VQ Source Models: Perceptual & Phase Issues

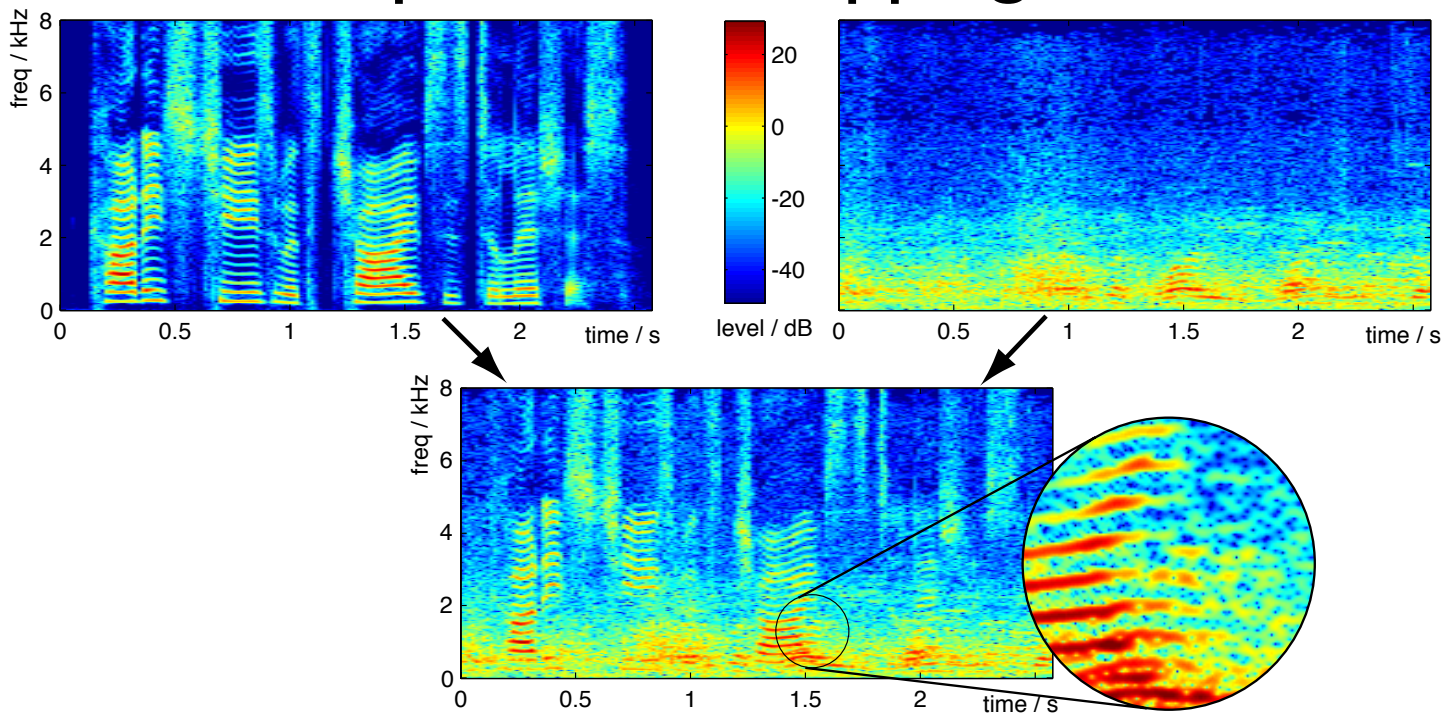
Dan Ellis & Ron Weiss
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

{dpwe,ronw}@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Source Models for Separation
2. VQ with Perceptual Weighting
3. Phase and Resynthesis
4. Conclusions

Single-Channel Scene Analysis

- How to separate overlapping sounds?



- underconstrained: infinitely many decompositions
- time-frequency overlaps cause obliteration
- .. no obvious segmentation of sources (?)

Scene Analysis as Inference

- Ideal separation is rarely possible
  - i.e. no projection can guarantee to remove overlaps
- Overlaps ⇒ Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

$$P(\{S_i\}|\, X) \propto P(X\,|\{S_i\})\; P(\{S_i\})$$

combination physics *source models*

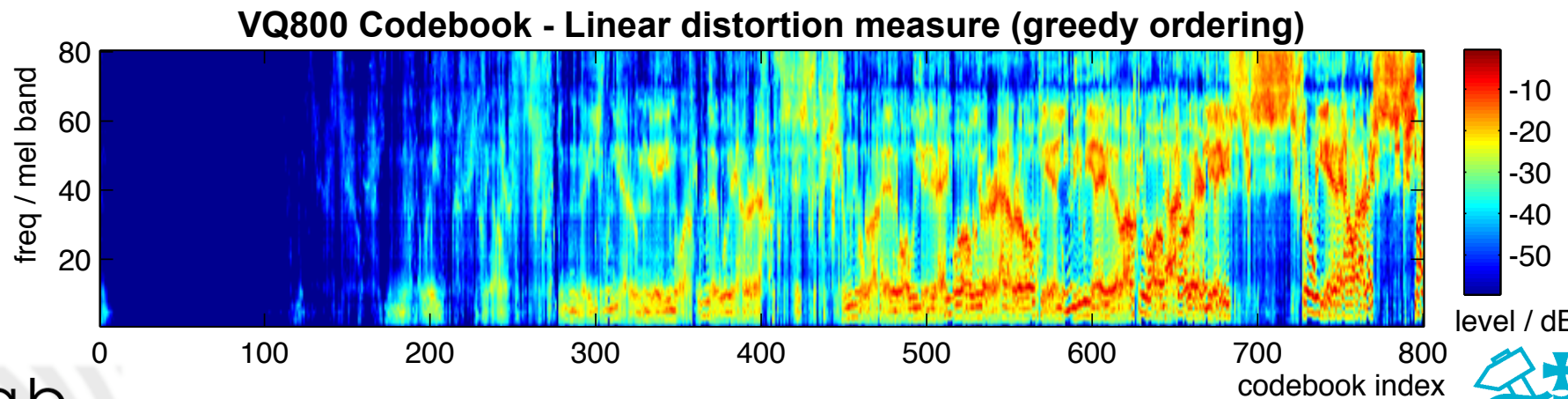
- Better source models → better inference
  - .. learn from examples?

Vector-Quantized (VQ) Source Models

- "Constraint" of source can be captured explicitly in a codebook (dictionary):

$$\mathbf{x}(t) \approx \mathbf{c}_{i(t)} \quad \text{where} \quad i(t) \in 1 \ldots N$$

  - defines the 'subspace' occupied by source
- Codebook minimizes distortion (MSE)
  - by k-means clustering



Simple Source Separation

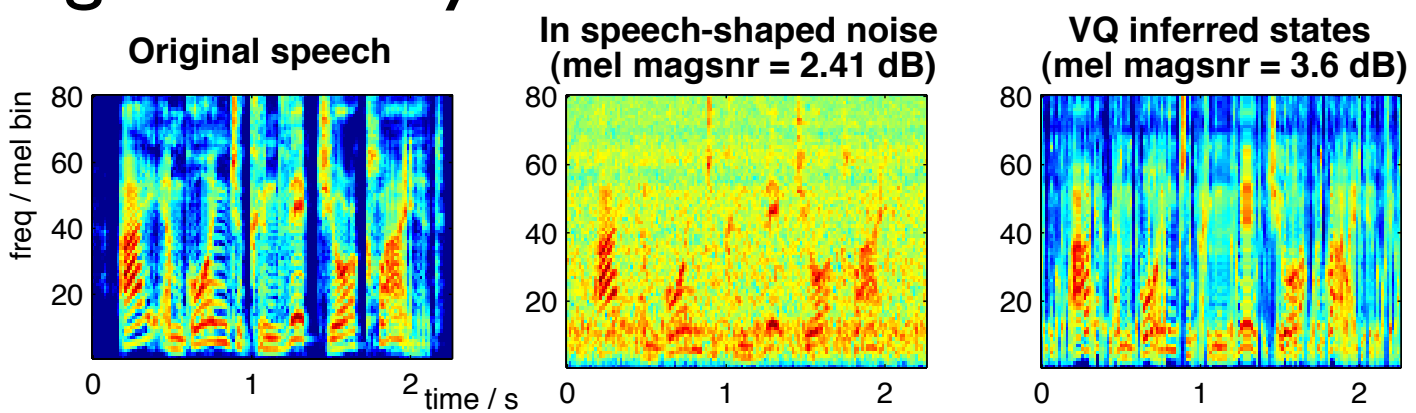
- Given models for sources,
  find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1,i_2) = \mathcal{N}(\mathbf{x};\mathbf{c}_{i1}+\mathbf{c}_{i2},\Sigma)$$ *combination model*

$$\{i_1(t),i_2(t)\} = argmax_{i_1,i_2} p(\mathbf{x}(t)|i_1,i_2)$$ *inference of source state*

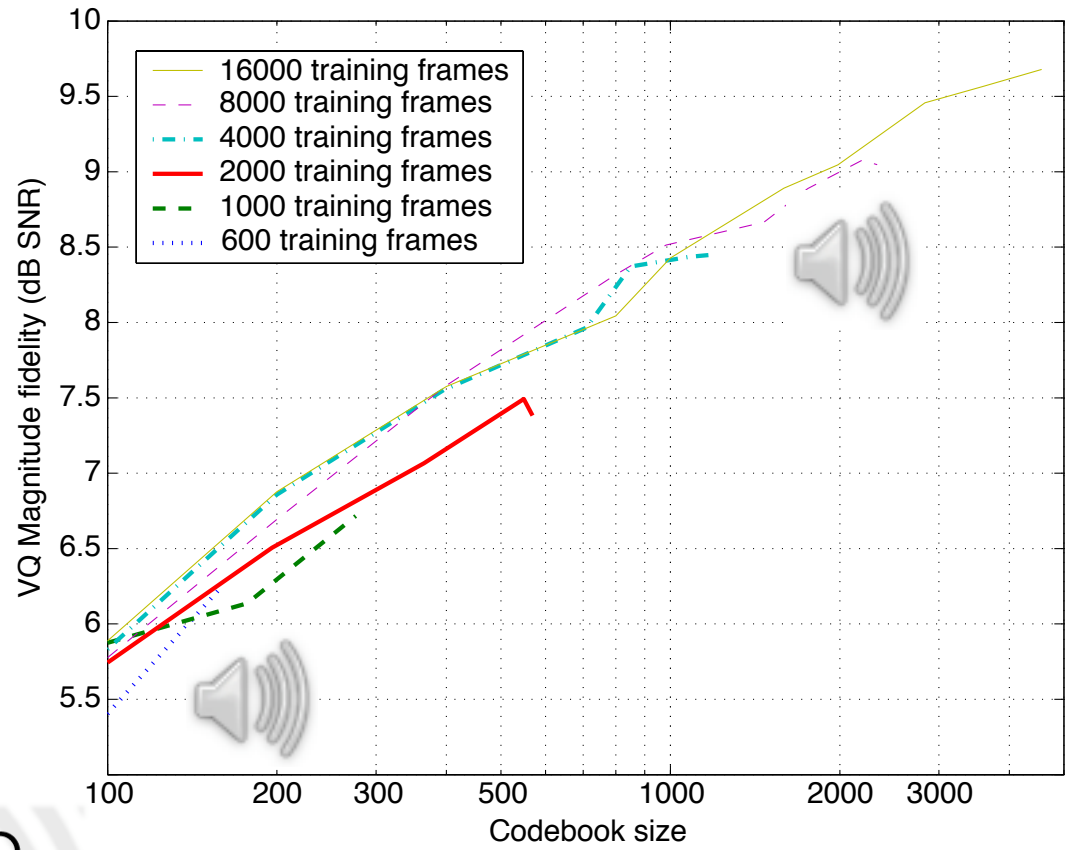
  - can include sequential constraints...
  - different domains for combining $\mathbf{c}$ and defining Σ

- E.g. stationary noise:



Codebook Size

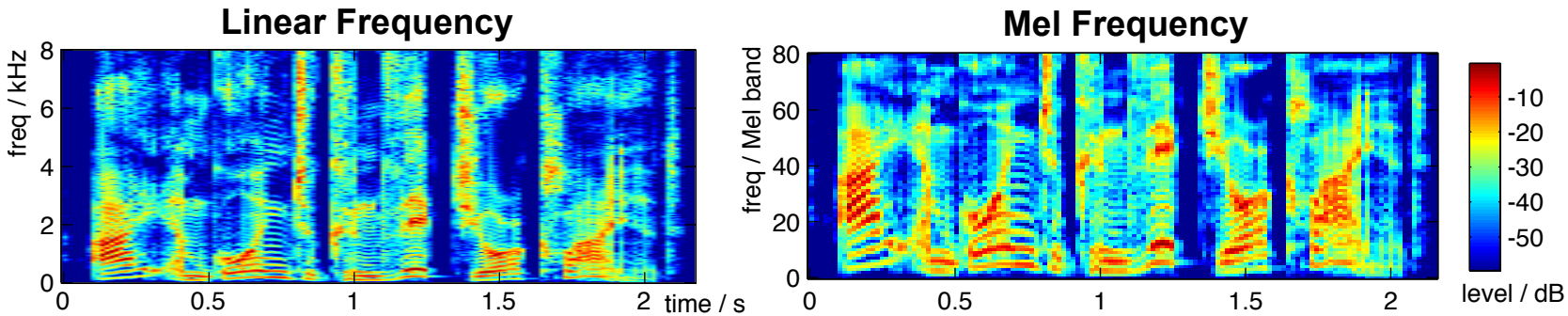
- Two (main) variables:
  - number of codewords
  - amount of training data
- Measure average accuracy (distortion):



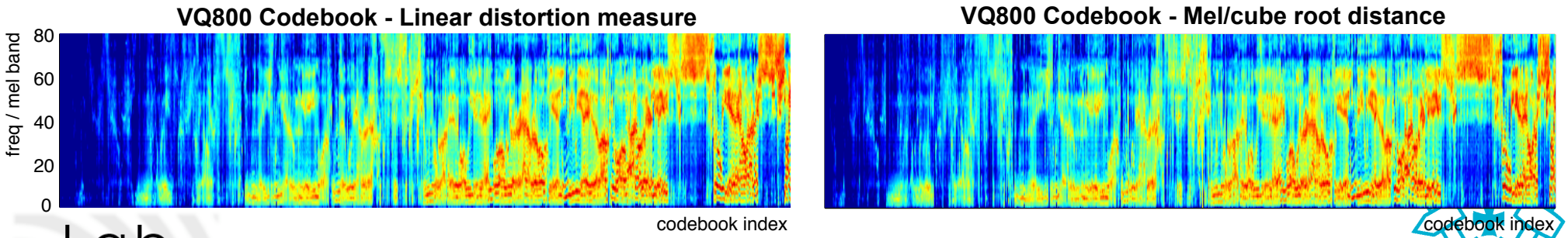
- main effect of codebook size
- larger codebooks need/allow more data
- (large absolute distortion values)

Distortion Metric

- Standard MSE gives equal weight by channel
  - excessive emphasis on high frequencies
- Try e.g. Mel spectrum
  - approx. log spacing of frequency bins

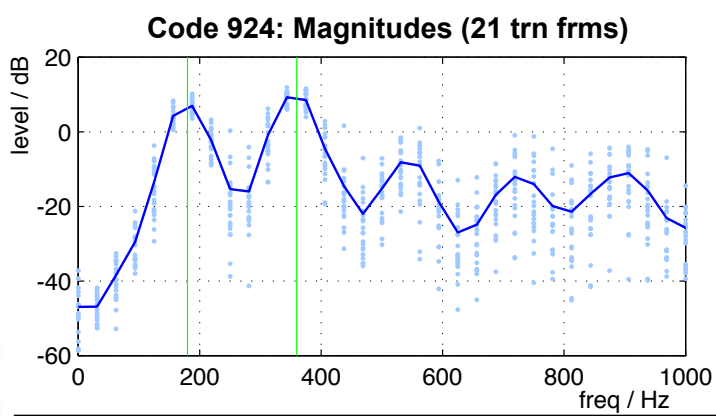


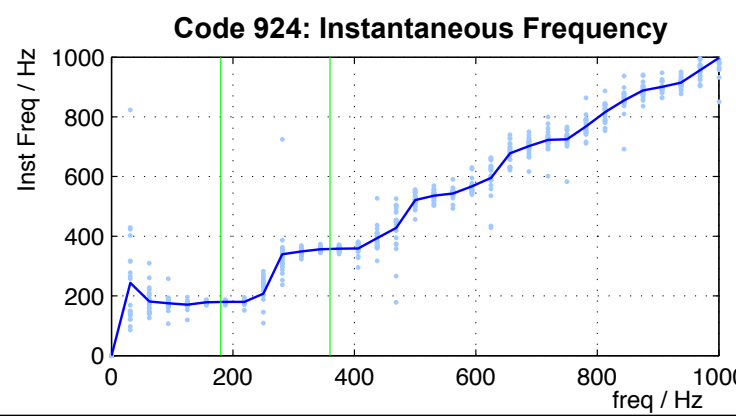
- Little effect (?):



Resynthesis Phase

- Codewords quantize spectrum magnitude
  - phase has arbitrary offset due to STFT grid
- Resynthesis (ISTFT) requires phase info
  - use mixture phase? no good for filling-in
- Spectral peaks indicate common instantaneous frequency $(\partial\phi/\partial t)$
  - can quantize and cumulate in resynthesis
  - .. like the "phase vocoder"





Resynthesis Phase (2)

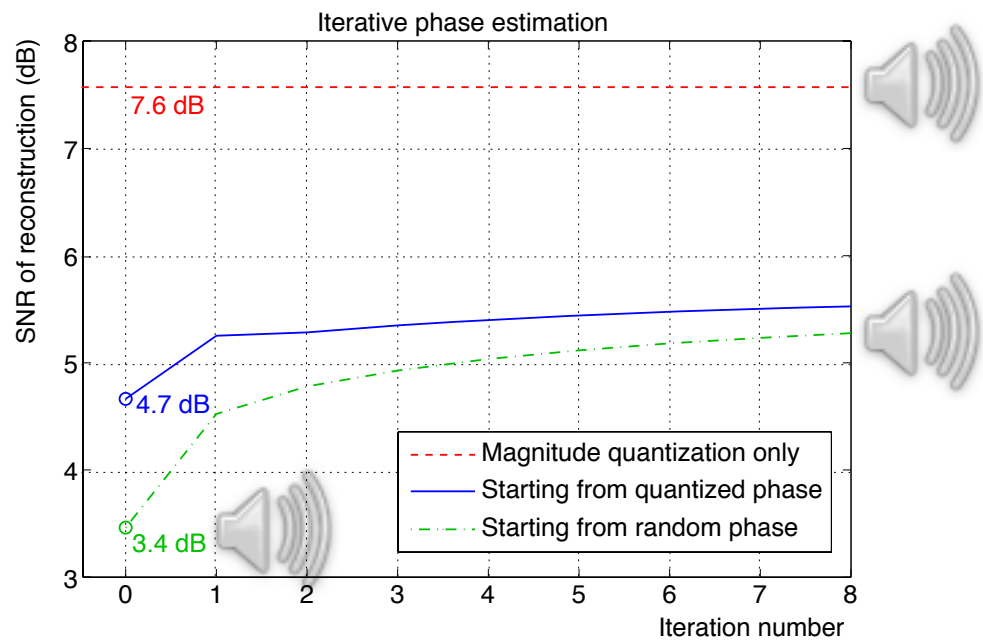
- Can also improve phase iteratively
  - repeat: $X^{(1)}(t,f) = |\hat{X}(t,f)| \cdot \exp\{j\phi^{(1)}(t,f)\}$

$$x^{(1)}(t) = \text{istft}\{X^{(1)}(t,f)\}$$

$$\phi^{(2)}(t,f) = \angle\left(\text{stft}\{x^{(1)}(t)\}\right)$$

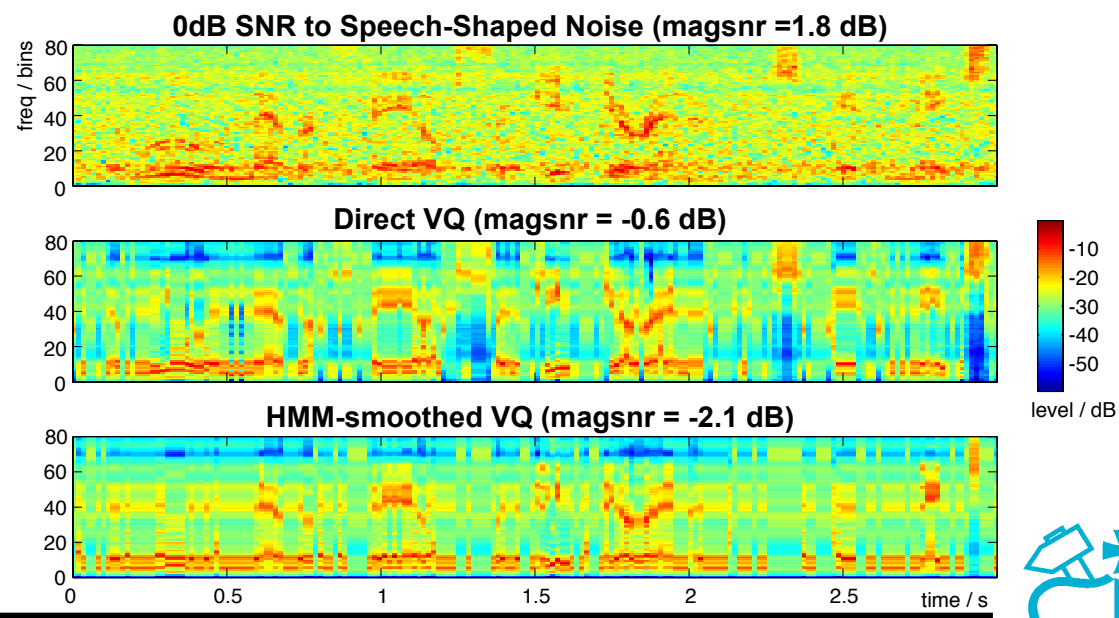
  - goal: $|X^{(n)}(t,f)| = |\hat{X}(t,f)|$

- Visible benefit:



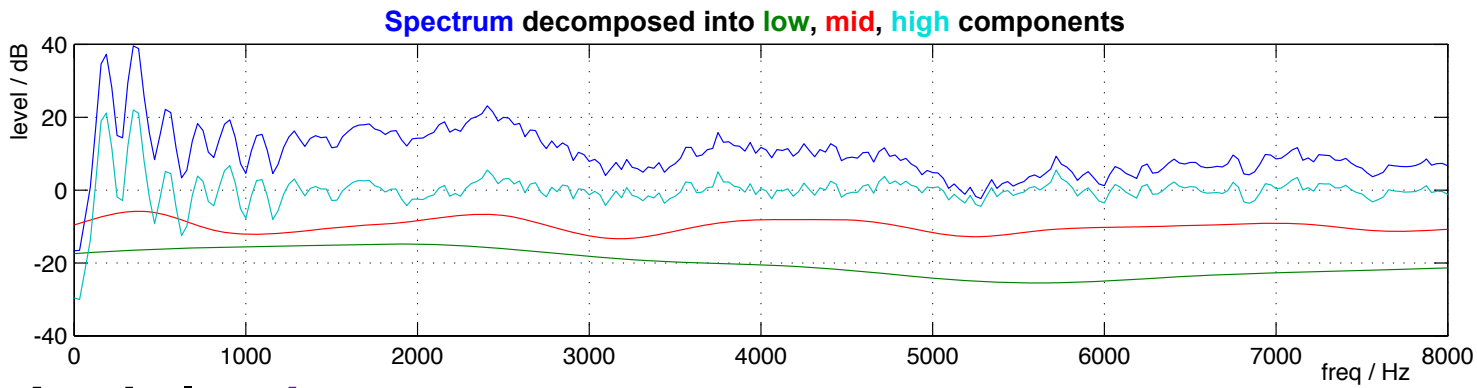
Evaluating Model Quality

- Low distortion is not really the goal; models are to constrain source separation
  - fit source spectra but reject non-source signals
- Include sequential constraints
  - e.g. transition matrix for codewords
  - .. or smaller HMM with distributions over codebook
- Best way to evaluate is via a task
  - e.g. separating speech from noise



Future Directions

- Factorized codebooks
  - codebooks too large due to combinatorics
  - separate codebooks for type, formants, excitation?



- Model adaptation
  - many speaker-dependents model, or...
  - single speaker-adapted model, fit to each speaker
- Using uncertainty
  - enhancing noisy speech for listeners:
    use special tokens to preserve uncertainty

Summary

- Source models permit separation of underconstrained mixtures
  - or at least inference of source state
- Explicit codebooks need to be large
  - .. and chosen to optimize perceptual quality
- Resynthesis phase can be quantized
  - .. using "phase vocoder" derivative
  - .. iterative re-estimation helps more

Extra Slides

Other Uses for Source Models

Projecting into the model's space:

- Restoration / Extension
  - inferring missing parts

- Generation / Copying:
  - adaptation + fitting

Example 2: Mixed Speech Recog.

- Cooke & Lee's Speech Separation Challenge
  - short, grammatically-constrained utterances:

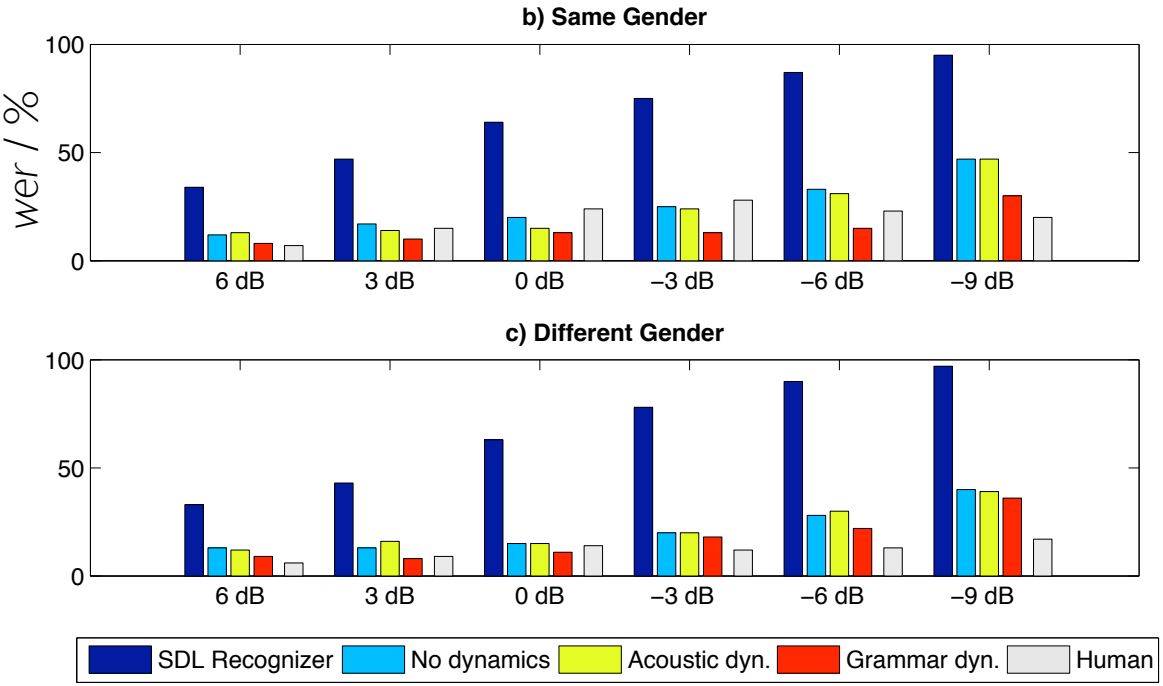
<command:4><color:4><preposition:4><letter:25><number:10><adverb:4>

e.g. "bin white at M 5 soon"

t5_bwam5s_m5_bbilzp_6p1.wav

- IBM's "superhuman" recognizer:

Kristjansson et al. Interspeech'06

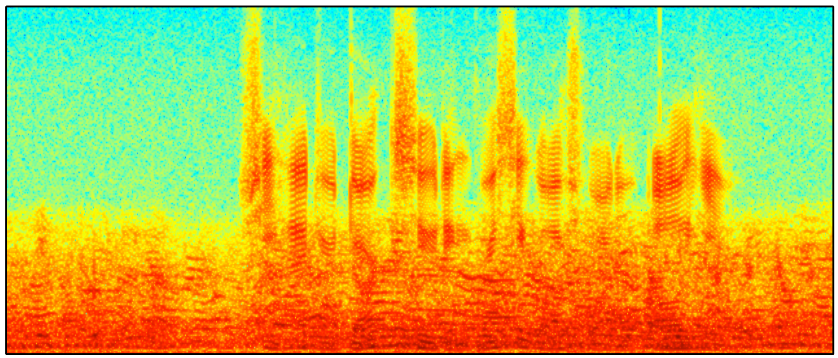


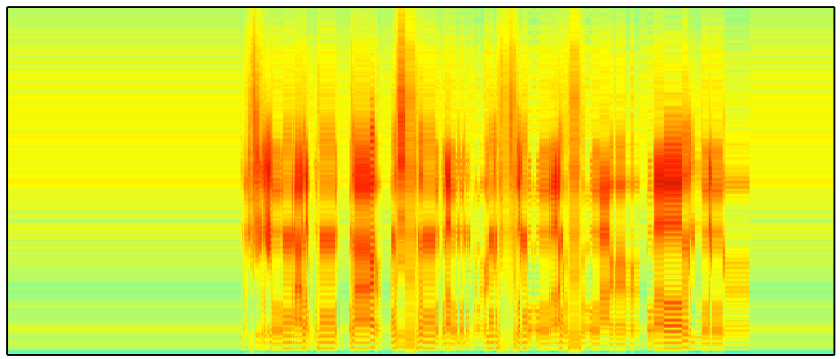
  - Model individual speakers (512 mix GMM)
  - Infer speakers and gain
  - Reconstruct speech
  - Recognize as normal...
- Grammar constraints a big help

Model-Based Separation

Varga & Moore'90
Roweis'03...

- Central idea:
  Employ strong learned constraints
  to disambiguate possible sources
  - $\{S_i\} = \text{argmax}_{Si}\ P(X \mid \{S_i\})$
- e.g. fit speech-trained Vector-Quantizer
  to mixed spectrum:



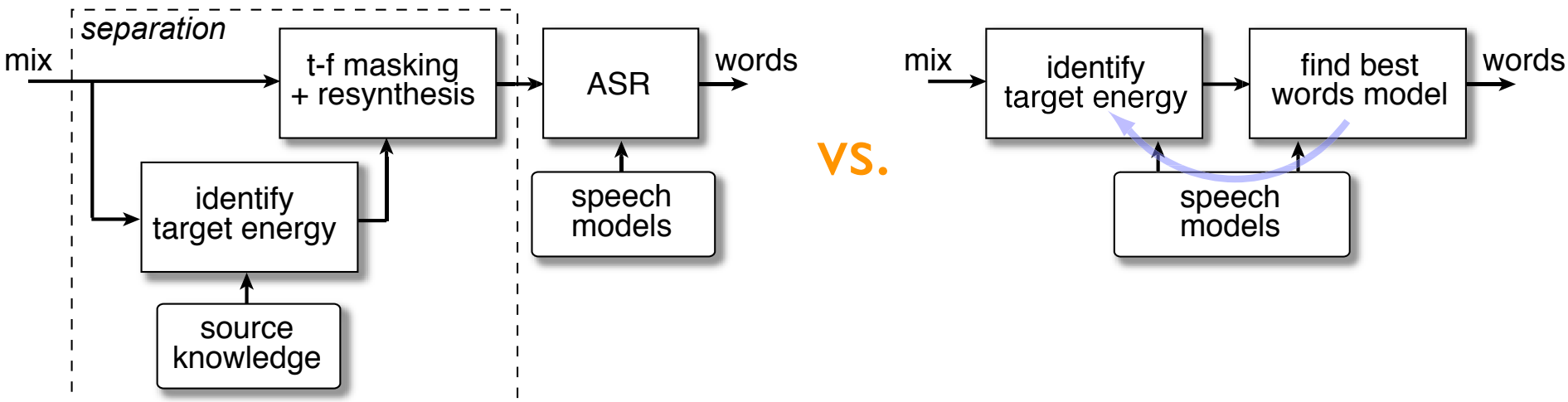


from Roweis'03

  - separate via T-F mask (again)

Separation or Description?

- Are isolated waveforms required?
  - clearly sufficient, but may not be necessary
  - not part of perceptual source separation!
- Integrate separation with application?
  - e.g. speech recognition



  - words output = abstract description of signal

Evaluation

- How to measure separation performance?
  - depends what you are trying to do
- SNR?
  - energy (and distortions) are not created equal
  - different nonlinear components *[Vincent et al. '06]*
- Intelligibility?
  - rare for nonlinear processing to improve intelligibility
  - listening tests expensive
- ASR performance?
  - separate-then-recognize too simplistic; ASR needs to accommodate separation