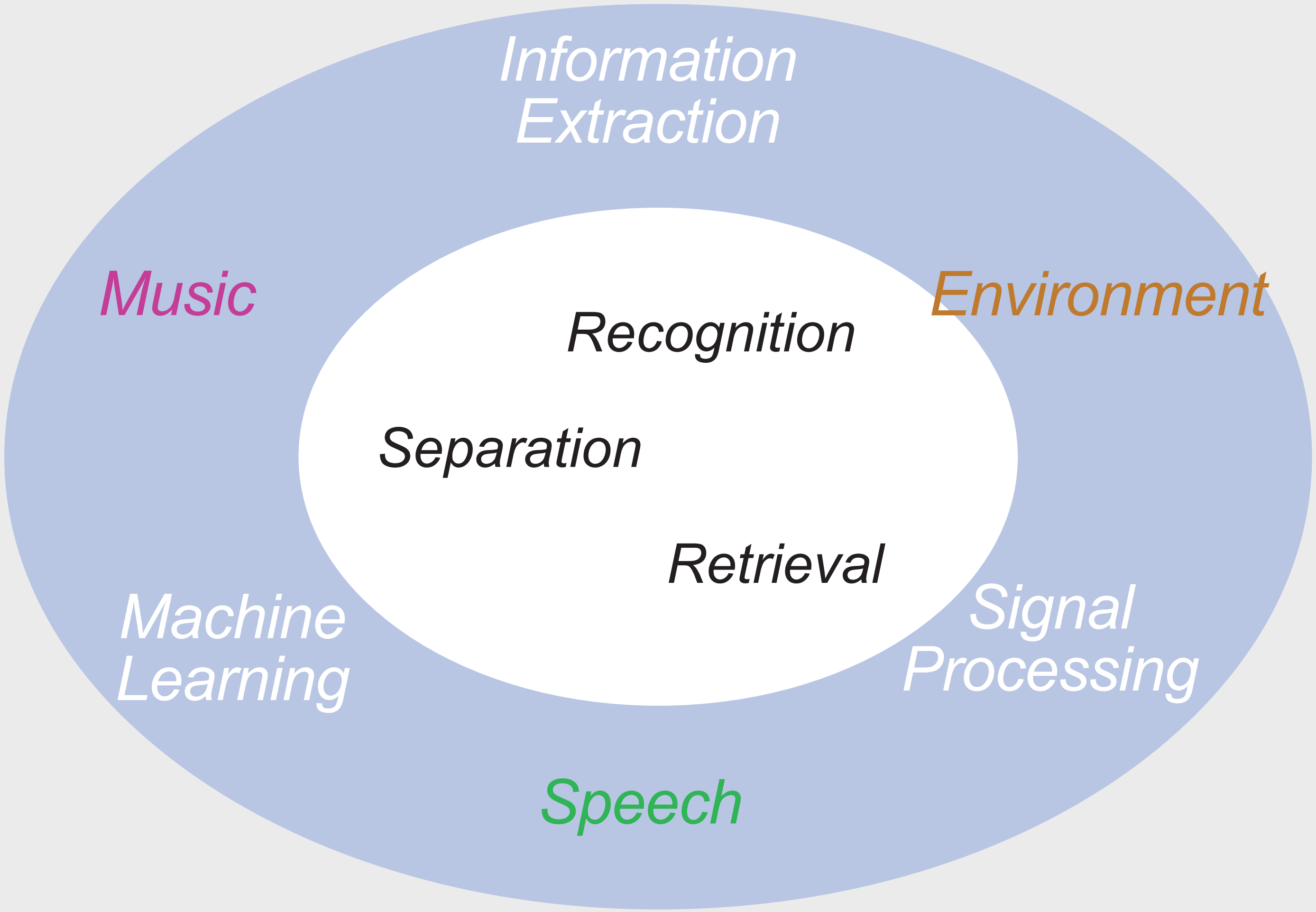
Handling Speech in the Wild

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia University, NY
and
International Computer Science Institute, Berkeley CA

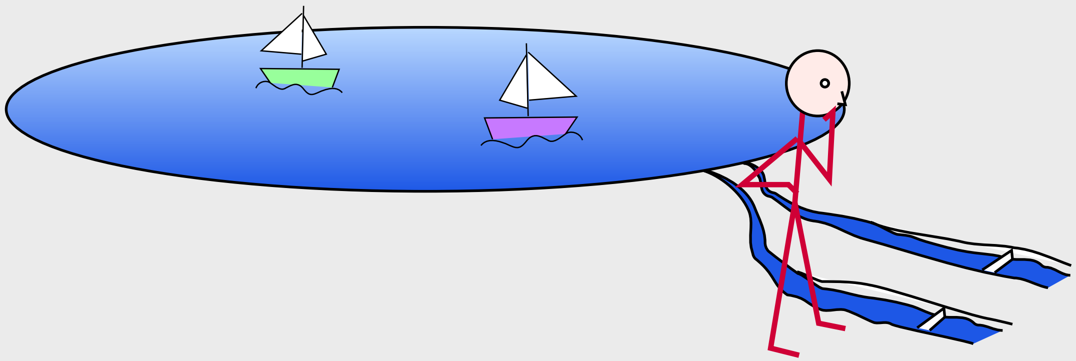
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Speech in the Wild
2. Separation by Space & Pitch
3. Separation by Source Model
4. Inharmonic Speech

LabROSA Overview



1. Speech in the Wild



- The world is cluttered
  sound is transparent
  - mixtures are inevitable

- Useful information is structured by ‘sources’
  - specific definition of a ‘source’:
    intentional independence

Speech in the Wild: Examples

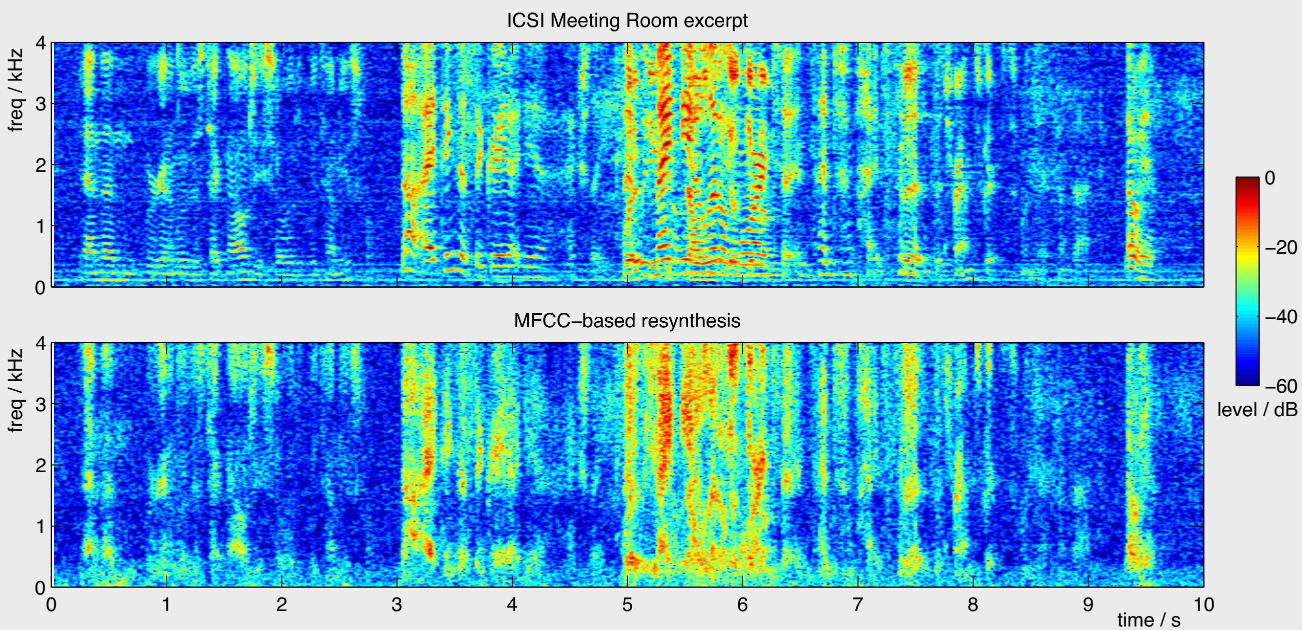
- Multi-party discussions

- Ambient recordings

- Applications:
  - communications
  - robots
  - lifelogging/archives

Recognizing Speech in the Wild

- Current ASR relies on low-D representations
  - e.g. 13 dimensional MFCC features every 10ms

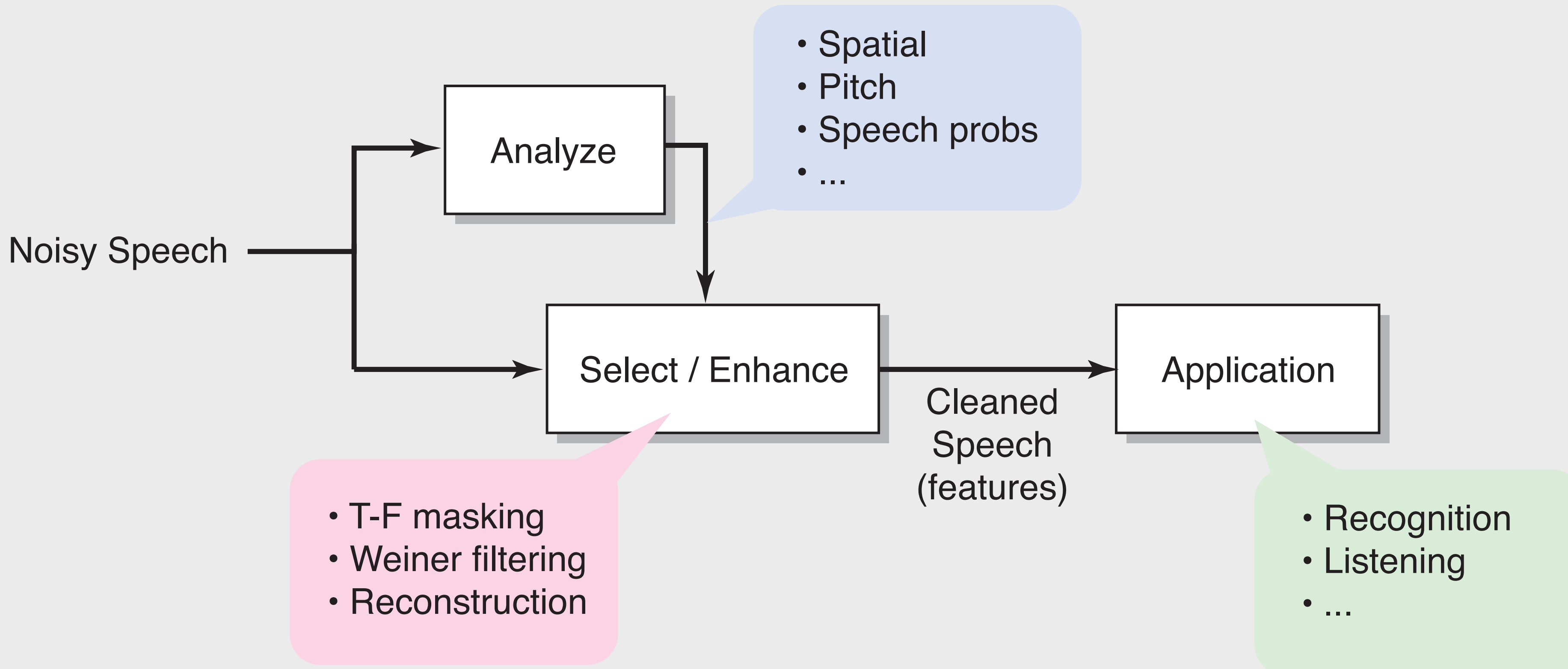


  - very successful for clean speech!
  - inadequate for mixtures

- We need separation!

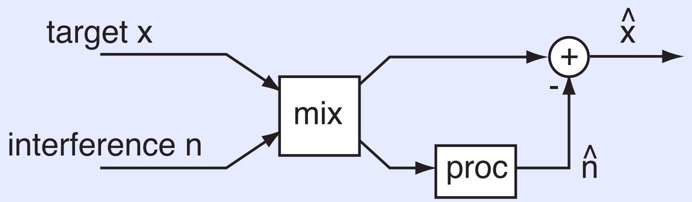
Speech Separation

- How can we separate speech information?

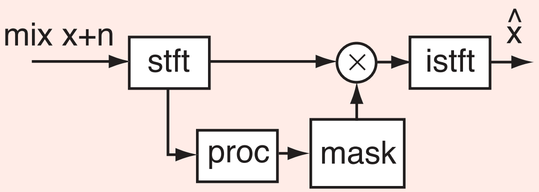


Approaches to Separation

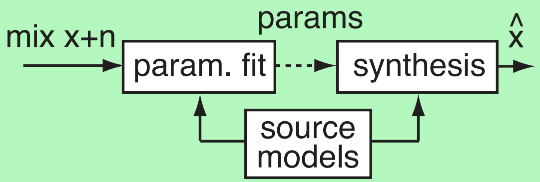
ICA
- Multi-channel
- Fixed filtering
- Perfect separation
  – maybe!



CASA
- Single-channel
- Time-var. filter
- Approximate separation



Model-Based
- Any domain
- Param. search
- Synthetic output?



Separation vs. Inference

- Ideal separation is rarely possible
  - many situations where overlaps cannot be removed
- Overlaps → Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

$$P(\{S_i\}|X) \propto P(X|\{S_i\}) \prod_i P(S_i|M_i)$$

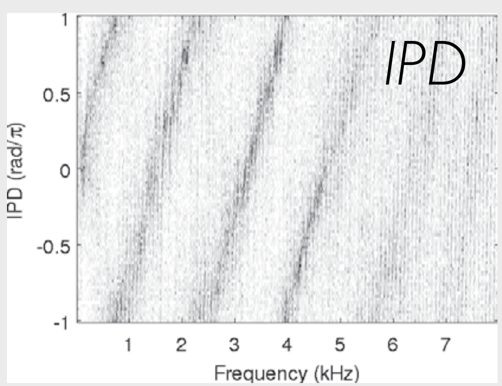
  - search over all source signal sets $\{S_i\}$??
- Better source models M_i → better inference

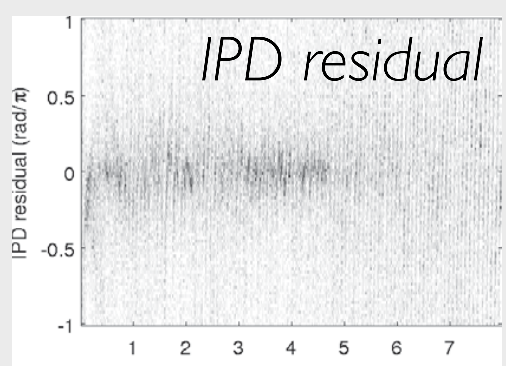
2. Separation by Spatial Info

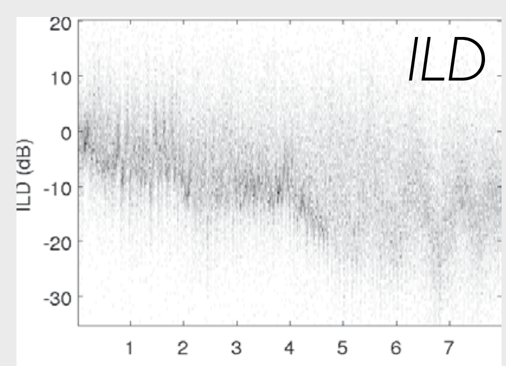
- Given multiple microphones,
  sound carries spatial information about source
- E.g. model interaural spectrum of each source
  as stationary level and time differences:

$$\frac{L(\omega, t)}{R(\omega, t)} = a(\omega) e^{j\omega\tau} N(\omega, t)$$

- e.g. at 75°, in reverb:

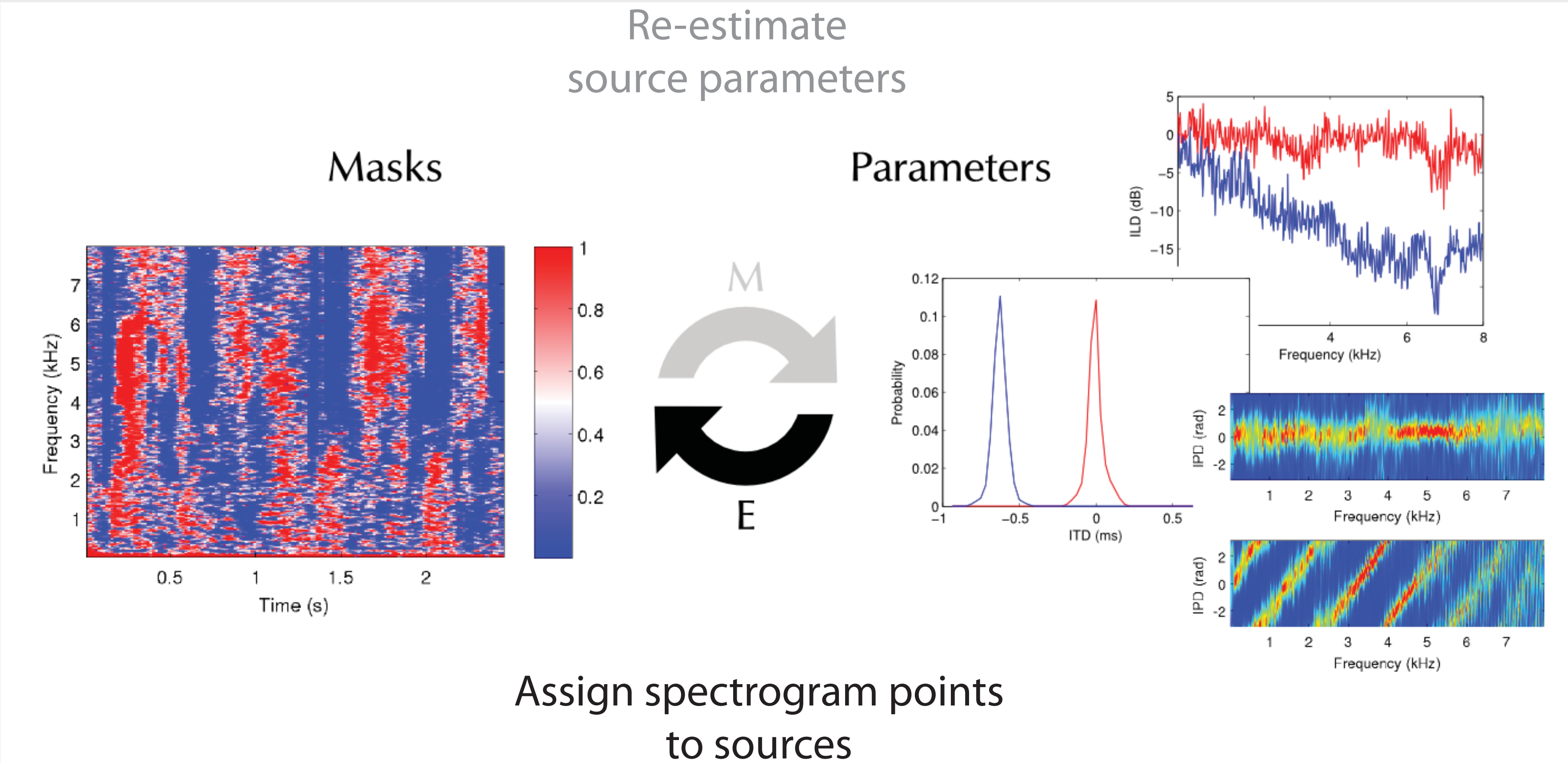






Model-based EM Source Separation and Localization (MESSL)

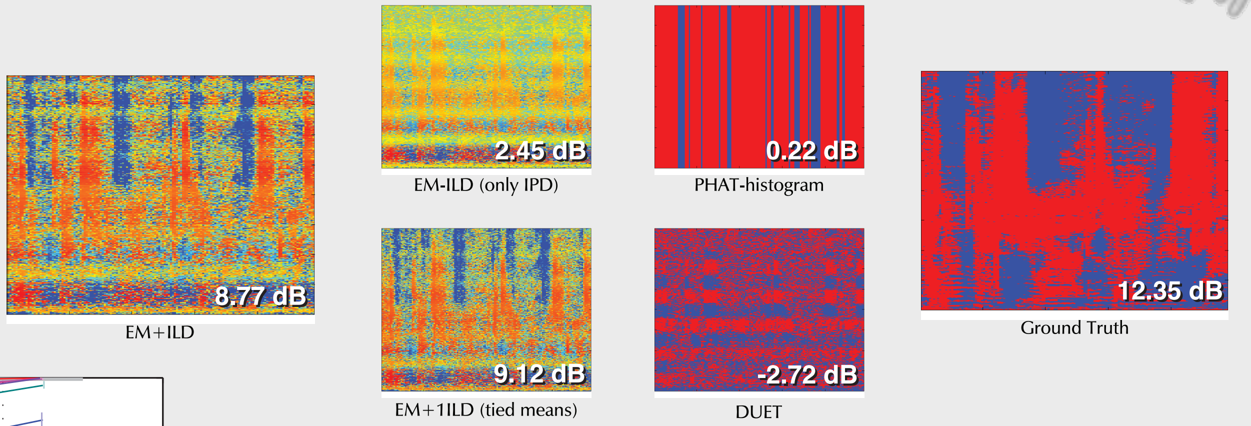
Mandel et al. '10

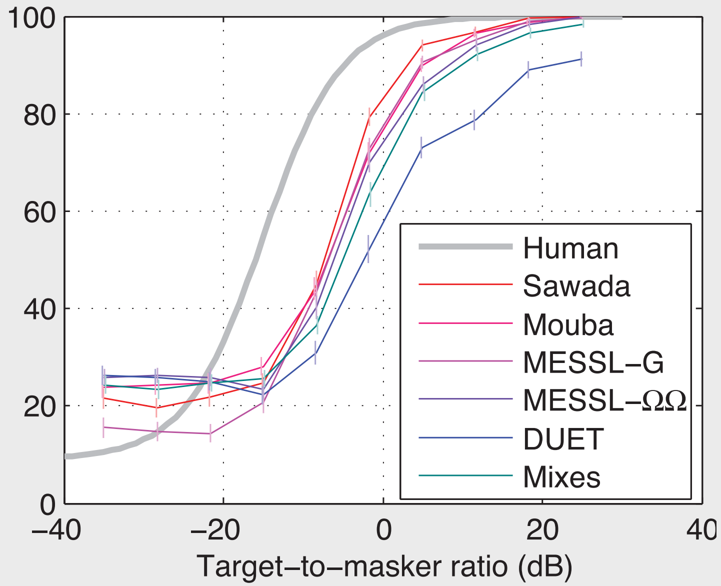


- can model more sources than sensors

MESSL Results

- Modeling uncertainty improves results
  - tradeoff between constraints & noisiness

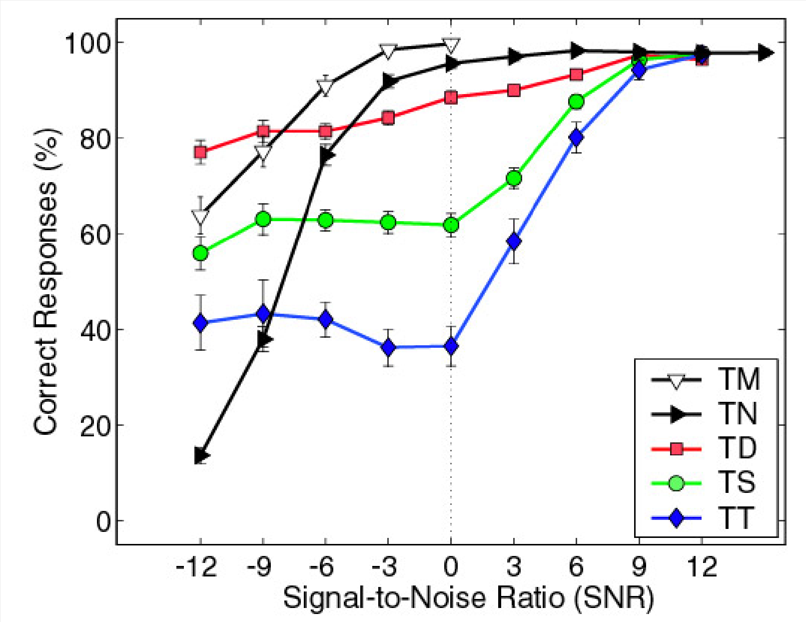




- Helps with recognition
  - digits accuracy

Separation by Pitch

- Voiced syllables have near-periodic "pitch"
  - perceptually salient
  - lost in MFCCs



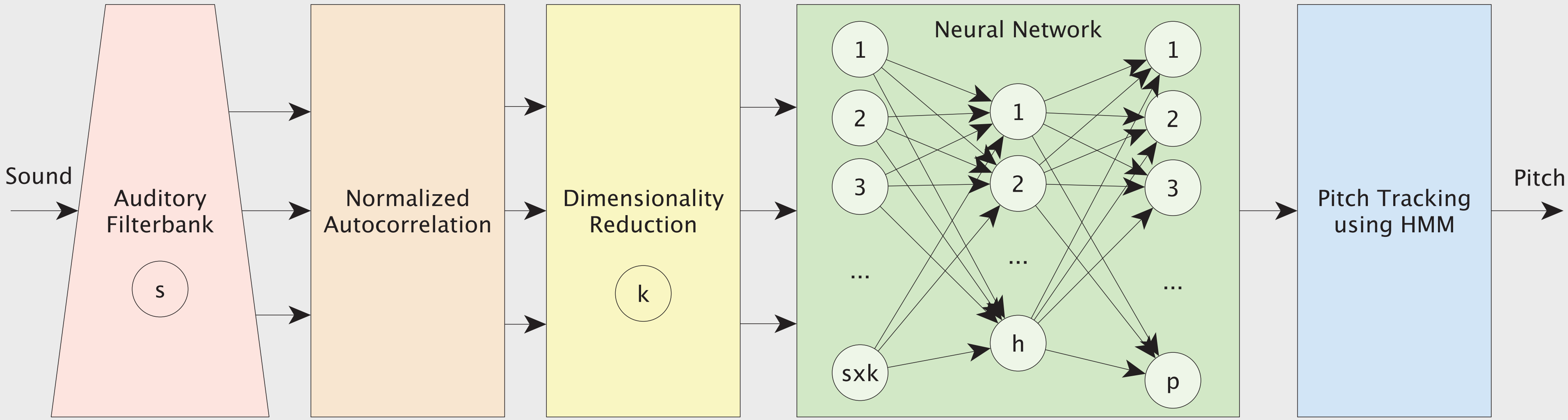
Brungart et al.'01

- Can we track pitch & use it for separation?
  - ... and other speech tasks?

SAcC Pitch Tracking

BS Lee & Ellis '12

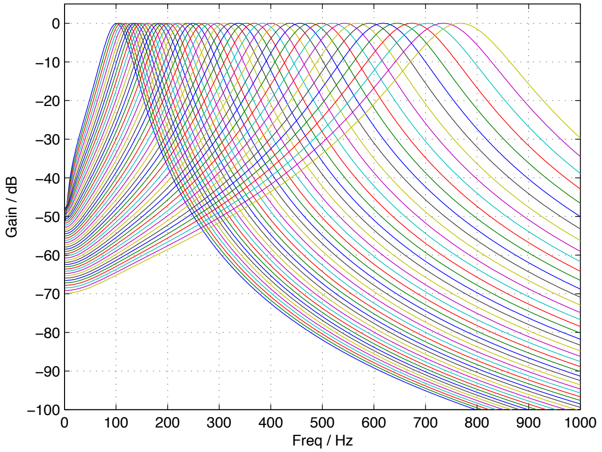
- Based on channel selection *Wu, Wang & Brown '03*
  - pitch from summary autocorrelation finds "good" bands

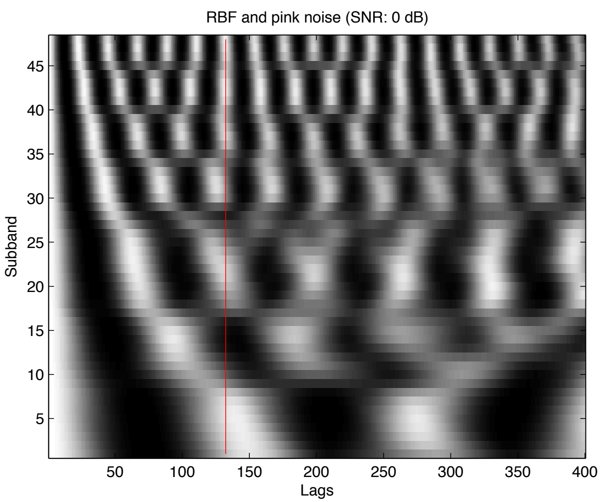


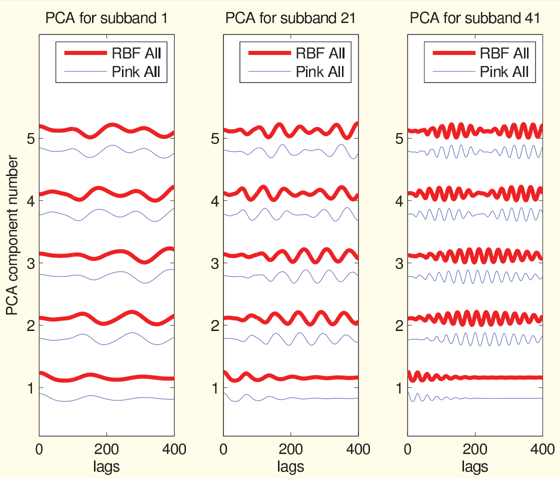
  - trained classifier decides pitch from evidence
  - Subband Autocorrelation Classification = SAcC

Subband Autocorrelation PCA

- Subband Autocorrelation is high-dimensional
  - e.g. 24 subbands x 200 lags
  - each subband's autocorrelation is highly redundant



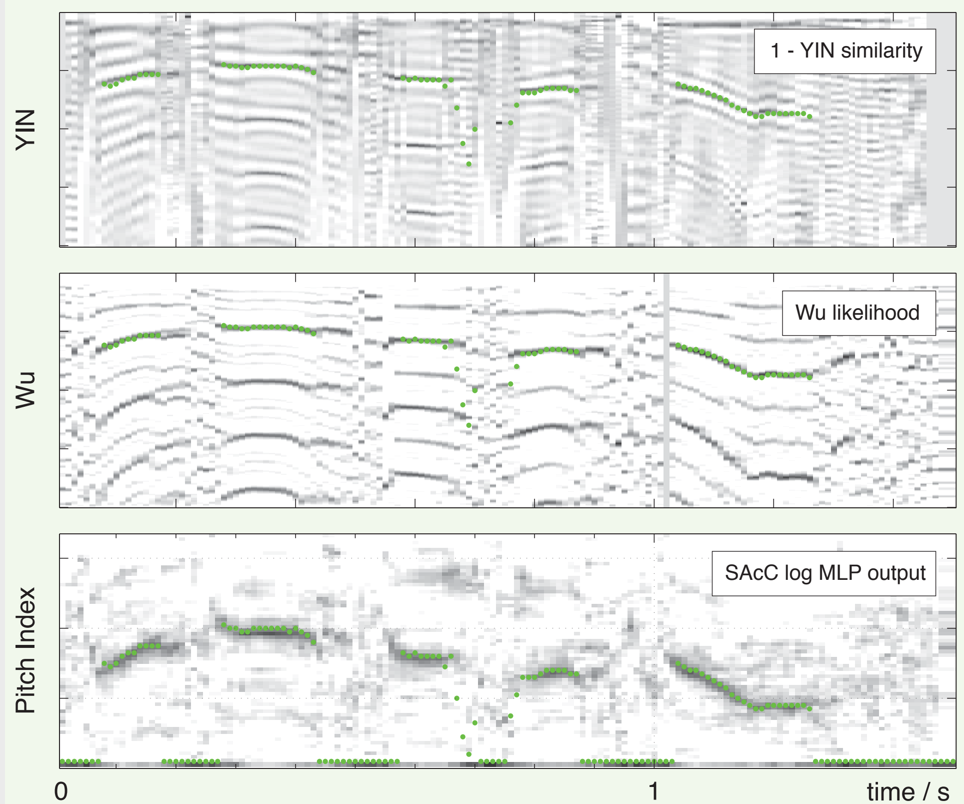




- Represent with PCA
  - 10 bases sufficient
  - bases don't much depend on training data

Trained Pitch Classifier

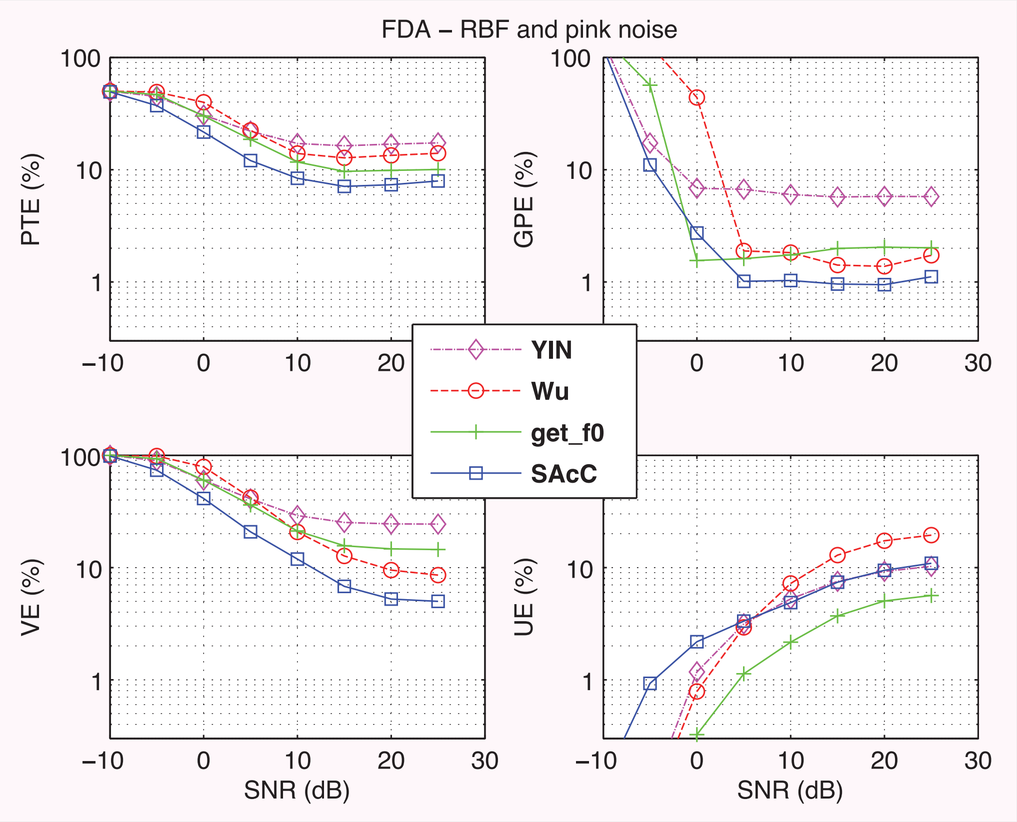
- Core of SAcC is MLP Classifier trained on noisy audio + ground-truth pitch to output one pitch bin per frame
    - discriminates against e.g. octave errors



Pitch “scores” from YIN, Wu, and SAcC

SAcC Results

- SAcC exploits in-domain data to do better than "general purpose" pitch trackers
  - generalization...



3. Separation by Models

Roweis '01, '03
Kristjannson '04, '06

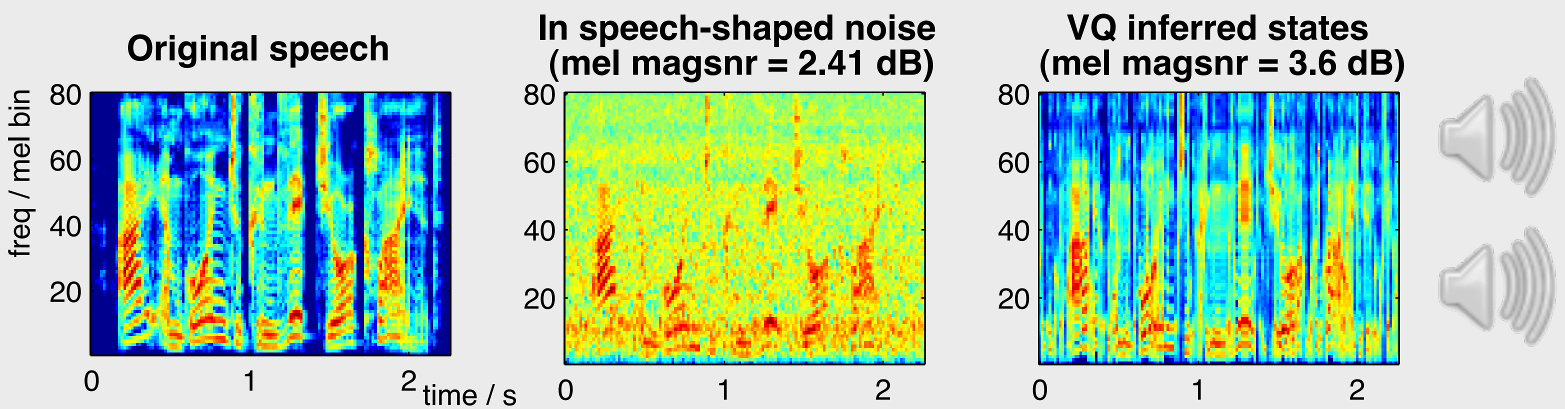
- Given models (codebooks) for sources, find "best" (most likely) states i for spectra:

$$P(\mathbf{x}|i_1, i_2) = \mathcal{N}(\mathbf{x}; \mu_{i_1} + \mu_{i_2}, \Sigma)$$ combination model

$$\{i_1(t), i_2(t)\} = \arg\max_{i_1, i_2} P(\mathbf{x}(t)|i_1, i_2)$$ inference of source state

  - can include sequential constraints...

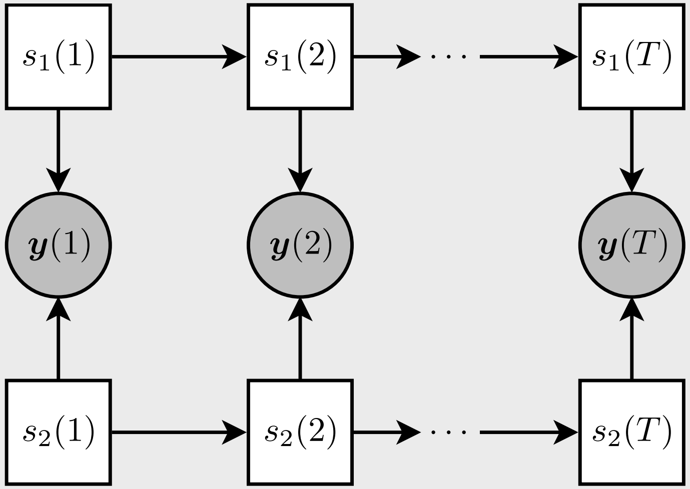
- E.g. stationary noise:

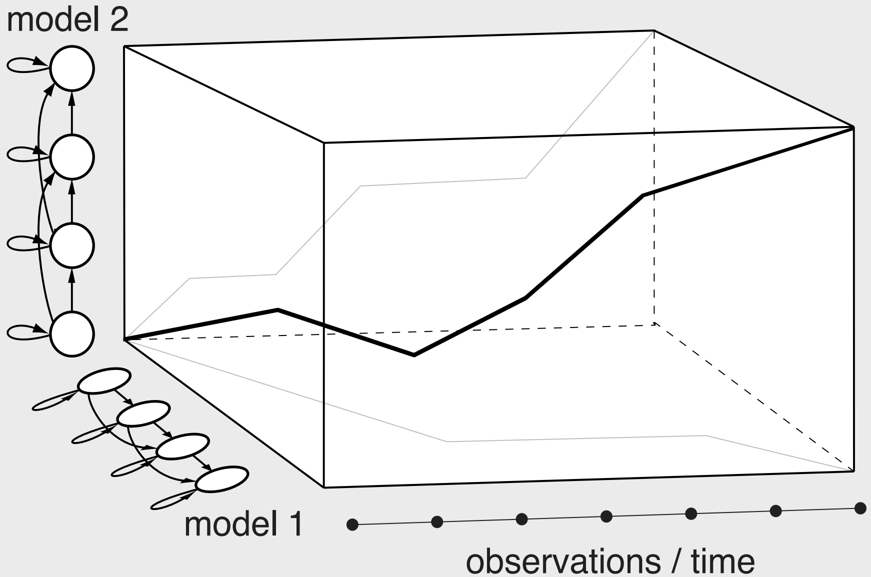


Separation by ASR Models

Varga & Moore, '90
Hershey et al., '10

- If ASR is finding best-fit parameters
  argmax $P(W \mid X)$ ...
- Recognize mixtures with Factorial HMM
  - model + state sequence for each voice/source
  - exploit sequence constraints, speaker differences



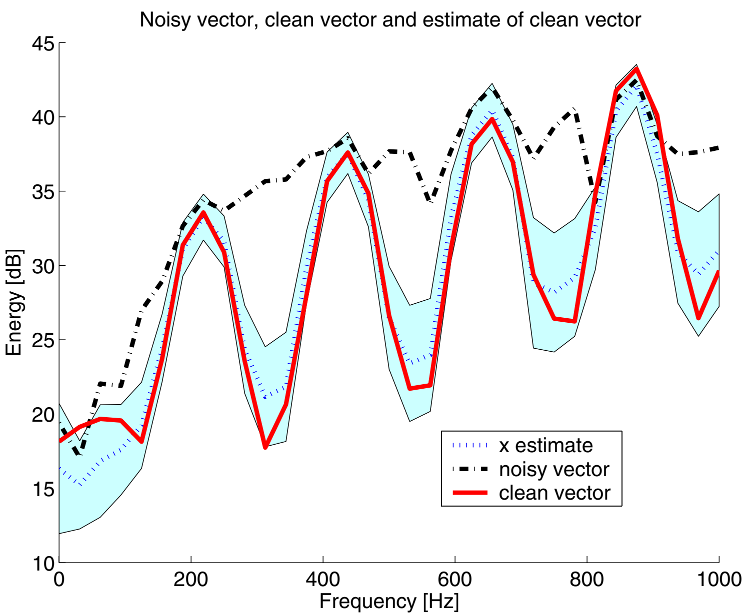


  - separation relies on detailed speaker model

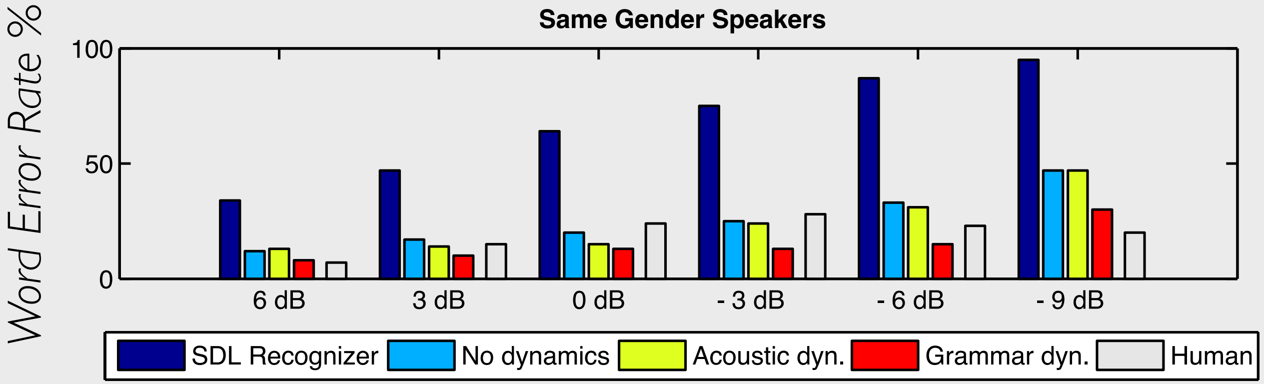
IBM "Superhuman" System

Kristjansson, Hershey et al. '06, '10

- **Iroquois speech separation system features:**
  - detailed state combinations
  - large speech recognizer
  - exploits grammar constraints
  - 34 per-speaker models



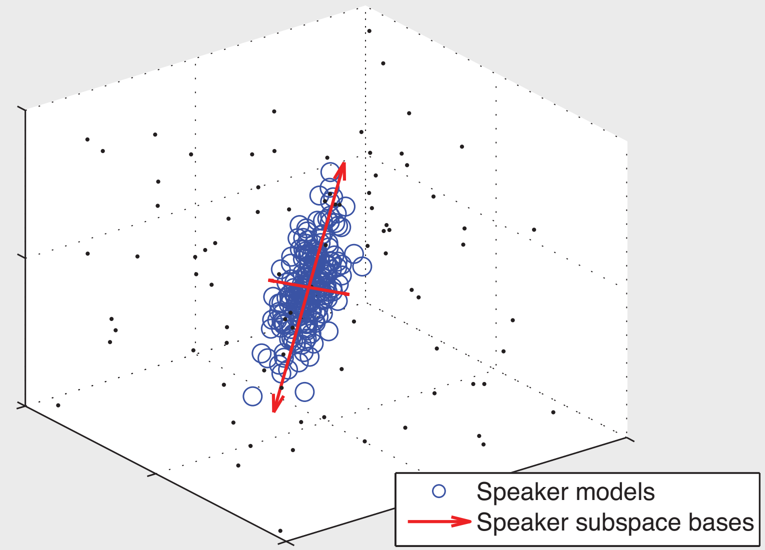
- **"Superhuman" performance**
  - ... in some conditions



Eigenvoices

Kuhn et al. '98, '00
Weiss & Ellis '10

- Idea: Find speaker model parameter space



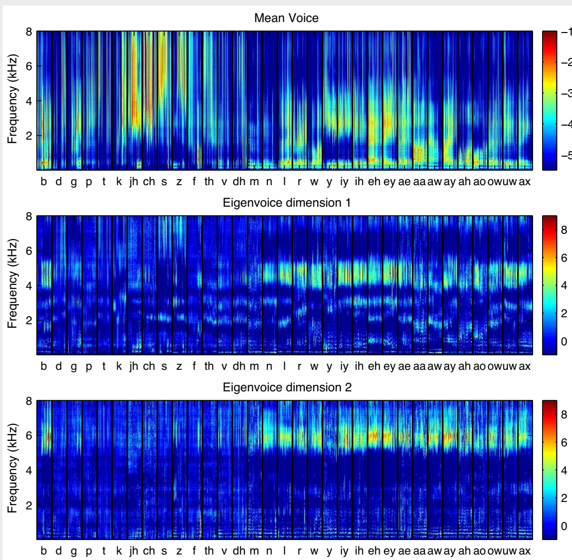
  - generalize without losing detail?

- Eigenvoice model:

$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} + U \quad \mathbf{w} + B \quad \mathbf{h}$$

adapted model	mean voice	eigenvoice bases	weights	channel bases	channel weights

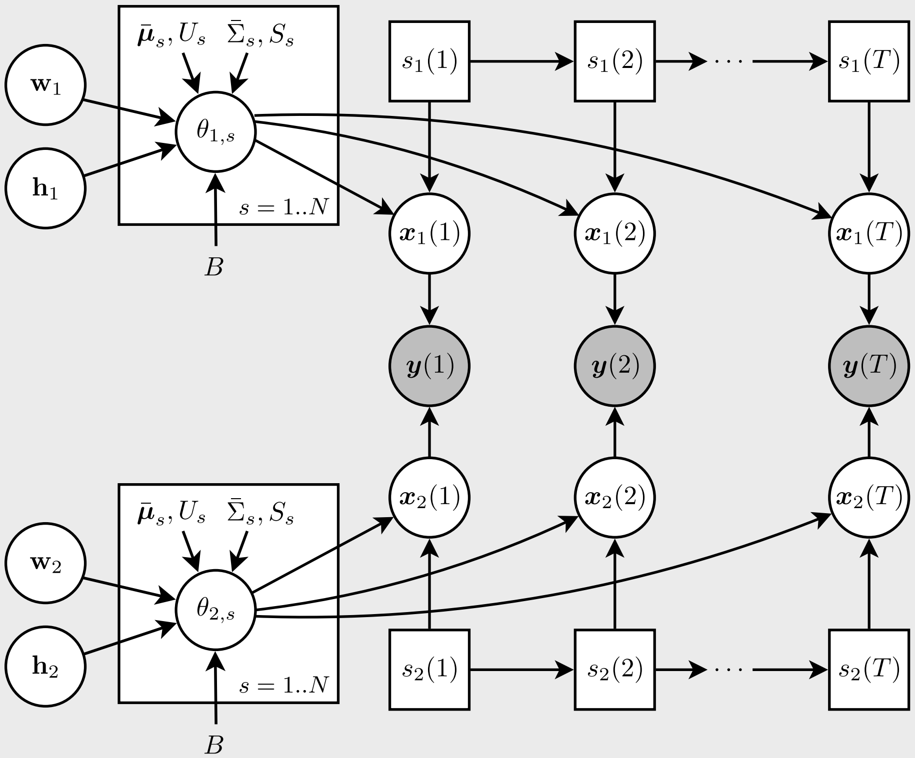
  - 89,600 dimensional space



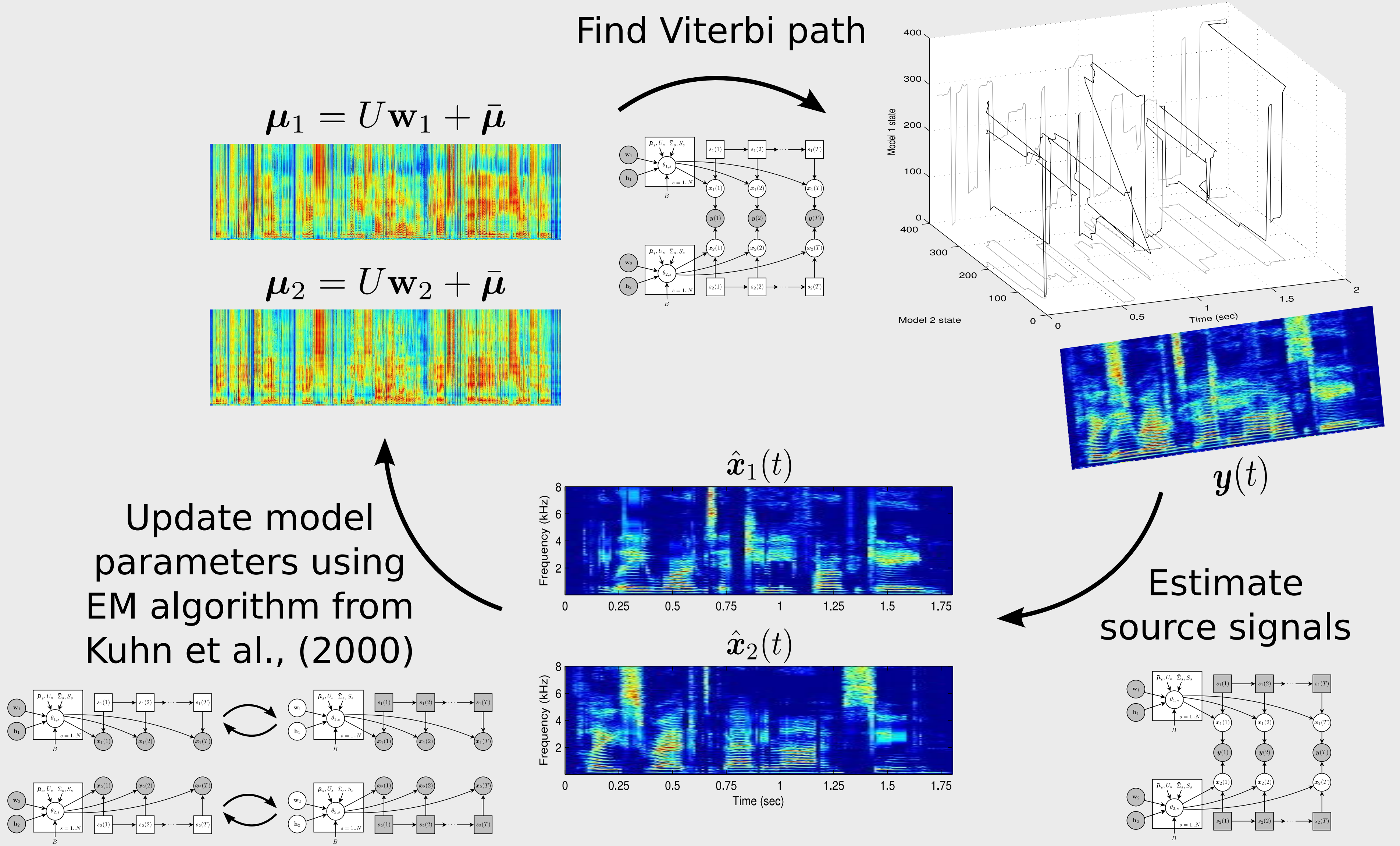
Eigenvoice Speech Separation

Weiss & Ellis '10

- Factorial HMM analysis
  with tuning of source model parameters
  = eigenvoice speaker adaptation

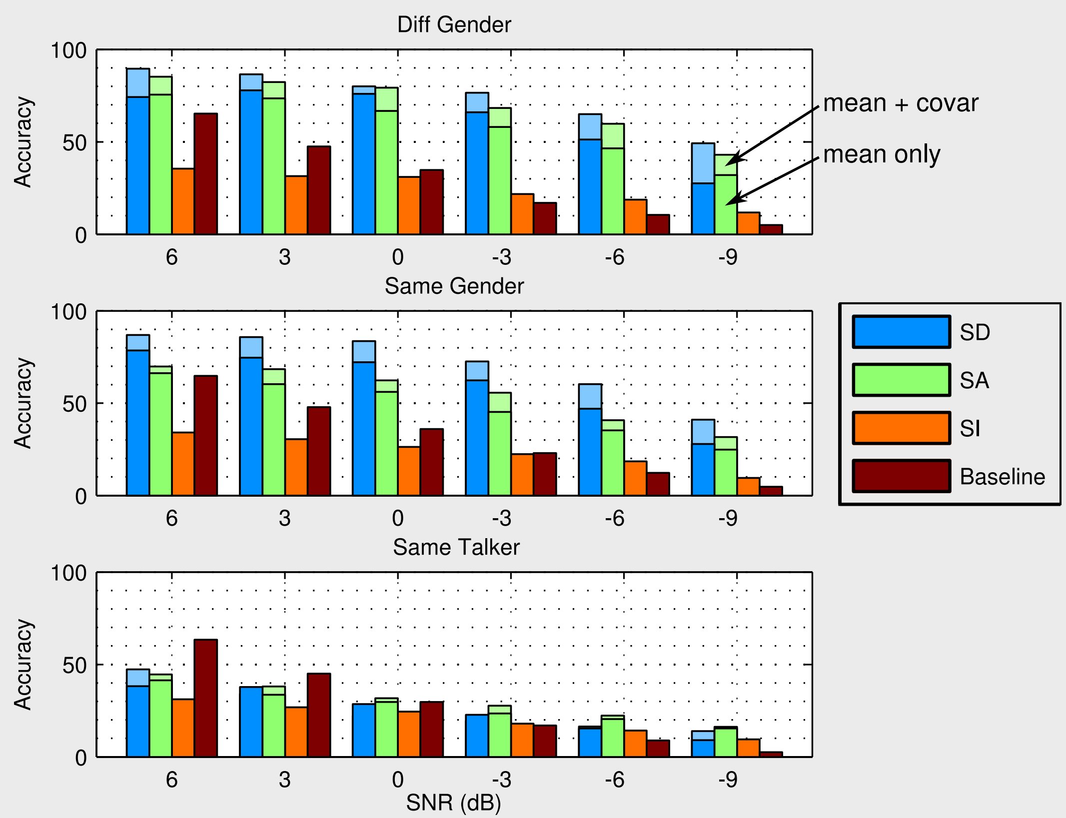


Eigenvoice Speech Separation



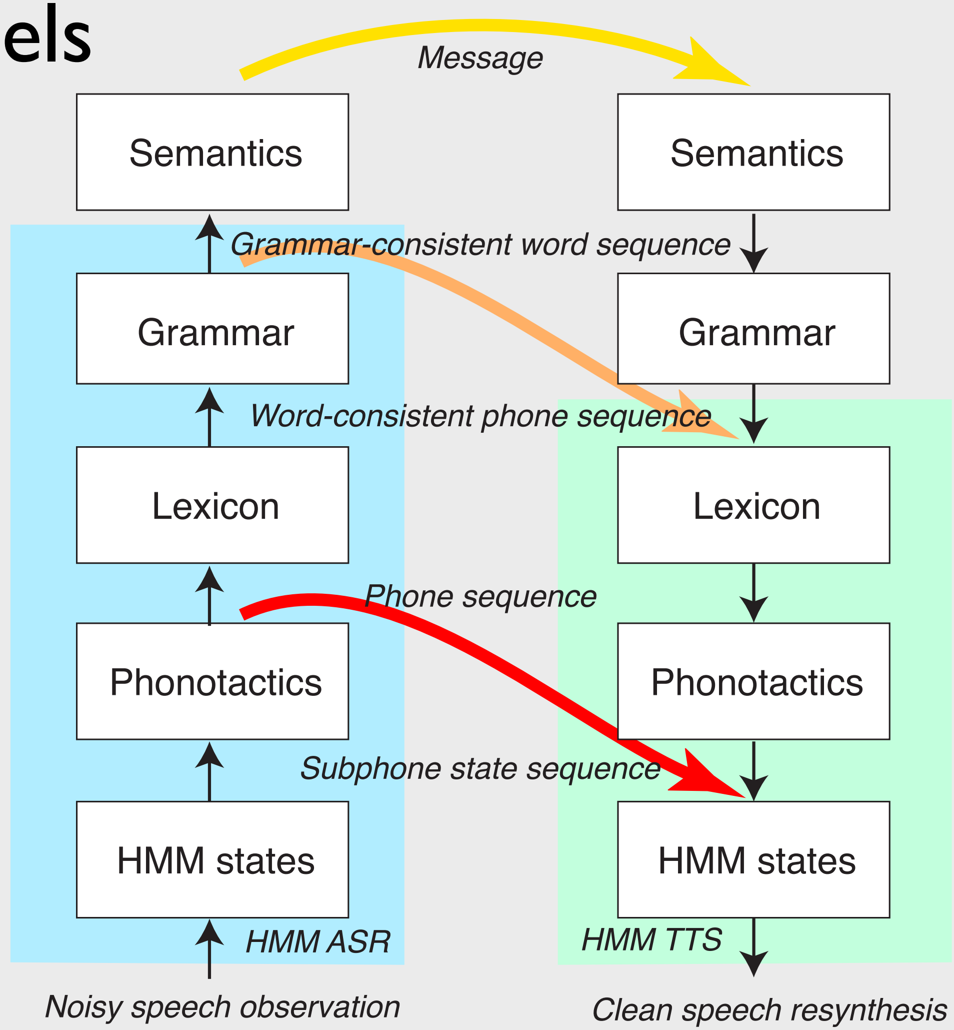
Eigenvoice Speech Separation

- Eigenvoices for Speech Separation task
  - speaker adapted (SA) performs midway between speaker-dependent (SD) & speaker-indep (SI)



Enhancement by Synthesis

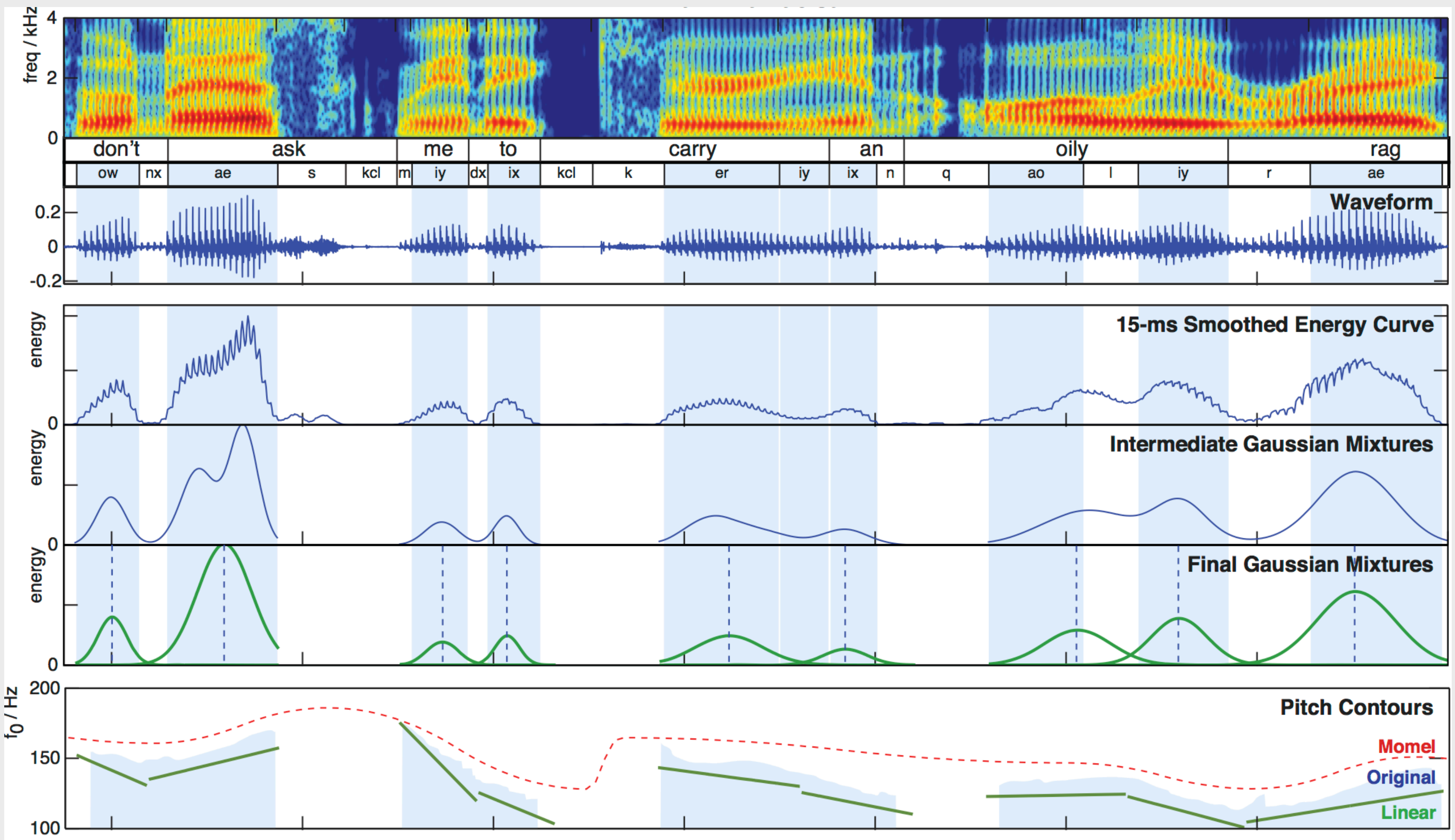
- Current speech synthesizers use ASR-like acoustic models
- Enhance noisy speech by partial recognition then speech synthesis ("copying")
- A novel kind of distortion...



Parametric Speech Models: Pitch

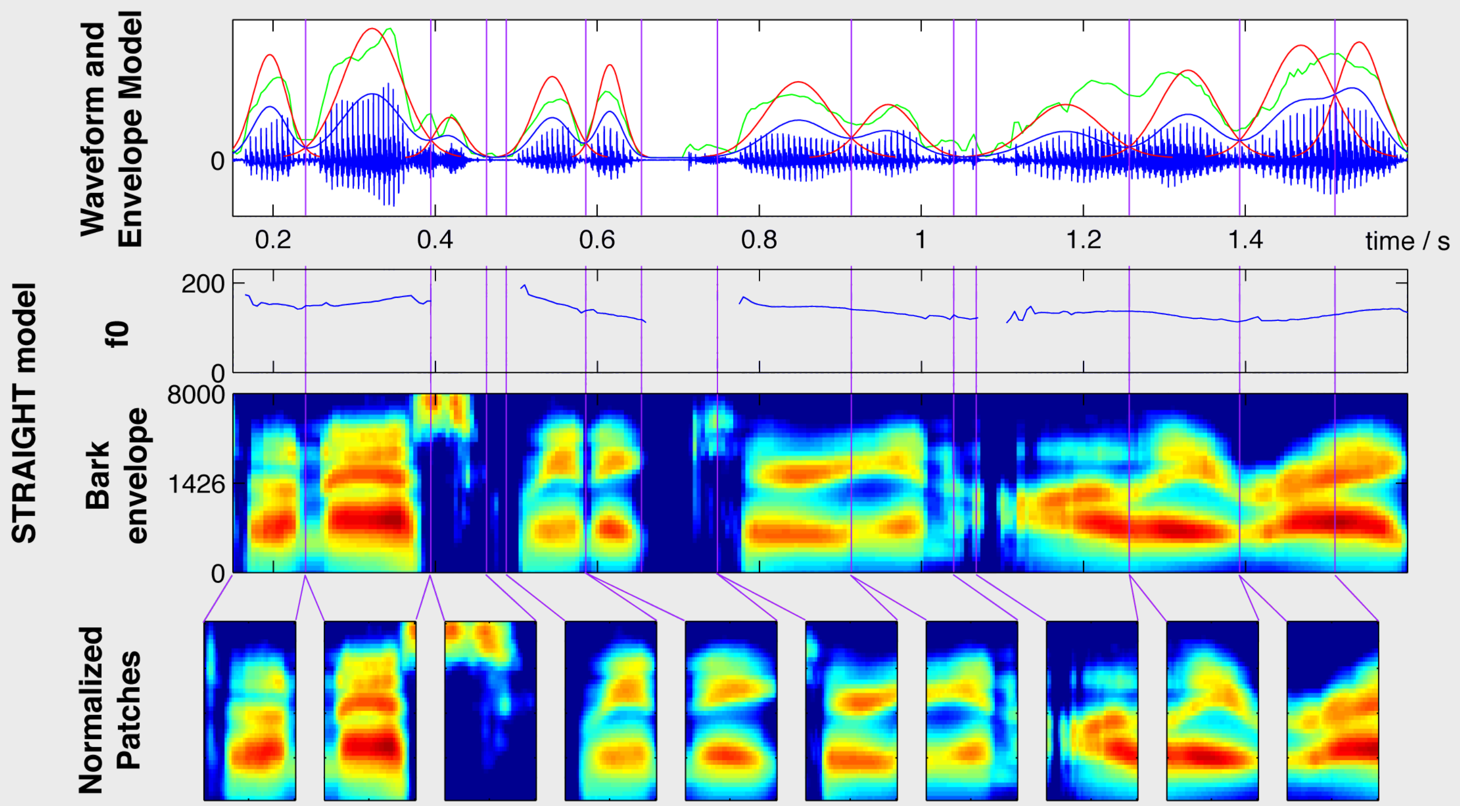
Ravuri & Ellis '08

- Segment into "syllable-like" units by energy
- Model pitch in each syllable as simple line



Parametric Speech: TF Envelope

- Use STRAIGHT for high-quality time-frequency envelope for each syllable
- Build codebook from duration-normalized TFEs



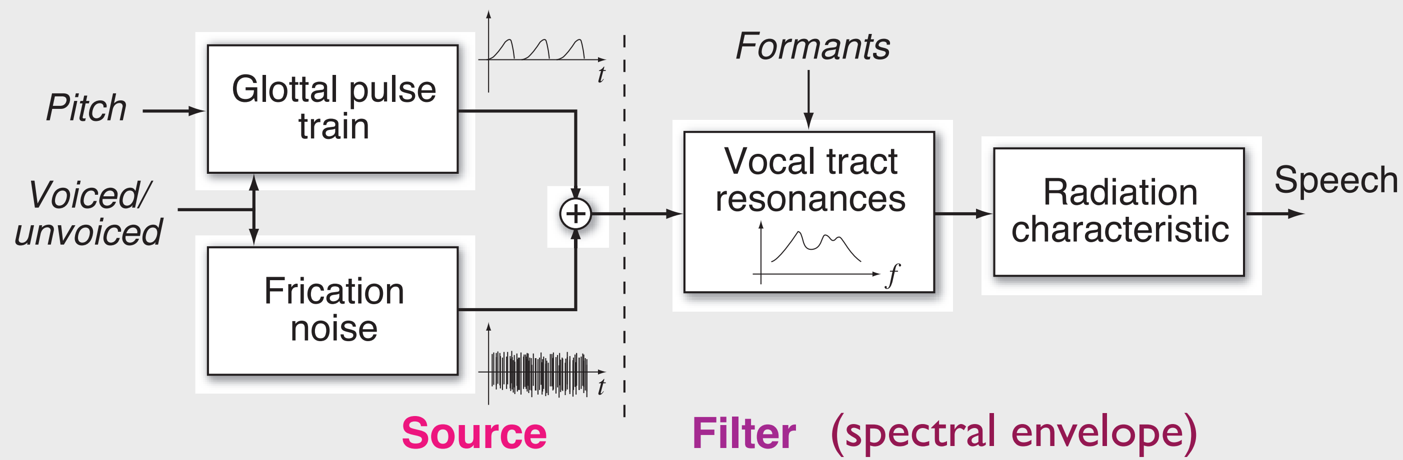
4. Inharmonic Speech

McDermott, Ellis, Kawahara '12

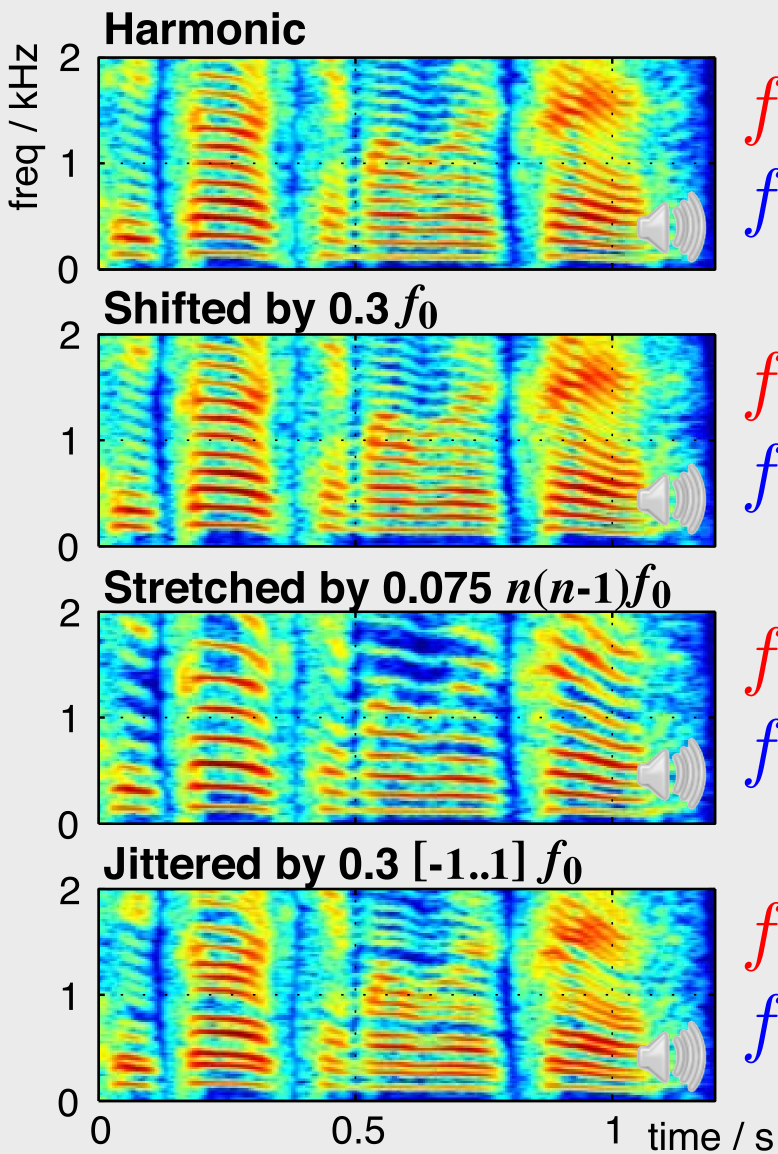
- **Harmonicity** is cited as cue for **fusion**
- Voiced speech has...
  - multiple (resolved) harmonics = "sparse" spectrum
  - .. with similar modulation properties
- How important is the "harmonic pattern"?

$$\sum_{n=-\infty}^{\infty} \delta(t - n\tau) = \frac{1}{\tau}\left(1 + \sum_{k=1}^{\infty} 2\cos k\frac{2\pi}{\tau}t\right)$$

- make "natural" inharmonic speech?



Inharmonic Speech



$$f_n = n f_0$$
$$f_{n+1} - f_n = f_0$$

$$f_n = n f_0 + a f_0$$
$$f_{n+1} - f_n = f_0$$

$$f_n = n f_0 + b(n^2 - n) f_0$$
$$f_{n+1} - f_n = (1 + 2bn) f_0$$

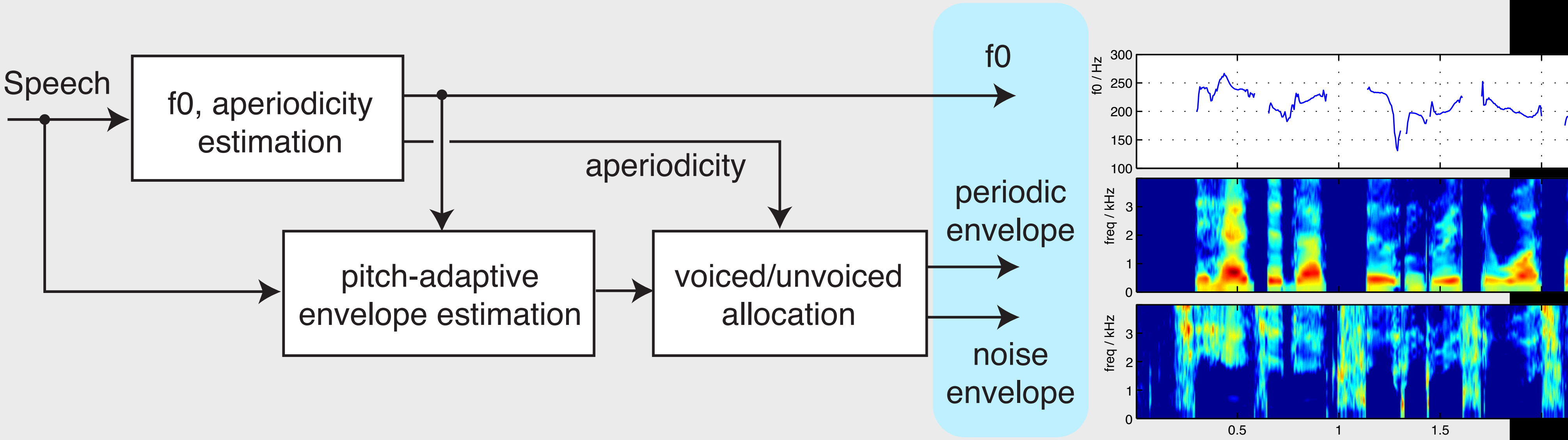
$$f_n = n f_0 + c r_n f_0 \quad r_n \in [-1 \ldots 1]$$
$$f_{n+1} - f_n = (1 + c \Delta r_n) f_0$$

source

$$s(t) = \sum_{n=1}^{N} \cos 2\pi f_n t$$

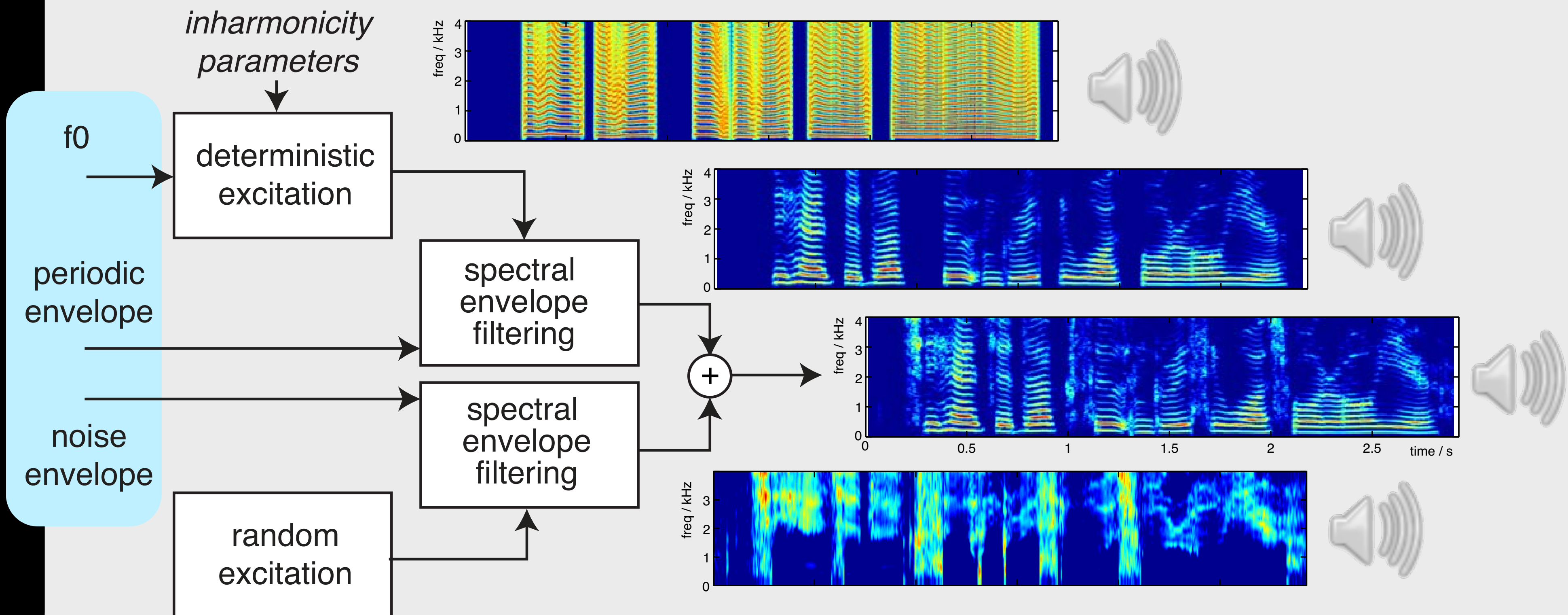
Synthesizing Inharmonic Speech

- Based on STRAIGHT *Kawahara 1999, 2006 ...*
  - decompose speech into:
    - f_0 (pitch track)
    - periodic envelope (voiced speech)
    - noise envelope (unvoiced speech component)



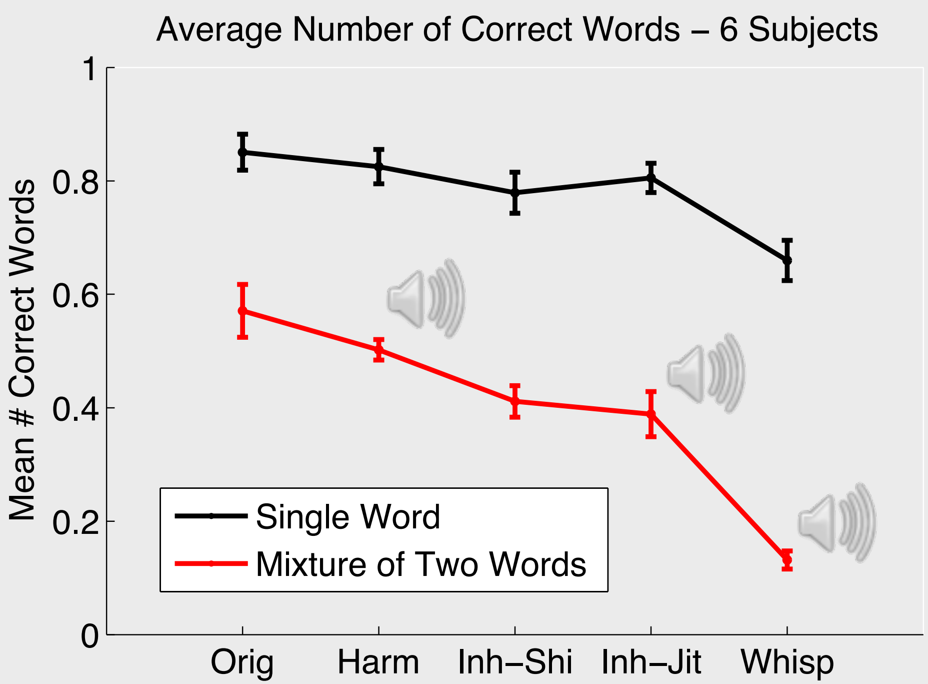
STRAIGHT Synthesis

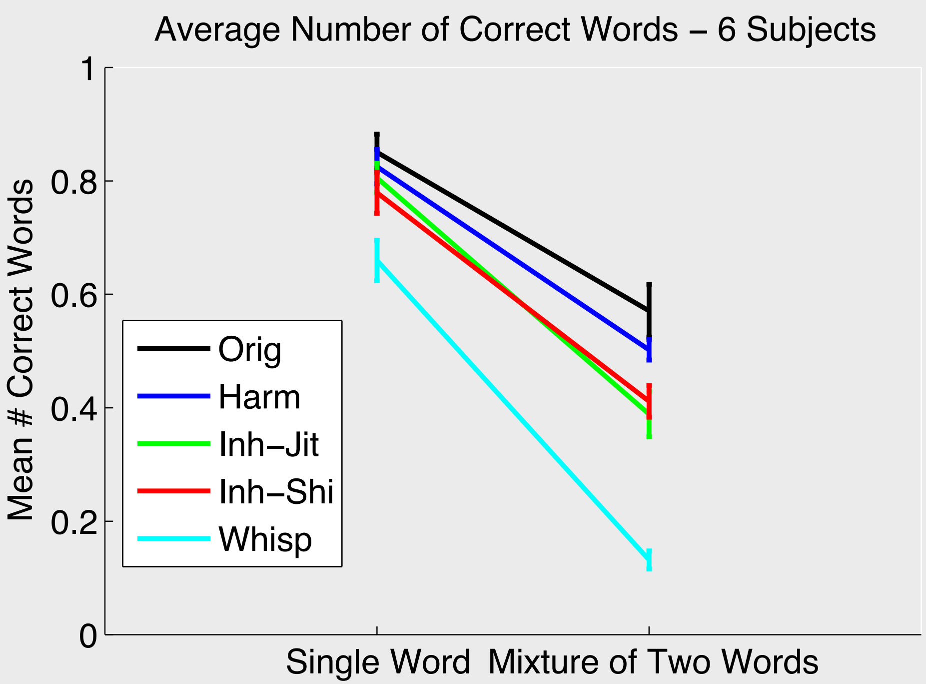
- STRAIGHT periodic source resynthesis
  - ... as individual pitch pulses
  - ... or as a set of Fourier components
    - which can be made inharmonic



Results

- **Harmonic tokens a little easier to understand**
  - but inharmonic tokens much better than whispered
  - different types of inharmonicity seem equivalent
  - Spectral sparsity is a big contributor to separation?





Summary

- Speech in the Wild
  ... real, challenging problem
  ... applications in communications, lifelogs ...

- Speech Separation
  ... by generic properties (location, pitch)
  ... via source models

- Inharmonic Speech
  ... 'natural' speech with inharmonic excitation

Separation vs. Recognition