B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Speech Recognition”, NIPS 2001

Everything you wanted to know about Algonquin but were afraid to ask (AKA The tricks conveniently left out of a NIPS paper)

Ron Weiss

Algonquin?

- Model based speech enhancement
- Denoise speech signal corrupted by non-stationary noise
- Needs a prior speech model trained on clean speech
- Can learn noise model directly from noisy signal with EM!

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

How to log spectra combine?

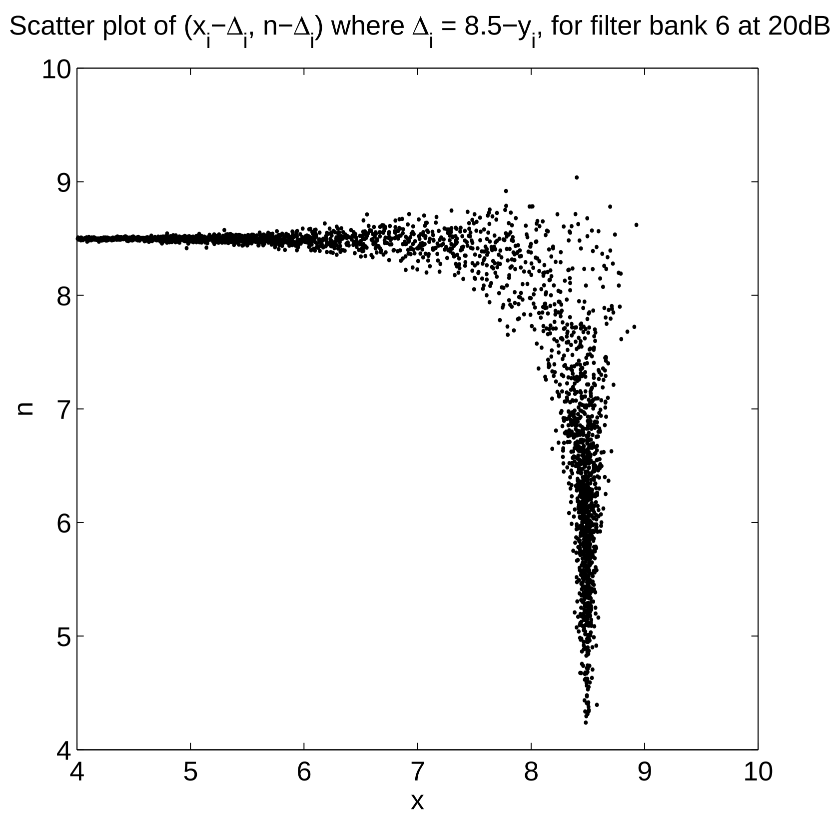
$$\begin{aligned} Y(f) &= S(f) + N(f) \\ |Y(f)|^2 &= |S(f)|^2 + |N(f)|^2 + 2\mathrm{Re}(N(f)^* S(f)) \\ &= |S(f)|^2 + |N(f)|^2 + E \end{aligned}$$

Let $y = \log |Y(f)|^2$

$$\begin{aligned} e^y &= e^s + e^n + \psi = e^s(1 + e^{n-s}) + \psi \\ y &= s + \log(1 + e^{n-s}) + \psi' \end{aligned}$$

B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

The Probability Model



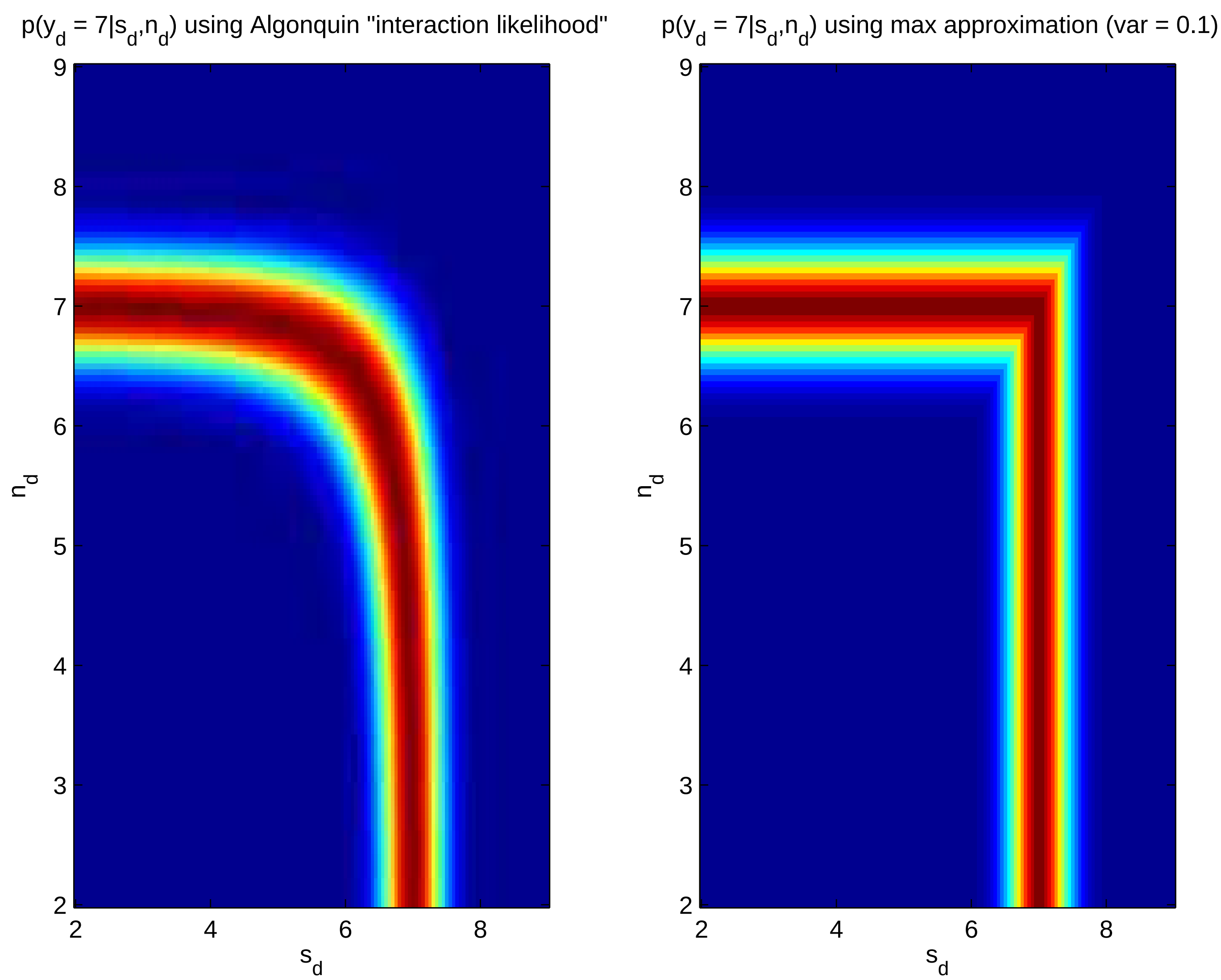
Model error due to phase cancellation as Gaussian noise:

$$p(y|s,n) = \mathcal{N}(y; s + \log(1 + e^{n-s}), \Psi)$$

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

The Probability Model



The rest of the probability model

- Separate GMMs for speech and noise
- assume dimensions are independent (i.e. diagonal covariance)
- no temporal dynamics

$$p(y, s, n, c_s, c_n) = p(y|s, n)p(s|c_s)p(c_s)p(n|c_n)p(c_n)$$

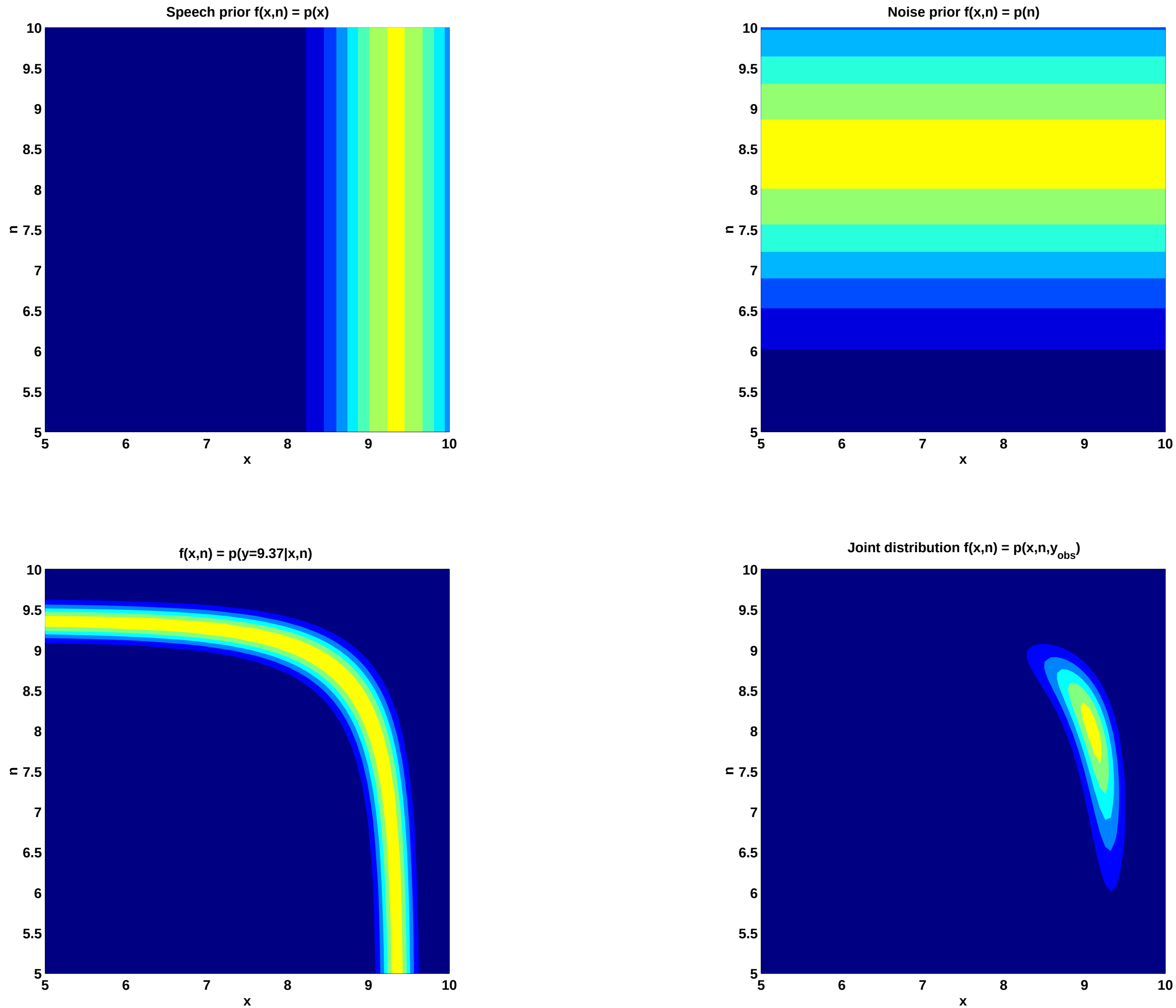
$$p(y|s, n) = \mathcal{N}(y; s + \log(1 + e^{n-s}), \Psi)$$

$$p(s|c_s)p(c_s) = \mathcal{N}(s; \mu^s_{c_s}, \sigma^s_{c_s})\pi_{c_s}$$

$$p(n|c_n)p(c_n) = \mathcal{N}(n; \mu^n_{c_n}, \sigma^n_{c_n})\pi_{c_n}$$

B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

The probability model in pictures



B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

How do we reconstruct the clean signal?

- Our good friend MMSE:

$$\hat{s} = \int s p(s|y) dx$$
$$= \sum_{c_s} p(c_s) \mu^s_{c_s} \quad \text{(if } p(s|y) \text{ is a mixture of Gaussians)}$$

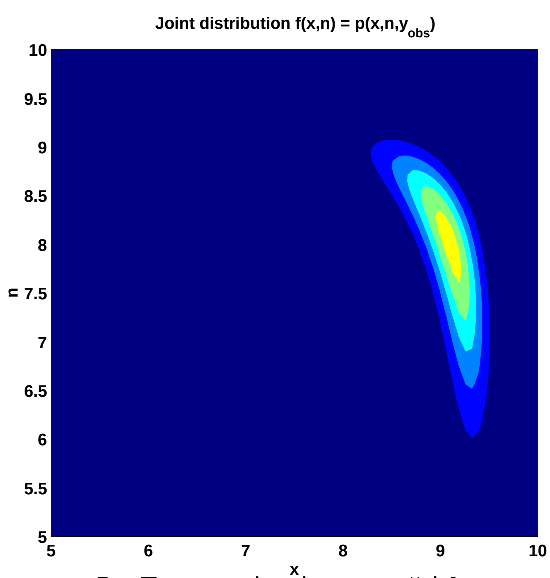
- But the posterior $p(s|y)$, isn't Gaussian due to non-linear likelihood

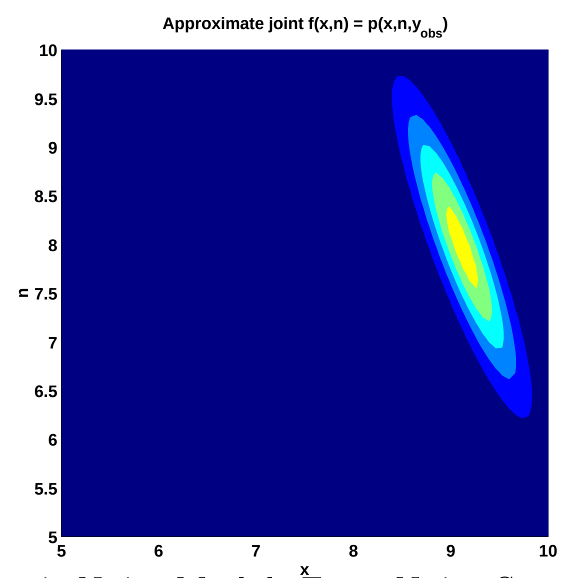
B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

Linearization

- Lets linearize the likelihood using 2nd order vector Taylor series!
  - Review: $f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a)(x-a)^n$
  - only want 2 terms: $f(x) \approx f(a) + f'(a)(x-a)$
- Let $g(s,n) = s + \log(1 + e^{n-s})$, and $z = [s; n]$

$$p(y|n,s) \approx p_l(y|n,s) = \mathcal{N}(y; \mathbf{g}(\mathbf{z_0}) + \mathbf{g}'(\mathbf{z_0})(\mathbf{z} - \mathbf{z_0}), \Psi)$$

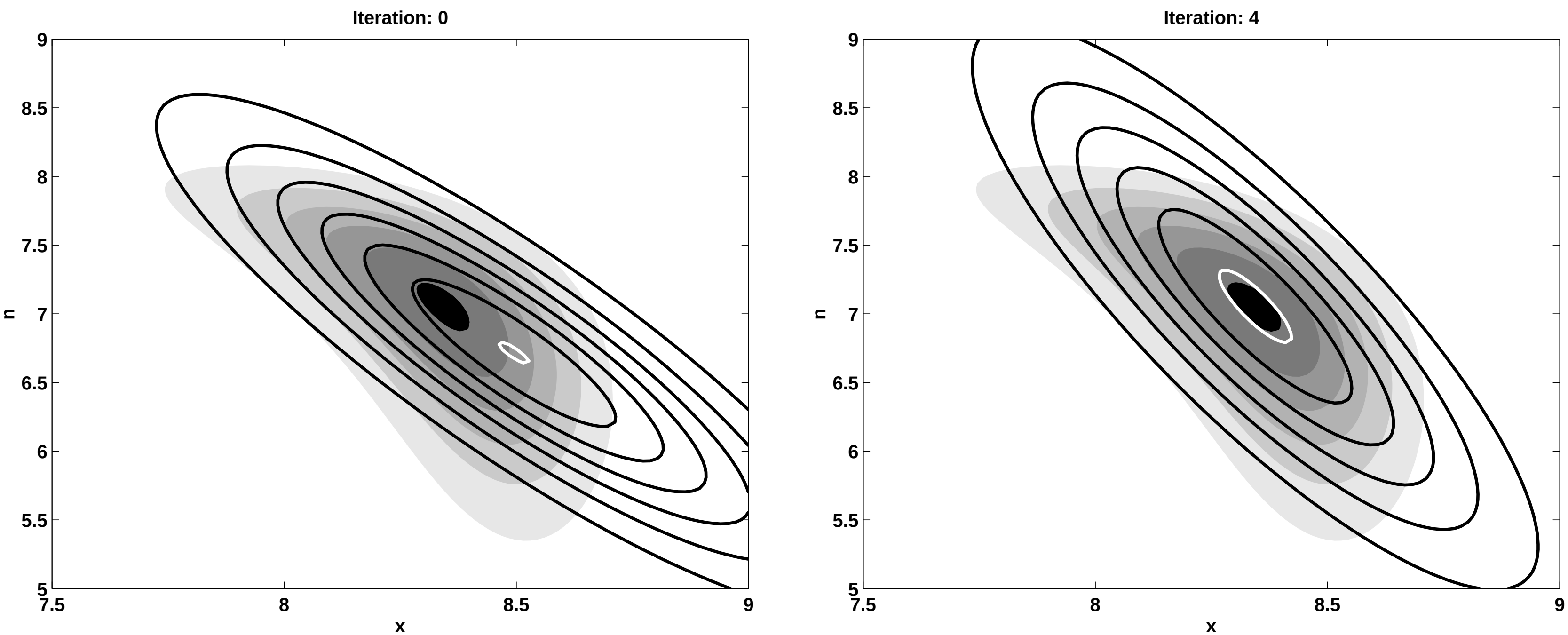




B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

Lets iterate!

Iteratively update $\mathbf{z_0}$



Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

Another Parametrization

- Now we have a Gaussian joint probability, so the posterior is a GMM
- "The form of the joint probability does not allow us to directly read off the mode of the distribution and the marginal"
- "The mode of the posterior $p_l(s, n|y)$ is not coincident with the modes of the priors or the interaction likelihood $(p_l(y|n, s))$"

$$p_l(s, n, c_s, c_n|y) \approx q(s, n, c_s, c_n) = q(s, n|c_s, c_n)q(c_s, c_n)$$

$$q(s, n|c_s, c_n) = \mathcal{N}([s; n]; [\eta_s; \eta_n], \mathbf{\Phi})$$

B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

Variational approximation

- Find the parameters of q in the standard variational way. Minimize the KL divergence (or equivalently maximize $\log p(y)$ - KL divergence) between p_l and q:

$$KL(p||q) = \sum_{c_s} \sum_{c_n} \int_s \int_n q(s, n, c_s, c_n) \log \frac{q(s, n, c_s, c_n}{p(s, n, c_s, c_n|y)}$$

$$\log p(y) - KL(p||q) = C - \sum_{c_s} \sum_{c_n} \int_s \int_n q(s, n, c_s, c_n) \log \frac{p(s, n, c_s, c_n, y}{q(s, n, c_s, c_n)}$$

- Take derivatives with respect to each parameter of q, set to zero and solve to find new parameters in terms of the old ones.
- Can now find MMSE estimate for s: $\hat{s} = \sum_{c_s} p(c_s) \eta_{c_s}^s$

Lab ROSA

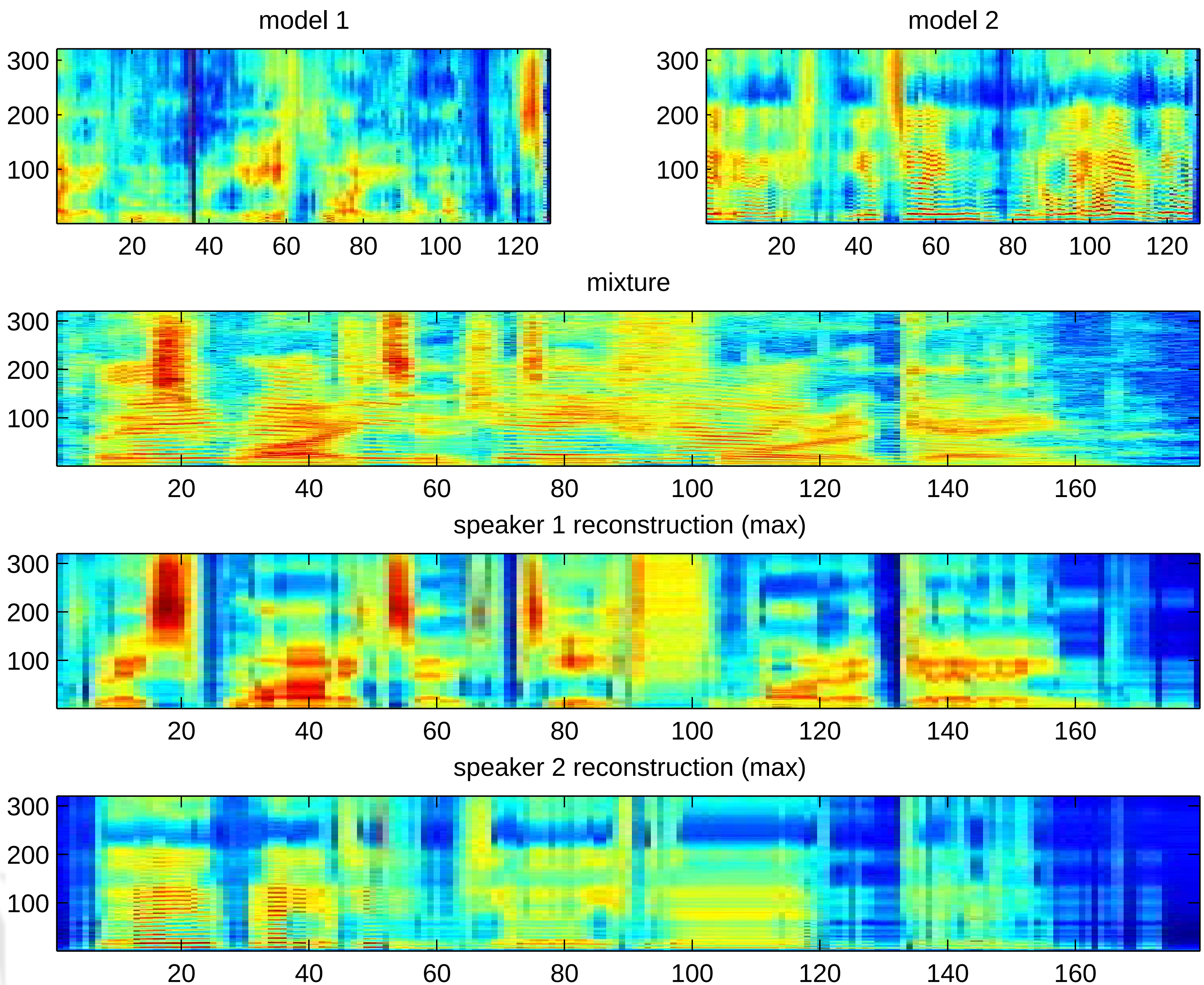
B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

So...

- Algonquin is the greatest thing since sliced bread
- Better approximation to the conditional likelihood of y given $s.n$ than the max approximation
- Its also a good bit slower (at least when you code it in MATLAB)
  - The new parametrization couples the means of the GMMs for s and n
  - η needs to be recalculated at every iteration
- Is it really worth it?

B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

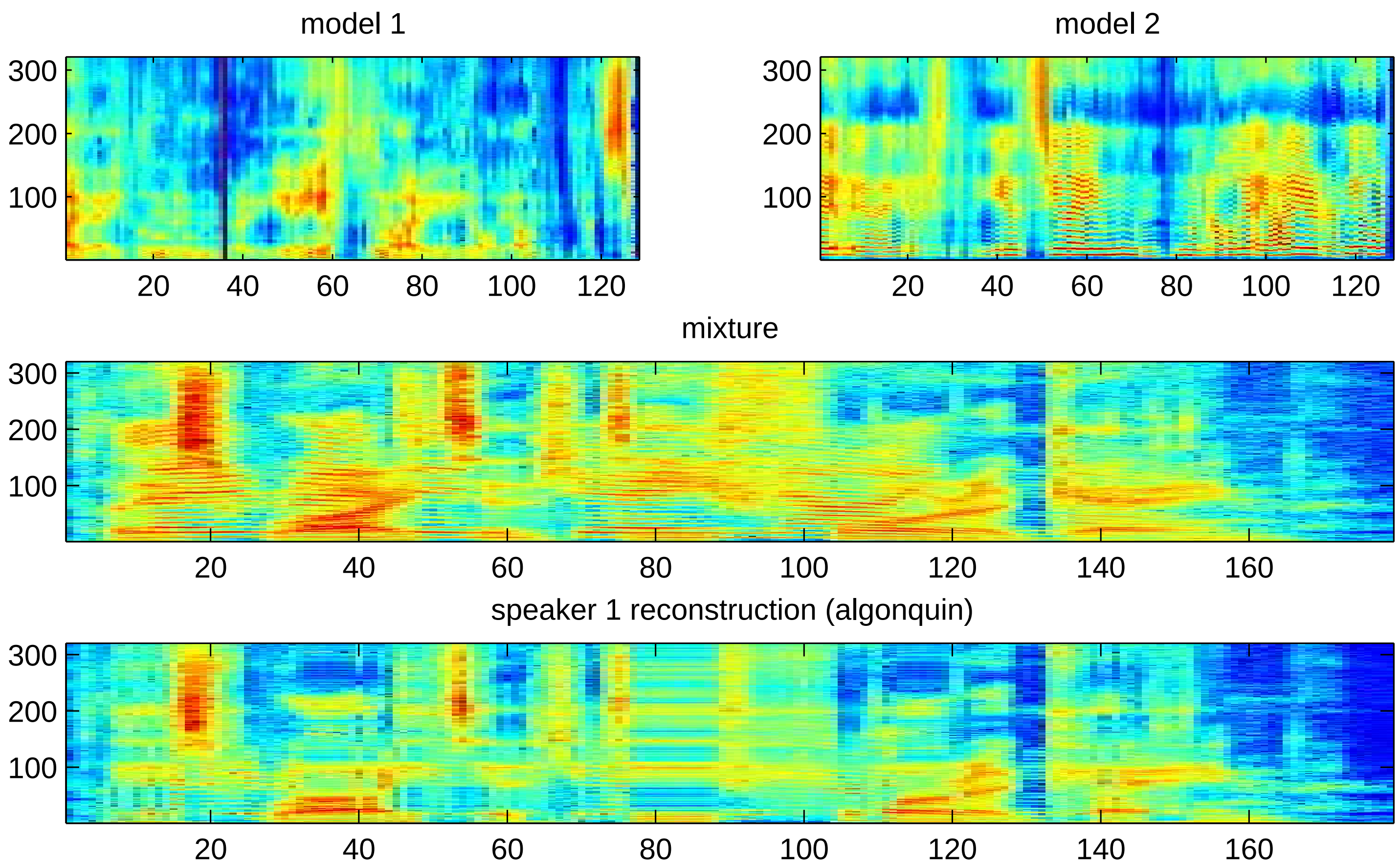
Max separation



Laboratory for the Recognition and
Organization of Speech and Audio

B. Frey, T. Kristjansson, L. Deng, A. Acero, "Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee

Algonquin separation



B. Frey, T. Kristjansson, L. Deng, A. Acero, “Algonquin - Learning Dynamic Noise Models From Noisy Speech For Robust Spee