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

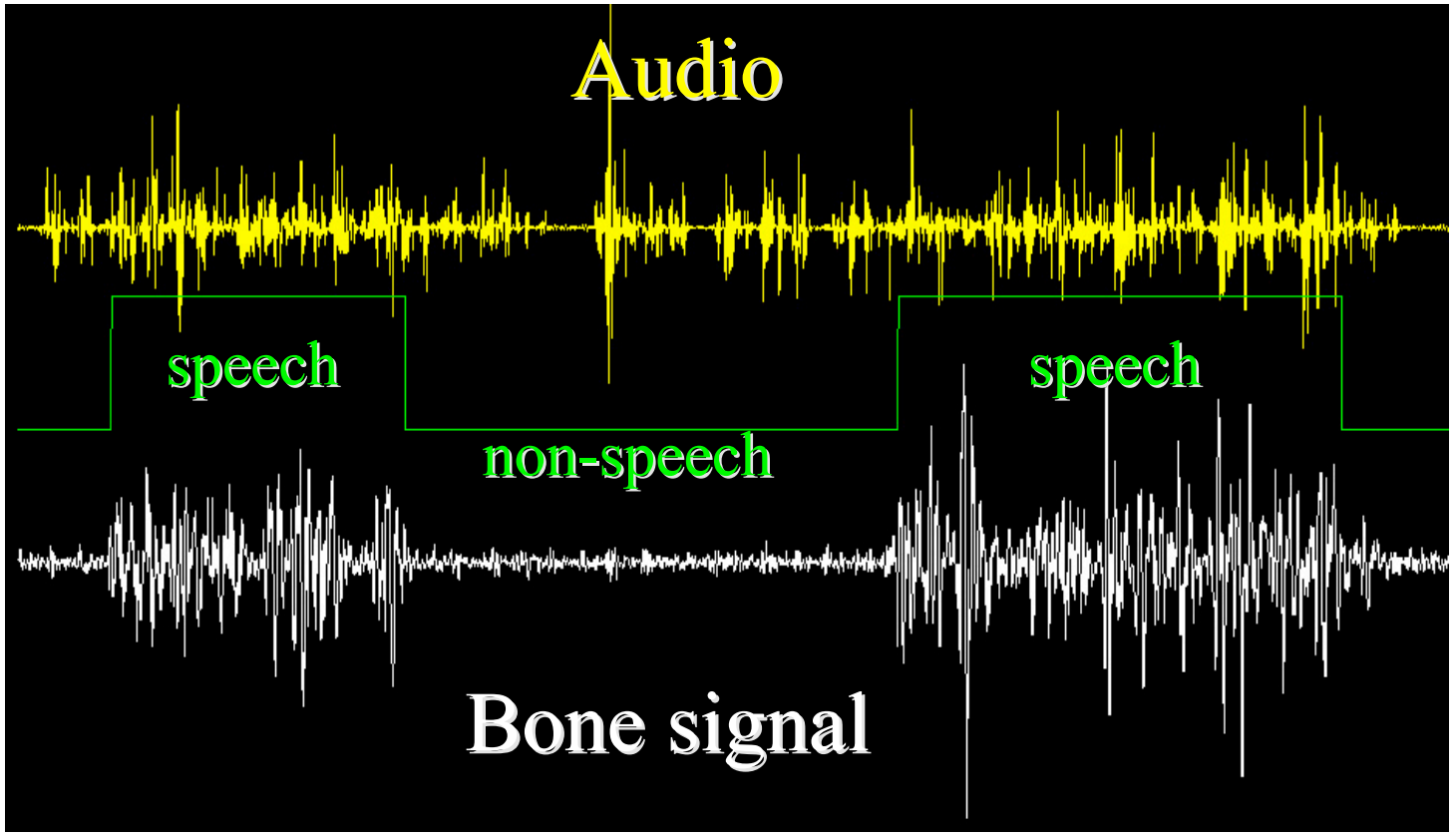
John Hershey, Trausti Kristjansson, Zhengyou Zhang,
Alex Acero: Microsoft Research
With help from Zicheng Liu and Asela Gunawardana

Bone Sensor for Robust Speech Recognition

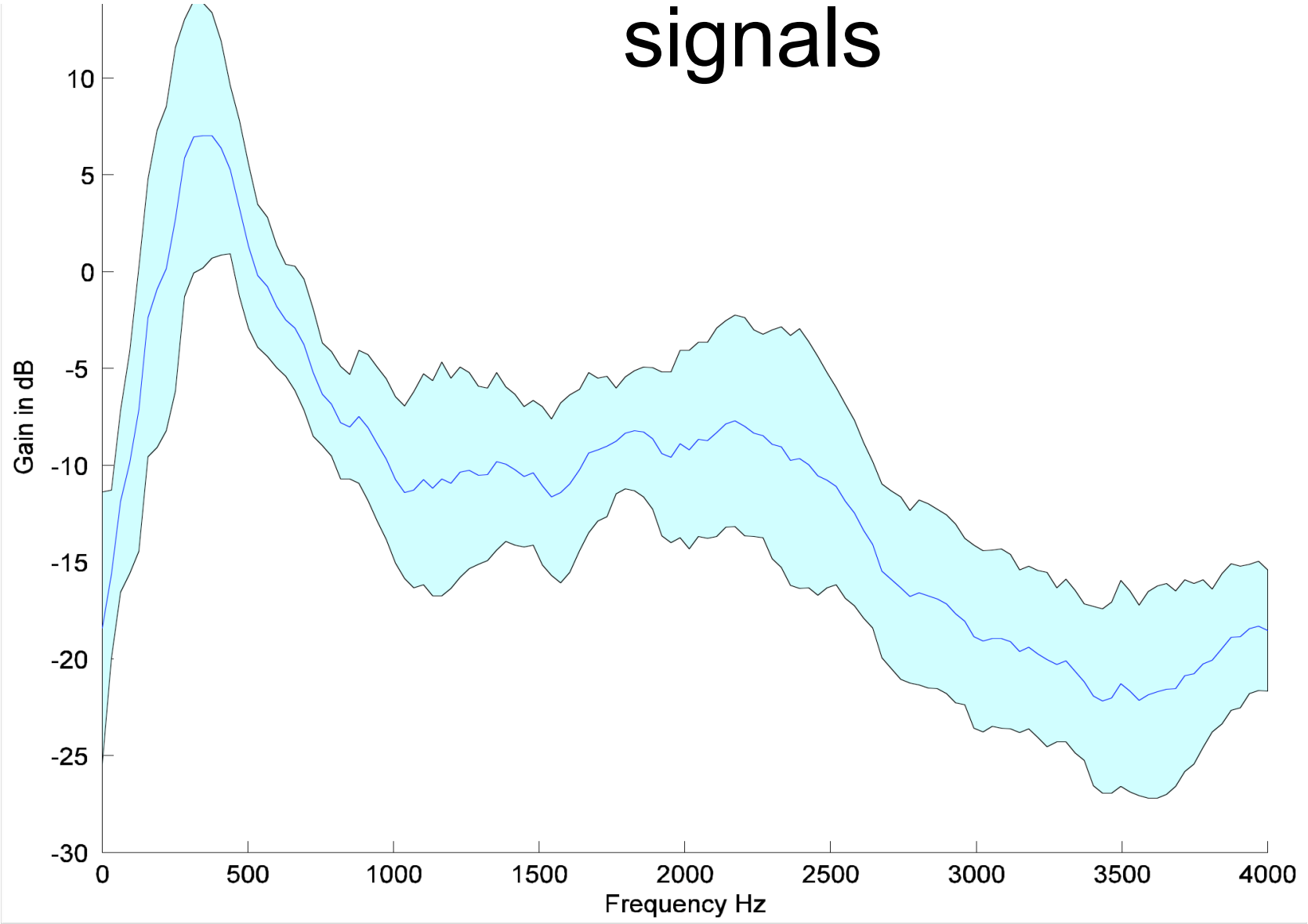
- Small amounts of interfering noise are disasterous for ASR.
- Standard air microphones are susceptible to noise.
- A bone sensor is more resistant to noise, but produces distortion.
- There is not enough data yet to train speech recognition using the bone sensor.
- Enhancement that fuses bone and air sensors allows us to use existing speech recognizers

Sensor Platform

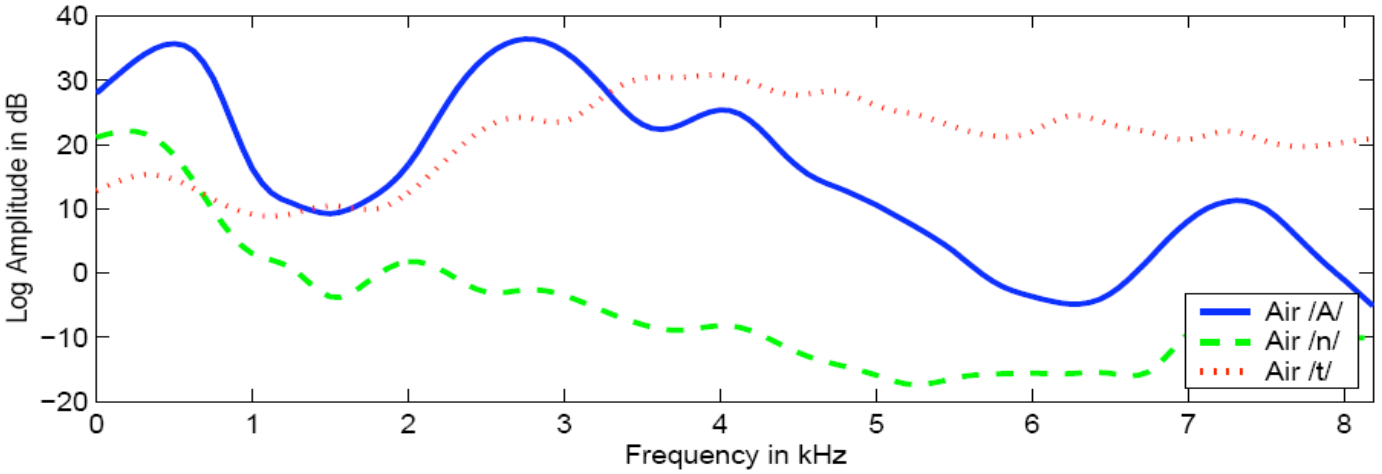
Noise Suppression in Bone Sensor



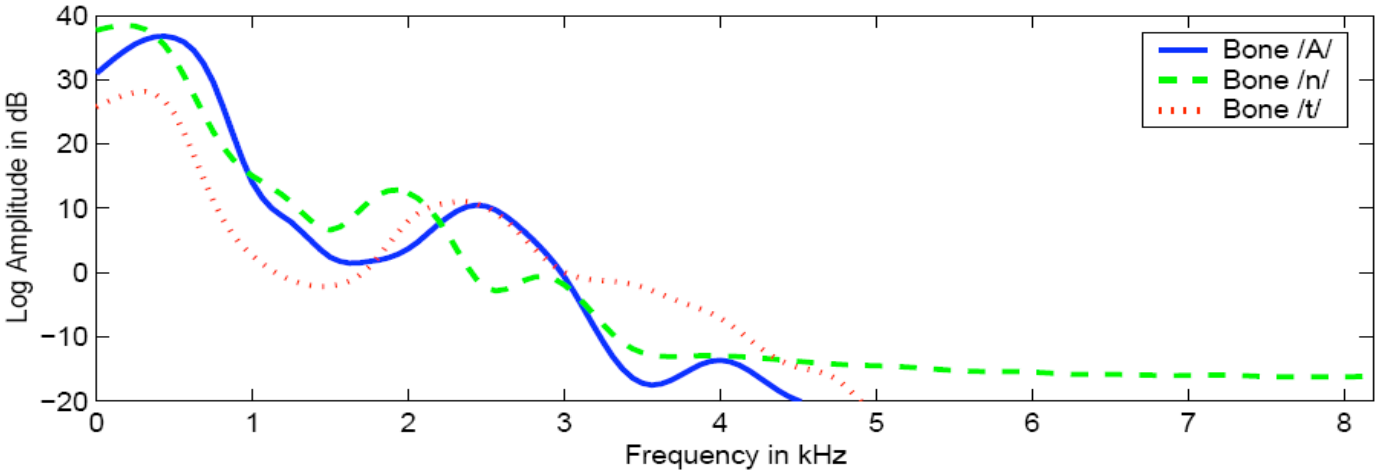
Relationship between air and bone signals



Articulation-Dependent Relationship Between Air and Bone Signals



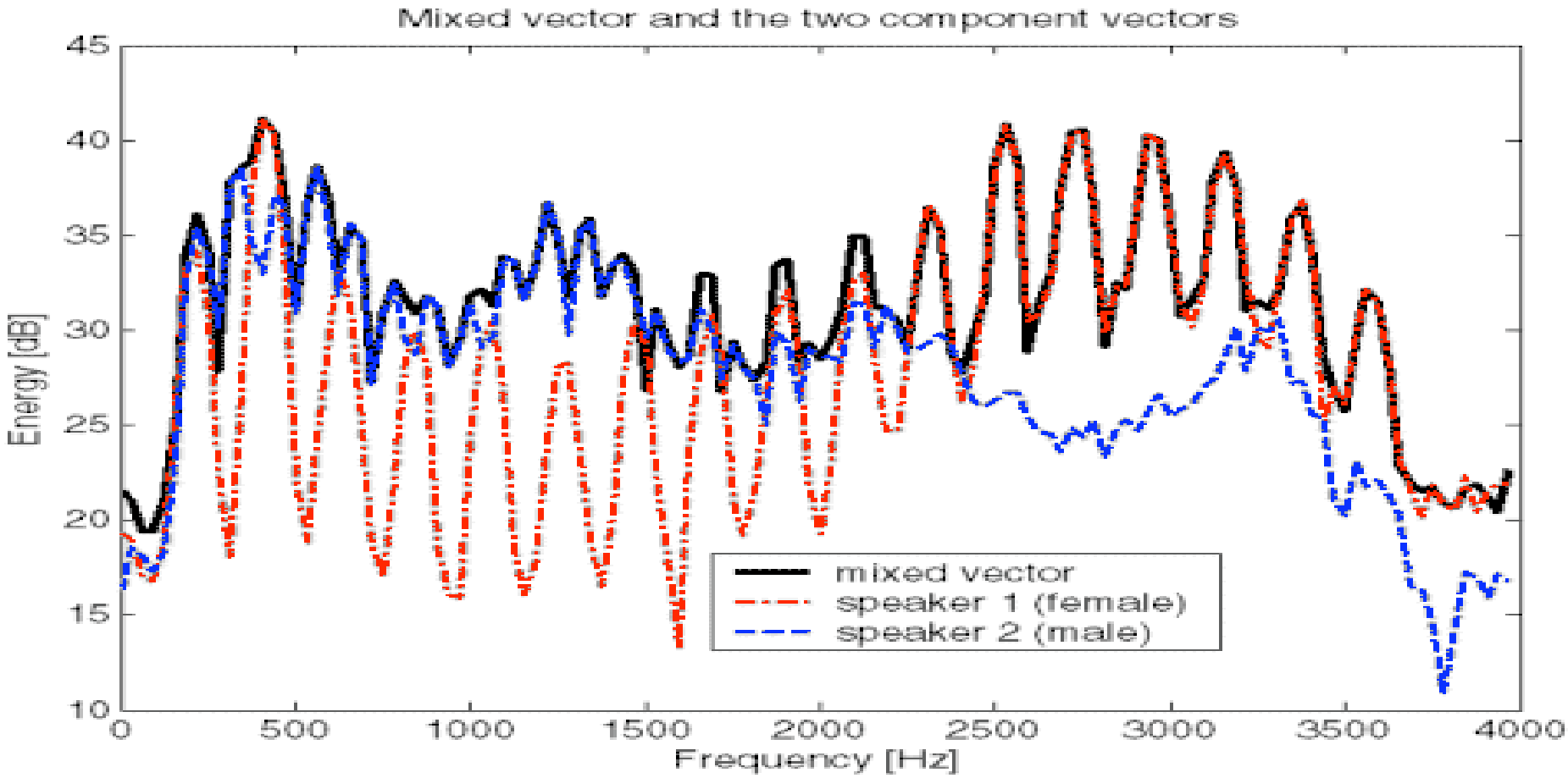
Air



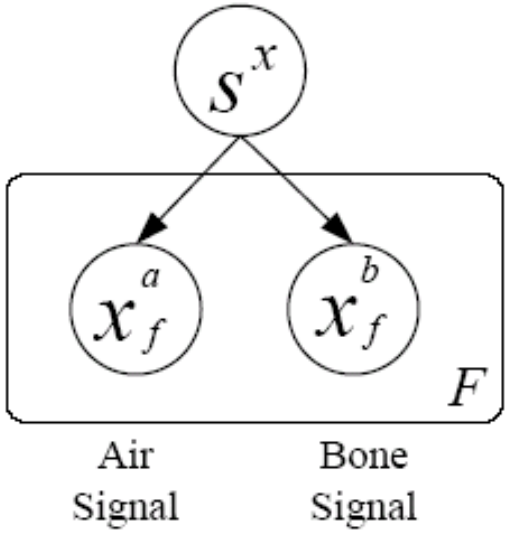
Bone

High Resolution Witty Enhancement

High-resolution log spectrum takes advantage of harmonics.



Speech Model



Pre-trained speech model:

(hundreds of states)

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

Adaptive noise model:

(handful of states)

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

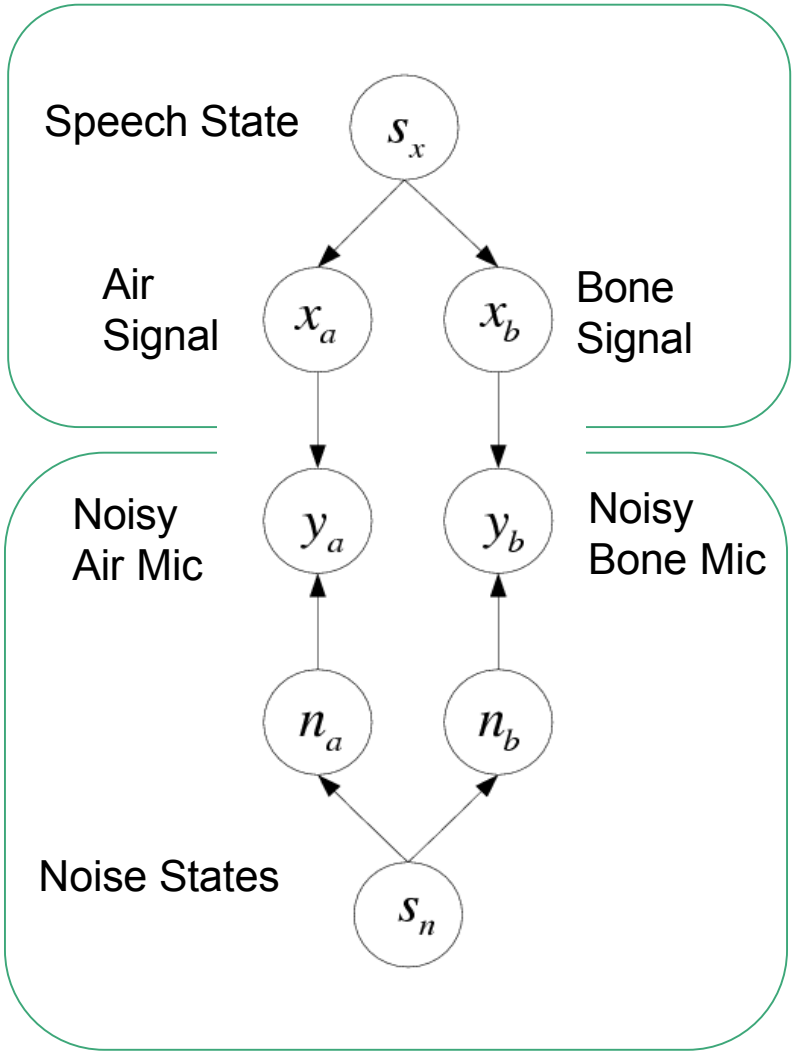
Noisy Speech Model

Speech Model

Trained on small speech database

Noisy Speech Model

Adapted at test time



Sensor Model

Frequency domain combination of signal and noise:

$$Y_f = X_f + N_f$$

Relationship between magnitudes:

$$|Y_f|^2 = |X_f|^2 + |N_f|^2 + 2|X_f||N_f|\cos(\theta)$$

Relationship between log spectra:

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

where $x_f \triangleq \ln|X_f|^2$

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

is treated as probabilistic error

Model distribution over sensors is thus:

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

Inference

Speech posterior:

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

This is intractable due to nonlinearity. Algonquin: Iterate Laplace method. Gaussian estimate of posterior of speech and noise $p(x,n|y,s)$ for each combination of states,

with mean and variance $\eta_s \triangleq \begin{bmatrix} \eta^x_{s^x s^n} \\ \eta^n_{s^n s^n} \end{bmatrix}$ $\Phi_s \triangleq \begin{bmatrix} \Phi^{xx}_{s^x s^n} & \Phi^{xn}_{s^x s^n} \\ \Phi^{nx}_{s^x s^n} & \Phi^{nn}_{s^x s^n} \end{bmatrix}$

and state posterior, $\gamma_s \triangleq \gamma_{s^x s^n}$ (see paper for details)

Compute posterior speech mean: $\hat{x^a} = \frac{\sum_s \gamma_s \eta^a_s}{\sum_s \gamma_s}$

Then resynthesize using lapped transform and noisy phases.

Adaptation

Since noise is unpredictable, we adapt the noise model at inference time, keeping the speech model constant.

The generalized EM algorithm yields the following update equations (see paper for details). The averages over time $\langle\rangle_t$ can be done either in batch or online.

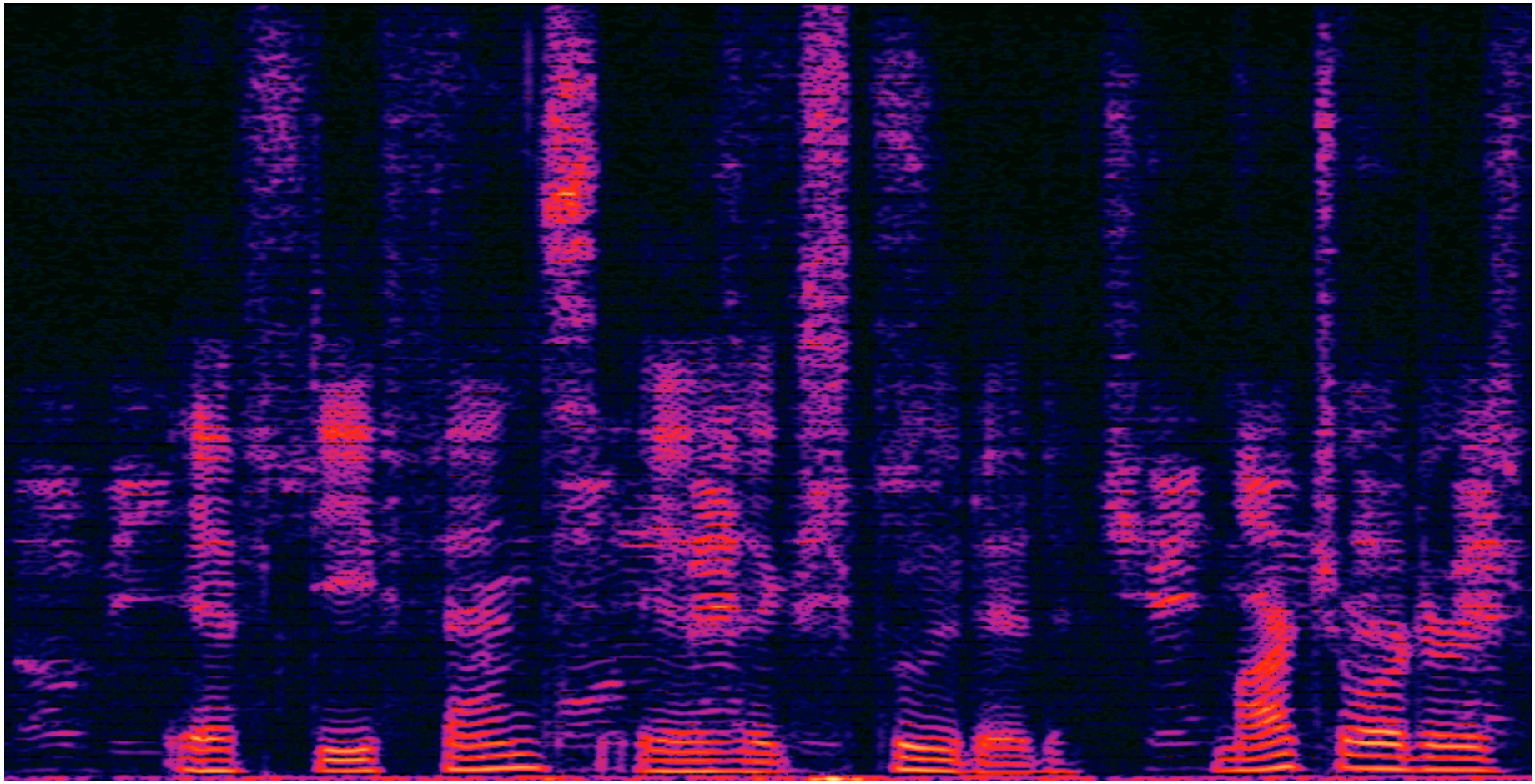
$$\widehat{\pi_{s^n}} \leftarrow \left\langle \sum_{s^x} \gamma_{s^x s^n t} \right\rangle_t$$

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

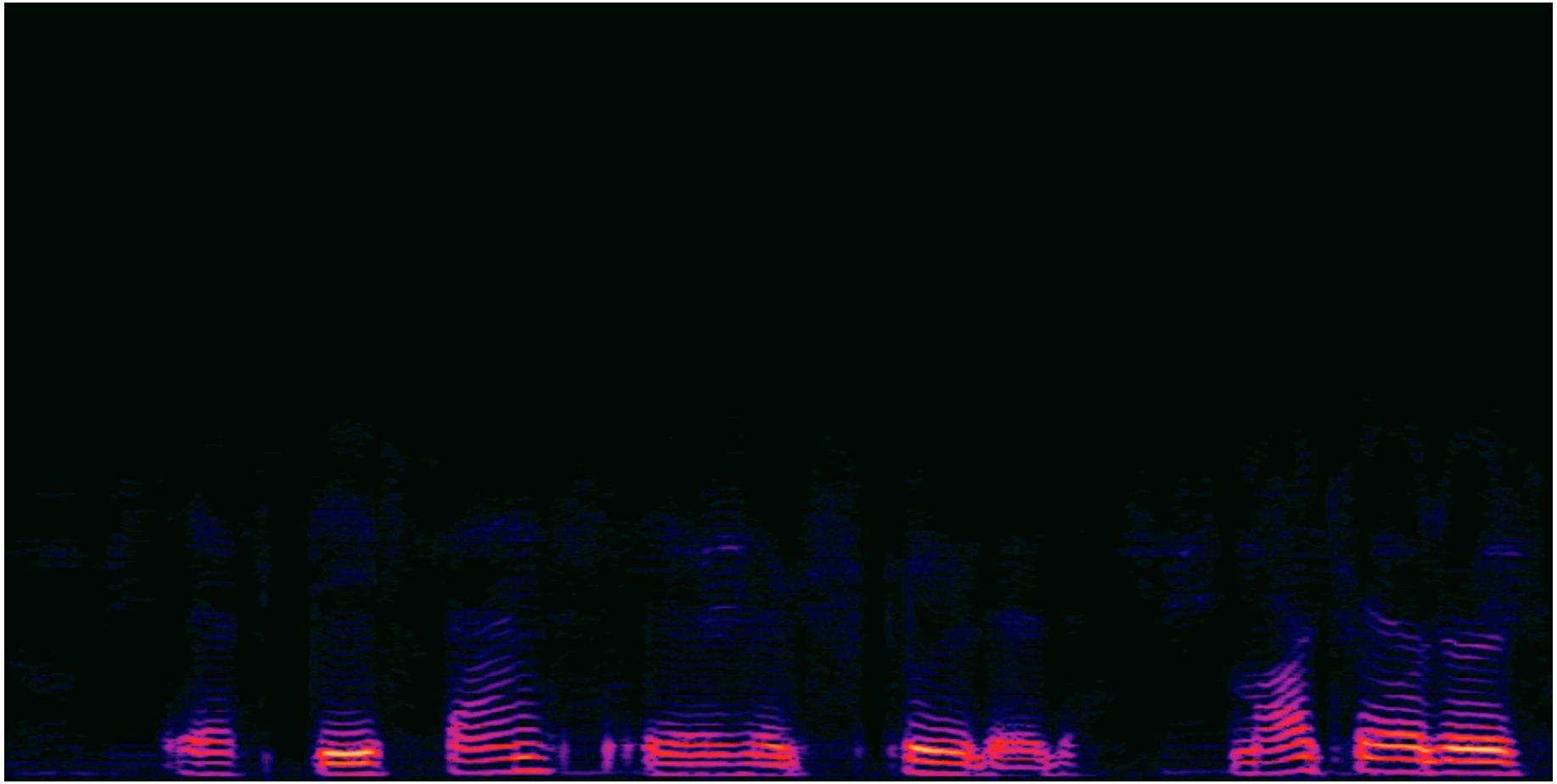
$$\widehat{\sigma^n_s} \leftarrow \left\langle \sum_{s^x} \frac{\gamma_{s^x s^n t}}{\widehat{\pi_{s^n}}} \mathrm{diag}\left[\Phi^{nn}_{s^x s^n} \quad (\eta^n_{s^x s^n t} - \mu_{s^n})(\eta^n_{s^x s^n t} - \mu_{s^n})^T \right] \right\rangle_t$$

Here, $\gamma_{s^x s^n t}$ is the posterior state probability, $\eta^n_{s^x s^n t}$, and $\Phi^{nn}_{s^x s^n t}$ are the posterior mean and covariance of the noise given the state.

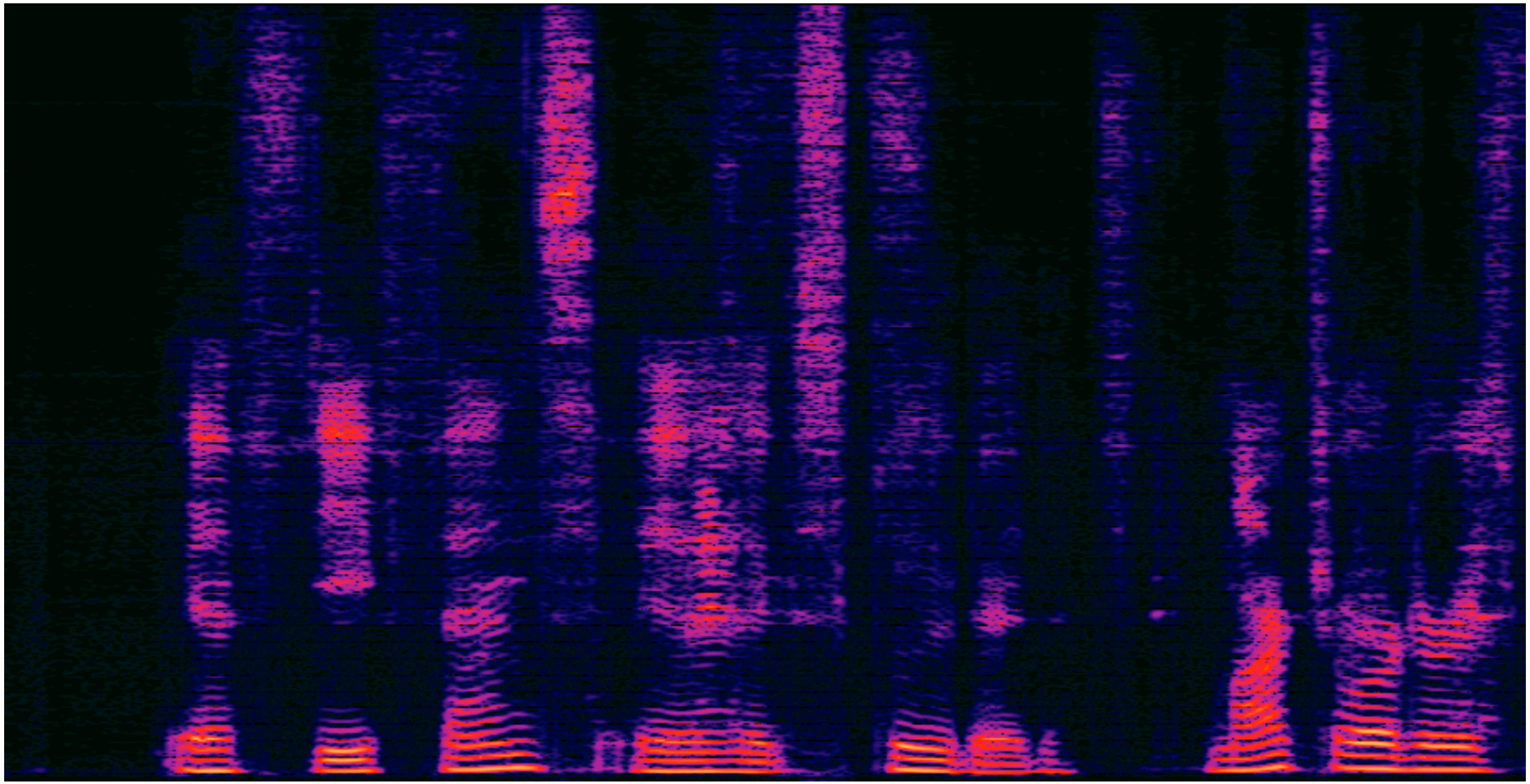
Air Signal



Bone Signal



Enhanced Signal



Evaluation

- Four subjects reading 41 sentences each of the Wall Street Journal.
- Speaker-dependent models of clean speech training set
- Test set: Same sentences read in same environment with interfering male speaker
- Recognizer: 500K vocabulary (Microsoft)

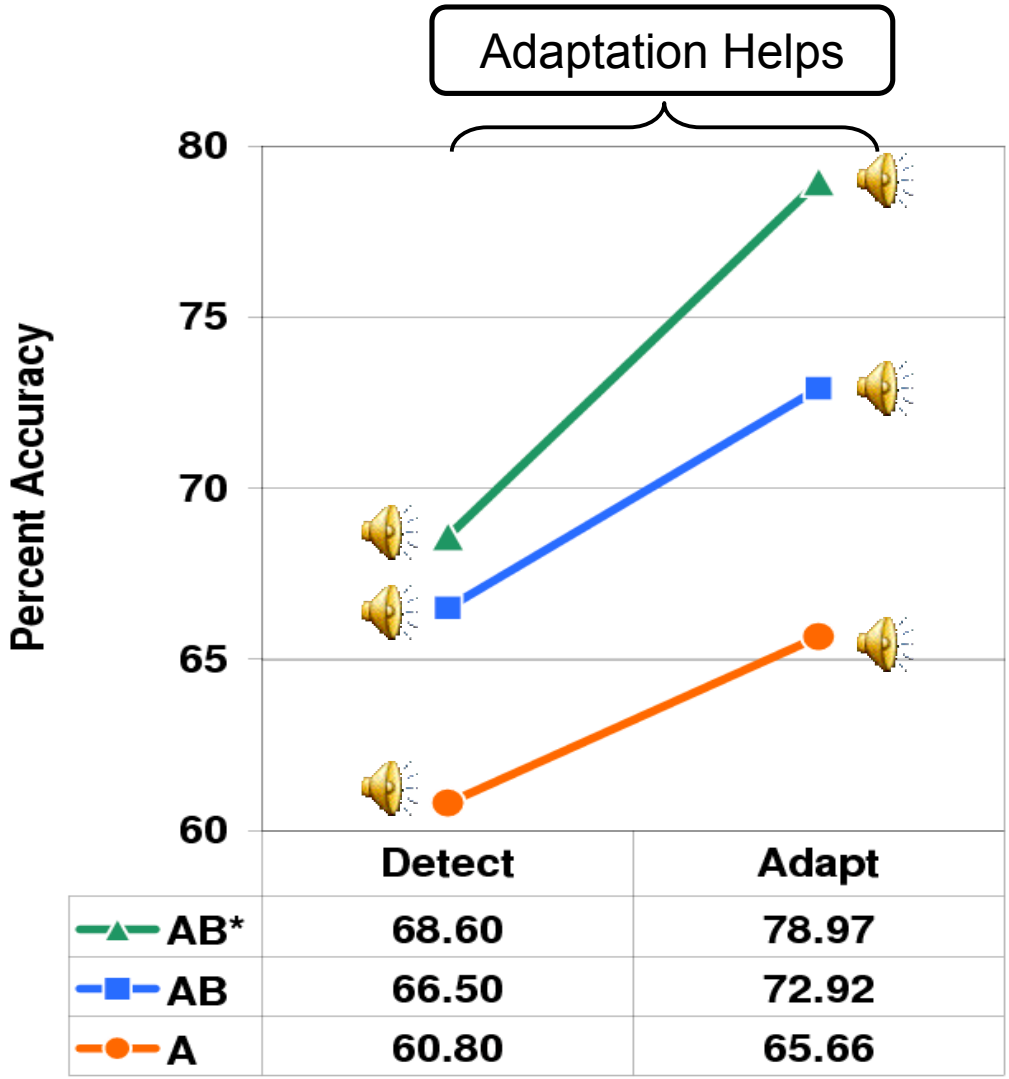
Parameters

- Audio was 16-bit, 16kHz
- Frames were 50ms with 20ms frame shift.
- Frequency resolution was 256 bins.
- Speech models had 512 states
- Noise models had 2 states
- Noisy log spectra were temporally smoothed prior to processing to reduce jitter.

Test Cases

- Sensor conditions:
  - Audio only speech model (***A***)
  - Audio plus Bone speech model (***AB***)
  - AB with state posteriors inferred from bone signal only (***AB+***)
- Adaptation conditions
  - Use speech detection on bone signal to train noise model (***Detect***)
  - Use EM algorithm to adapt noise model (***Adapt***)

Results



Inferring states from bone mic only

Inferring states from both bone and air mics

Inference only using air mic

Original Noisy Signals

Air 57.2%

Bone 28.4%

Clean: 92.7%

Future Directions

- We have shown a strong potential for model-based noise adaptation in concert with a bone sensor.
- Improve the speed of such systems by employing sequential approximations.
- Incorporate state-conditional correlations between air and bone signals.
- Deal with phase dependencies.