (see [7] for a derivation) to estimate the state posteriors for each frame. We then retained only the four most likely speech states for each frame.

We explored two independent variables, *sensor condition* and *noise mode*. In order to compare enhancement with just the air sensor (*A* sensor condition) to enhancement with both the air and bone sensors (*AB* sensor condition), we trained both air sensors models and air plus

bone sensor models. After looking at preliminary results, we wondered if perhaps the bone sensor wasn't having enough influence on the inferred air signal. So we added a third condition in which the speech state posterior was determined solely on the basis of the bone sensor observation, rather than using both the air and bone sensors (*AB** sensor condition). Apart from the computation of the state posteriors, the rest of the inference in this condition is the same as in the AB condition.

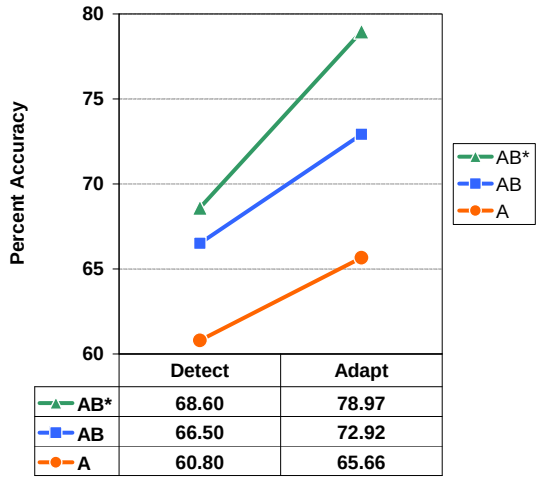
The noise model was always initialized using speech detection on the bone microphone to determine how much of the beginning segment of the file was free of speech, with a minimum of 300ms being used. The model was initialized by training on this detected noise segment. We then compared performance without adaptation (*Detect* noise mode) to performance with the adaptation (*Adapt* noise mode) described in Section 5.

The enhanced results were tested using a commercial speech recognition system, and are shown in Figure 6 averaged across the four subjects. Baseline percent accuracy [2] in the air microphone was 57.2% for the noisy condition, and 92.66% for the clean. As we had hoped, performance appears to be better with the bone sensor than without, and better with adaptation than without. In the AB* condition where we relied more heavily on bone signals to do inference, performance was better still. The best condition, AB* with adaptation yield accuracies of around 79%, or a relative improvement in word error rate of about 51%. Although we cannot make a strict comparison for lack of a standardized dataset, these results compare favorably with prior art in [2], and [1].

7. Conclusion

We have demonstrated a model that exploits a bone sensor combined with air microphone to produce noise adaptation speech enhancement results that are much better than could be achieved with an air microphone alone. The model we proposed posits conditional independence of the bone and air signals given the state. We are currently working on several improvements, including a model with a direct state-dependent correlation between the air and bone sensor. In informal analysis it seemed that this correlation could help to distinguish the right speech state. In the same context we are working on online noise adaptation as well as adaptation to varying channel characteristics, which may be important for the development of a speaker independent system. In general, although results are preliminary, the bone sensor technology and the proposed models appear very promising.

[2] Percent accuracy is $100\frac{N-I-S-D}{N}$, where where I is the number of inserted words, S is the number of substituted words, D is the number of deleted words and N is the total number of words in the transcription.



	Detect	Adapt
AB*	68.60	78.97
AB	66.50	72.92
A	60.80	65.66

Figure 6: Results: All numbers are percent accuracy (defined in Footnote 2) of word recognition. Sensor conditions are, *A*: air sensor only; AB: Air and bone sensors; AB*: air and bone sensor with state posteriors derived only from the bone sensor. Noise mode conditions are, *Detect*: initialize the noise model with speech detection on the bone sensor; *Adapt*: adapt to the noise in the test sentence

8. References

[1] Zhengyou Zhang, Zicheng Liu, Mike Sinclair, Alex Acero, Li Deng, Jasha Droppo, Xuedong Huang, and Yanli Zheng, "Multi-sensory microphones for robust speech detection, enhancement, and recognition," in *ICASSP*, Montreal, May 17-21 2004.

[2] M. Graciarena, Franco H., 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.

[3] Trausti Kristjansson and John Hershey, "High resolution signal reconstruction," *ASRU*, 2003.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Proceedings of the Royal Statistical Society*, vol. B 39, pp. 1–38, 1977.

[5] T. Kristjansson, *Speech Recognition in Adverse Environments: A Probabilistic Approach*, Ph.D. thesis, University of Waterloo, Waterloo, Ontario, Canada, April 2002.

[6] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero, "Learning dynamic noise models from noisy speech for robust speech recognition," *Advances in Neural Information Processing (NIPS)*, 2001.

[7] Mark .J.F Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Phd, Cambridge University, 1995.