

Spectral Signal Processing for ASR

Melvyn Hunt

Dragon Systems UK R&D

(a subsidiary of *Dragon Systems, Inc.*)




Issues in This Talk

- 📁 Should we expect modeling human auditory properties to help ASR?
- 📁 The “standard” acoustic representation
 - And some successful variants
- 📁 Why are these representations successful?
 - Is it because they model hearing?

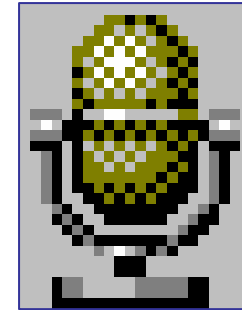
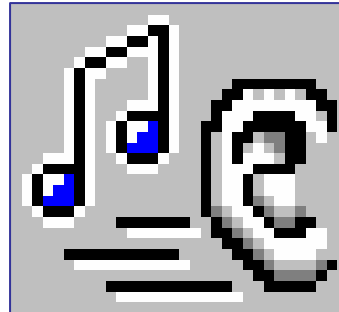
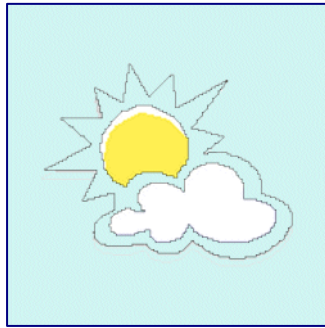
“Acoustic Representation” Is a Hazy Concept

- 📁 No sharp boundary between acoustic front-end and acoustic pattern matching
 - Choice of metric;
 - Spectral subtraction *vs.* Model Conditioning
- 📁 In humans, there may be no single acoustic representation, but rather multiple representations (Pols)


Why Should Modeling Human Hearing Help ASR?

-  Humans are better than machines at recognizing speech (USC notwithstanding!)
 - Even without syntactic/semantic help




Airplanes don't flap their wings,
so why should ASR copy our ears?
... A false analogy!




Why Should Modeling Human Hearing Help ASR? contd.


-  Ears evolved before speech
 - So speech evolved under the constraints of hearing.
 - It would be surprising if there is anything useful in the speech waveform that human ears can't perceive.
 - That's why the "aircraft don't flap their wings" argument is fallacious.

Why Modeling Human Hearing Might *Not* Help ASR





-  Techniques suited to human physiology may be unsuitable for current hardware.
-  Modeling only part of human auditory perception may be counter-productive.
-  What we measure as superior human phonetic classification may depend on a production-based analysis in the brain.

Progress in ASR: the EuroSpeech Paradox

 If so many researchers are discovering so many improvements, how come ASR isn't perfect??

 The reality is:
Progress has been solid but slow
—especially if the effects of more powerful hardware and bigger training corpora are excluded.

Why “Advances” aren’t Always Real

-  May simply not be statistically significant.
-  May depend on the particular conditions, speaker population or task.
-  May depend on other peculiar features of the complete system.
-  The standard alternative may not have been implemented optimally.

I am going to consider only techniques that have proved effective in multiple:

- ⊗ implementations
- ⊗ sites
- ⊗ tasks
- ⊗ *etc.*

The “Standard” Acoustic Representation —and a Few Successful Variants




Standard:

- mel-scale cepstrum and δ and $\delta\delta$ cepstrum derived from log energies in a (simulated) filter-bank sampled every 10-20 ms.

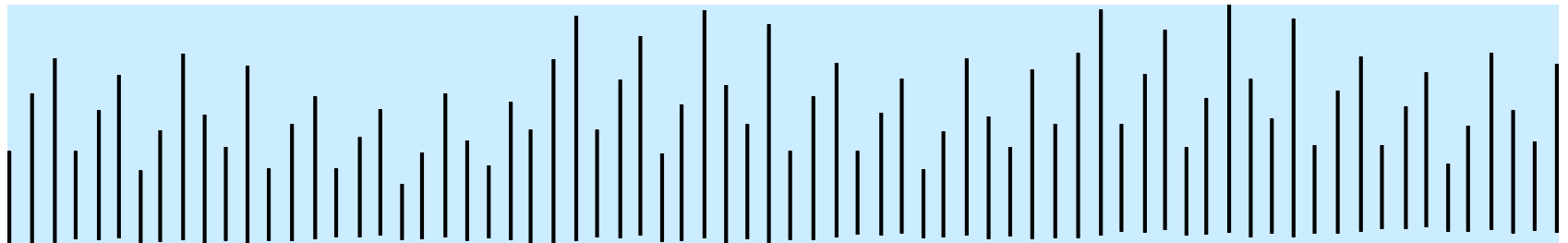
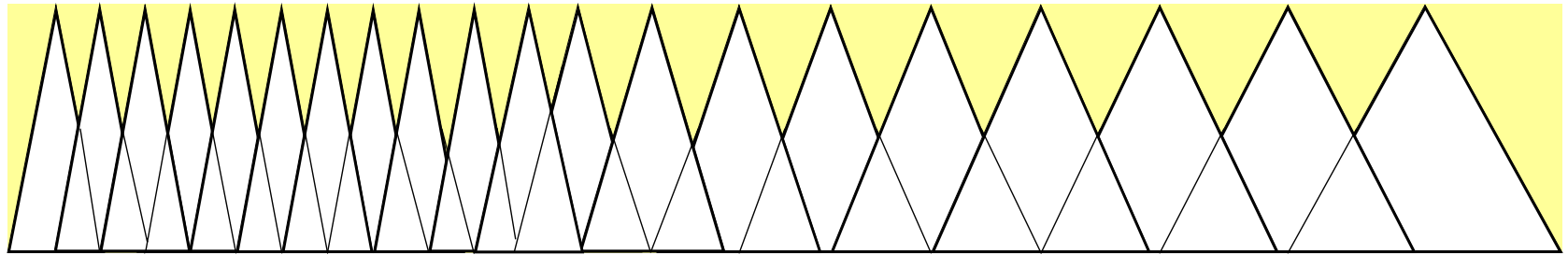
Variants:

- PLP
- LDA
- (Cube) Root Power Energy Representations
- Cepstrum Correlation Measures

Why Use a Filter-Bank?

-  Smooths out irrelevant fine structure (pitch harmonics, noise) before taking logs.
-  Allows control of frequency resolution and of the weight given to different parts of the spectrum.
-  Usually use triangular filters crossing at 3 dB points on technical mel scale

Technical Mel-Scale Filter-Bank



Why Use the Mel Scale?

 Conventional answer:

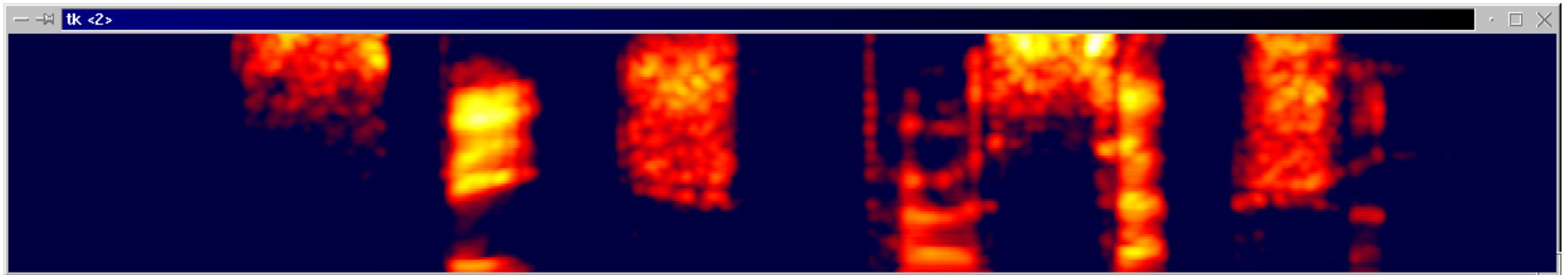
—because it reflects the frequency resolution of the human ear.

 BUT exact replication of the mel scale doesn't seem to be important.

 Need for a mel-like scale is apparent purely from the speech signal itself.

Spectrogram of "speech perception"


Note that main activity at higher frequencies is in the unvoiced sounds, which are broadband and noisy.



Motivations from the Speech Signal for a Mel-like Scale

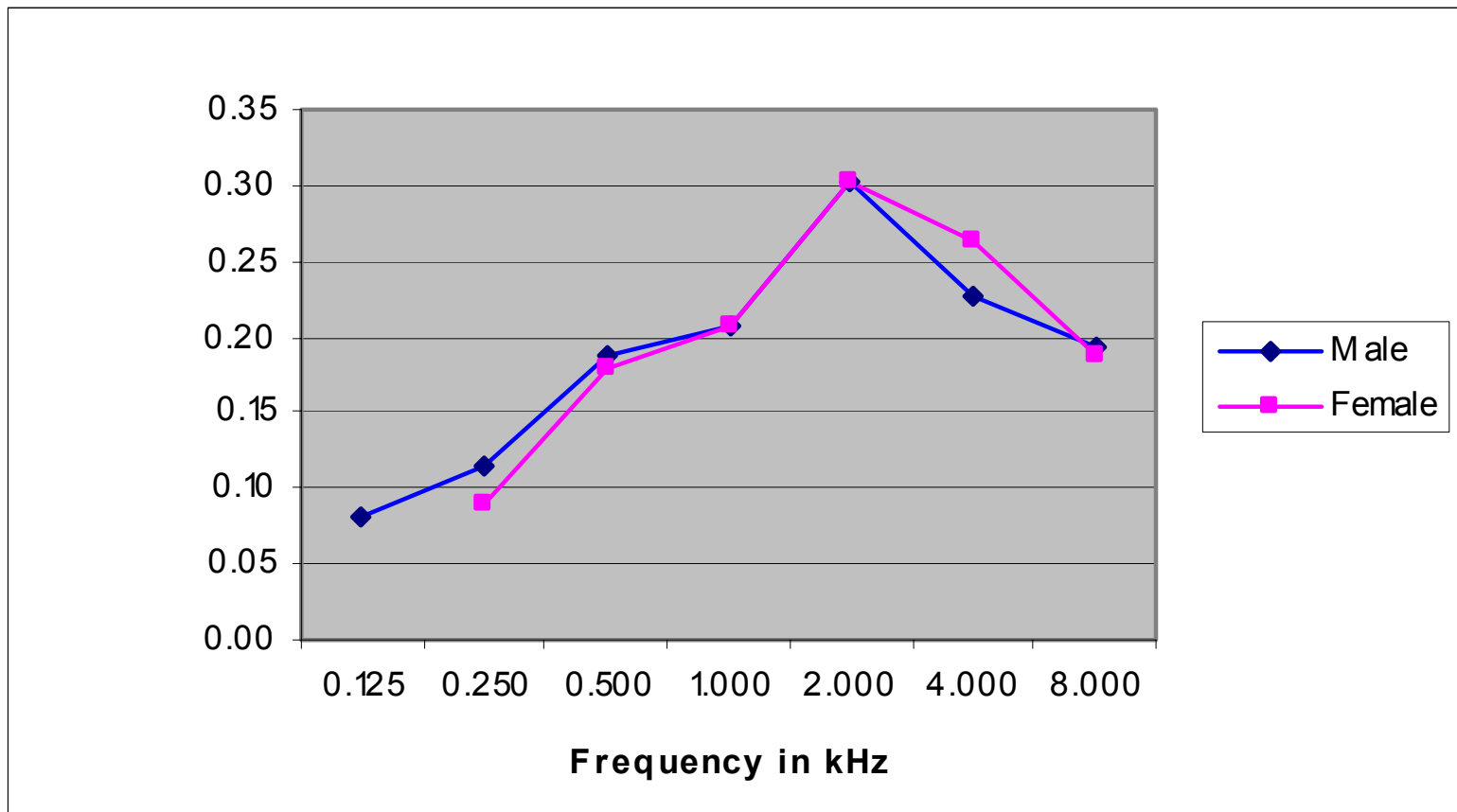
- ➡ Formant bandwidths increase with frequency
- ➡ Higher formants ($> F_3$) contribute little to intelligibility of voiced sounds
 - and are difficult to control independently of the lower formants.
- ➡ Only unvoiced sounds depend on distinctions at higher frequencies
 - and they have broad bandwidths and a “noisy” spectrum.

Motivations from the Speech Signal for a Mel-like Scale

 All the motivations just listed stem from acoustical and physiological properties of speech production;

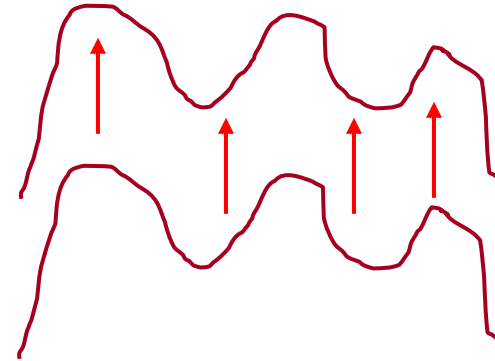
 They don't depend on human auditory properties.

Contributions to speech intelligibility from each octave band using CVC nonsense words (Steeneken):
not flat on a technical mel scale
— maybe need even wider high-frequency bands.



Why Use Log Energies?

- 📁 Gain invariance
 - spectrum shape is preserved under gain changes



- 📁 Normal distribution
 - log power spectrum of speech signal has approx. multivariate Gaussian distribution
 - convenient for linking spectral distances to log probabilities

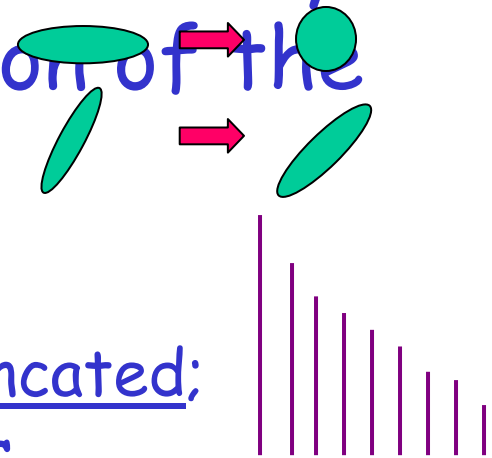
- 📁 But downside is excessive sensitivity to low-energy (low SNR) parts of spectrum

Why Apply a Cosine Transform to the Log Channel Energies?



📖 It's not to remove harmonic structure.

📖 It's because it's an approximately uncorrelated representation of the speech spectrum:

- Can apply weighting schemes;
- Allows representation to be truncated;
- Isolates overall energy level in C_0 .




Dynamic Cepstrum Coefficients

-  Dynamic coefs. also uncorrelated.
-  Regression computation is a band-pass filtering of the time sequence
 - with “DC” term removed
 - so dynamic coefs. are unaffected by constant linear filtering
 - RASTA reintegrates dynamic coefs, giving a “static” representation unaffected by slowly varying filtering


Variants on the Standard Analysis: Linear Discriminant Analysis (LDA)

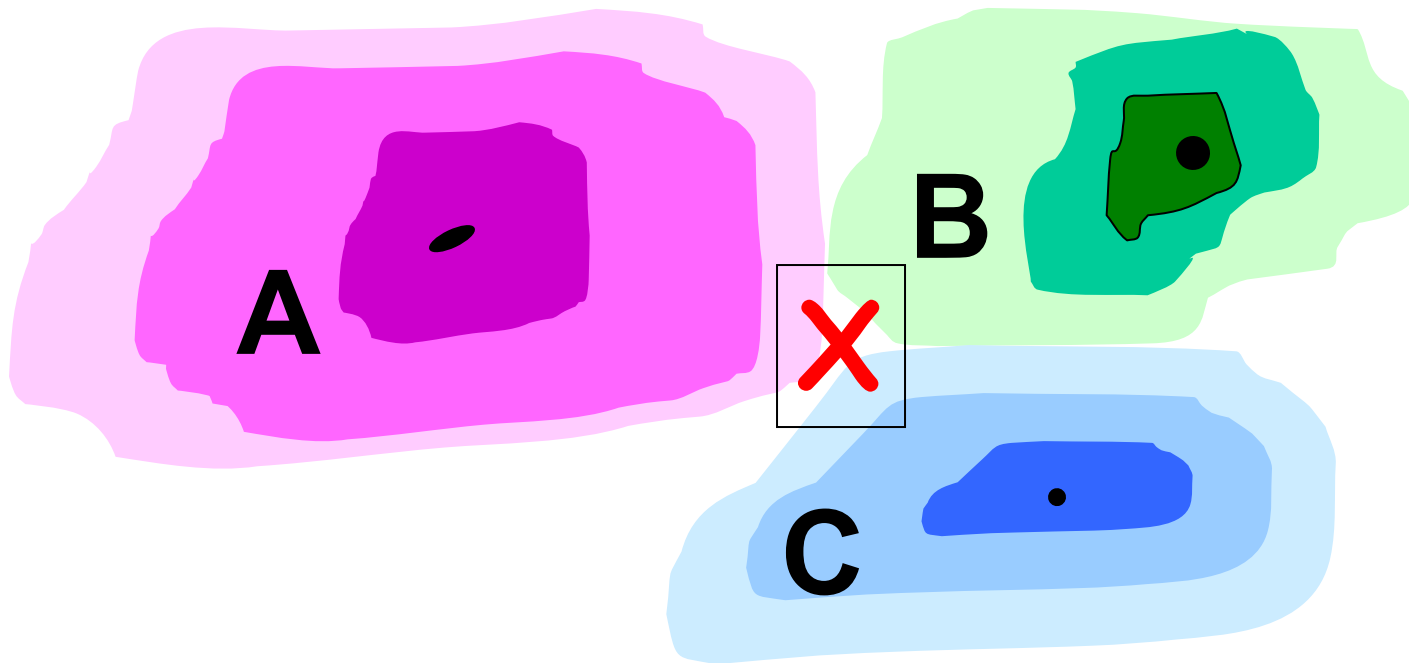
 A linear transformation of any set of coefficients representing the speech spectrum

 The transformed coefficients are:

- arranged to make Euclidean distances correspond to log probabilities.
- uncorrelated
- ordered according to their usefulness.

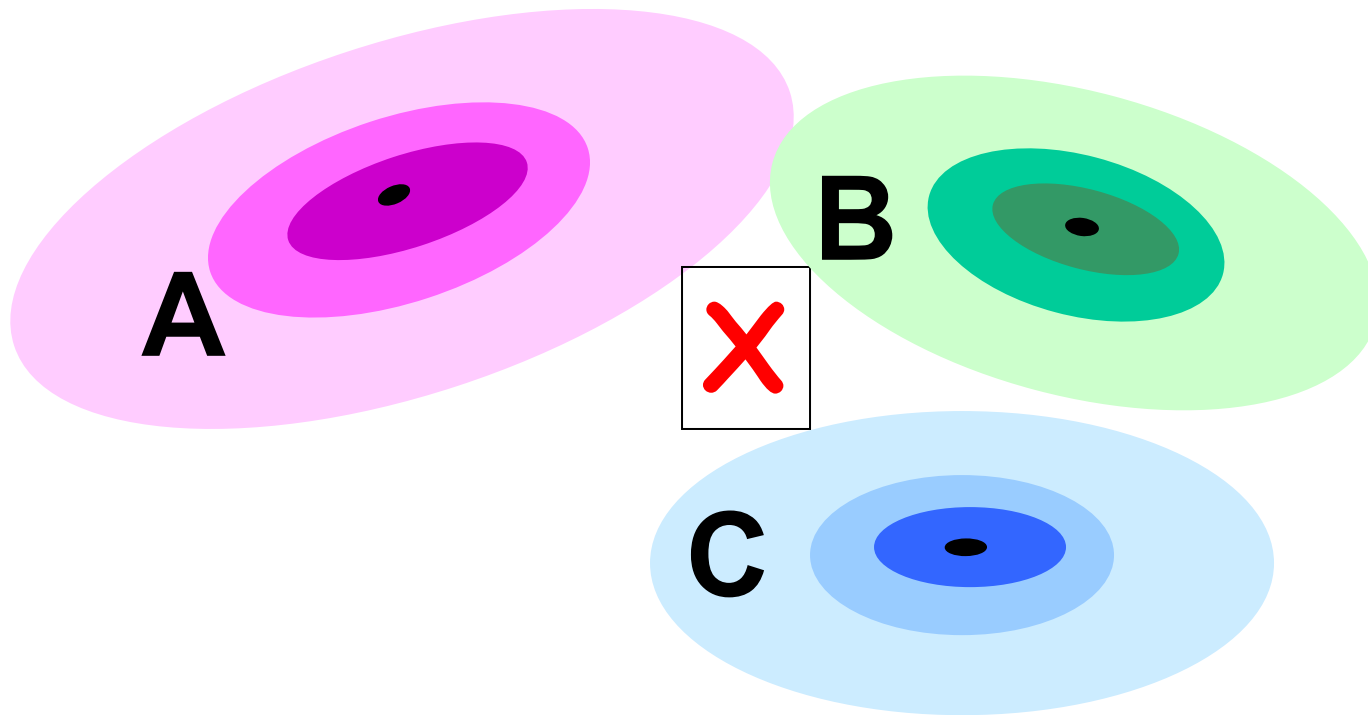
Derivation of LDA: Pattern Classification

 The problem is to assign a sample to one of several classes



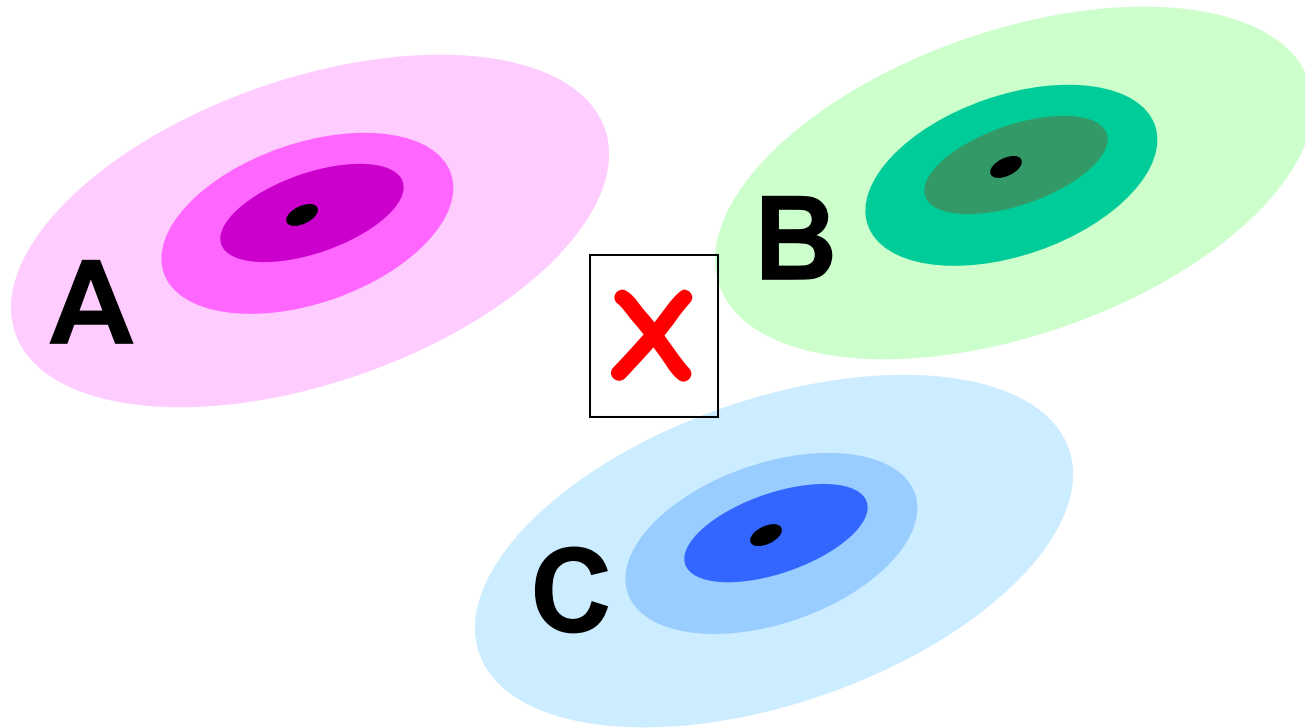
Derivation of LDA:

 We normally assume that the classes have multivariate Gaussian distributions




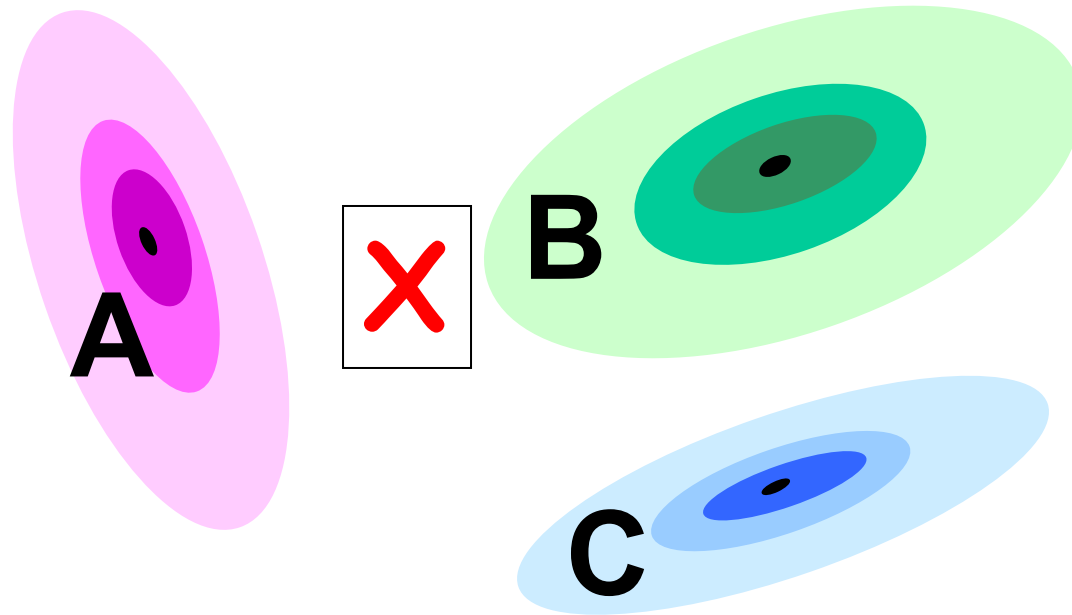
Derivation of LDA:

 LDA assumes that the classes have identical multivariate Gaussian distributions



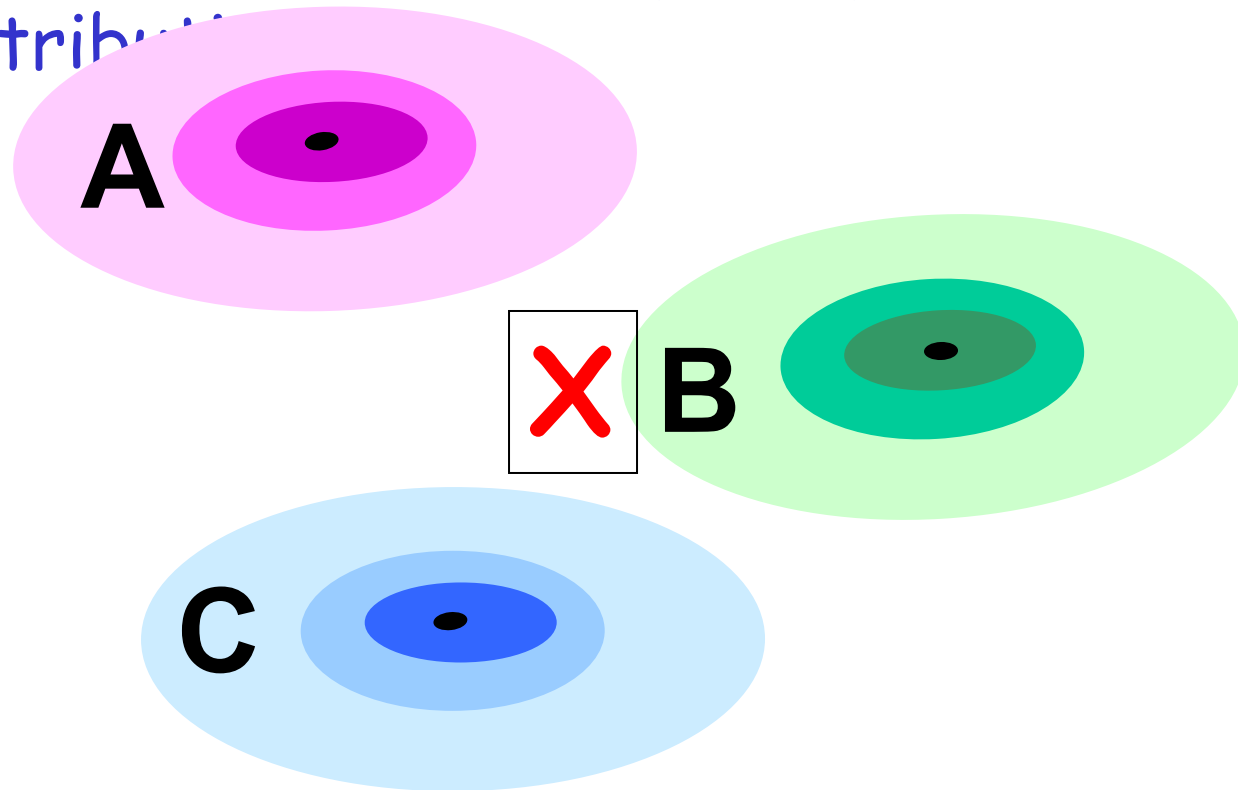
An Extension of LDA: the Heteroscedastic Transform

 The heteroscedastic transform assumes that the classes have Gaussian distributions with identical principal axes but different variances.




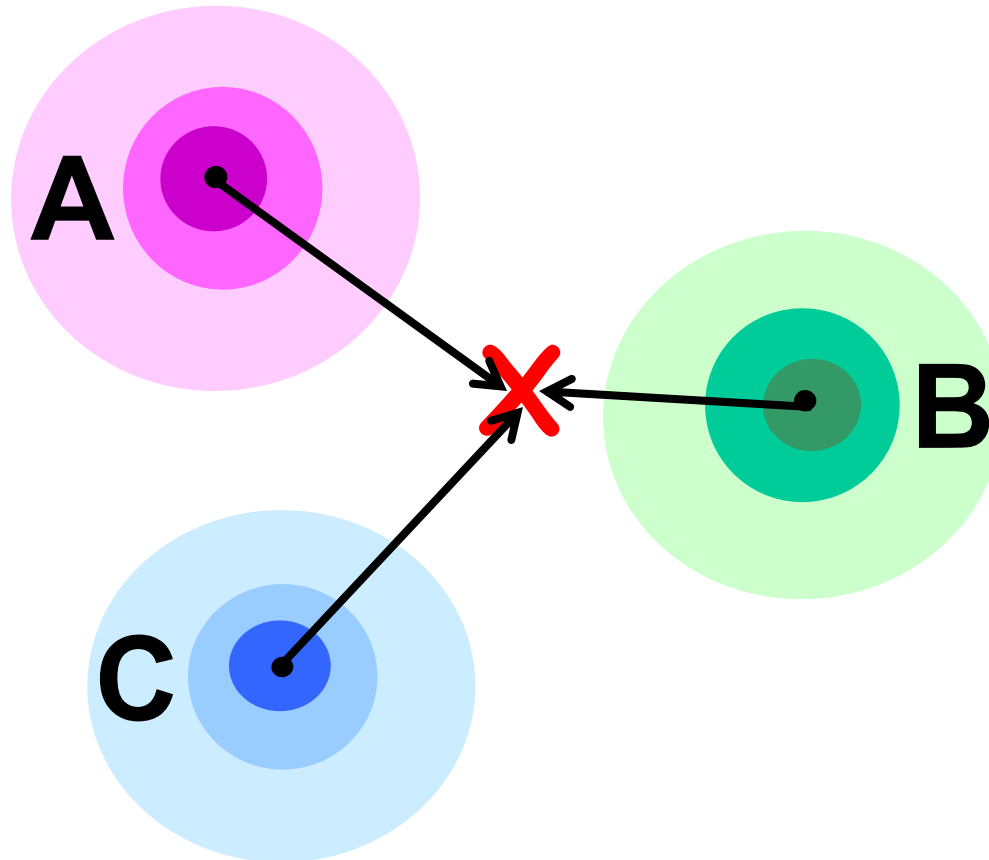
Derivation of LDA:

 To scale the distributions to hyperspheres, we first have to rotate the axes to coincide with the principal axes of the distributions.




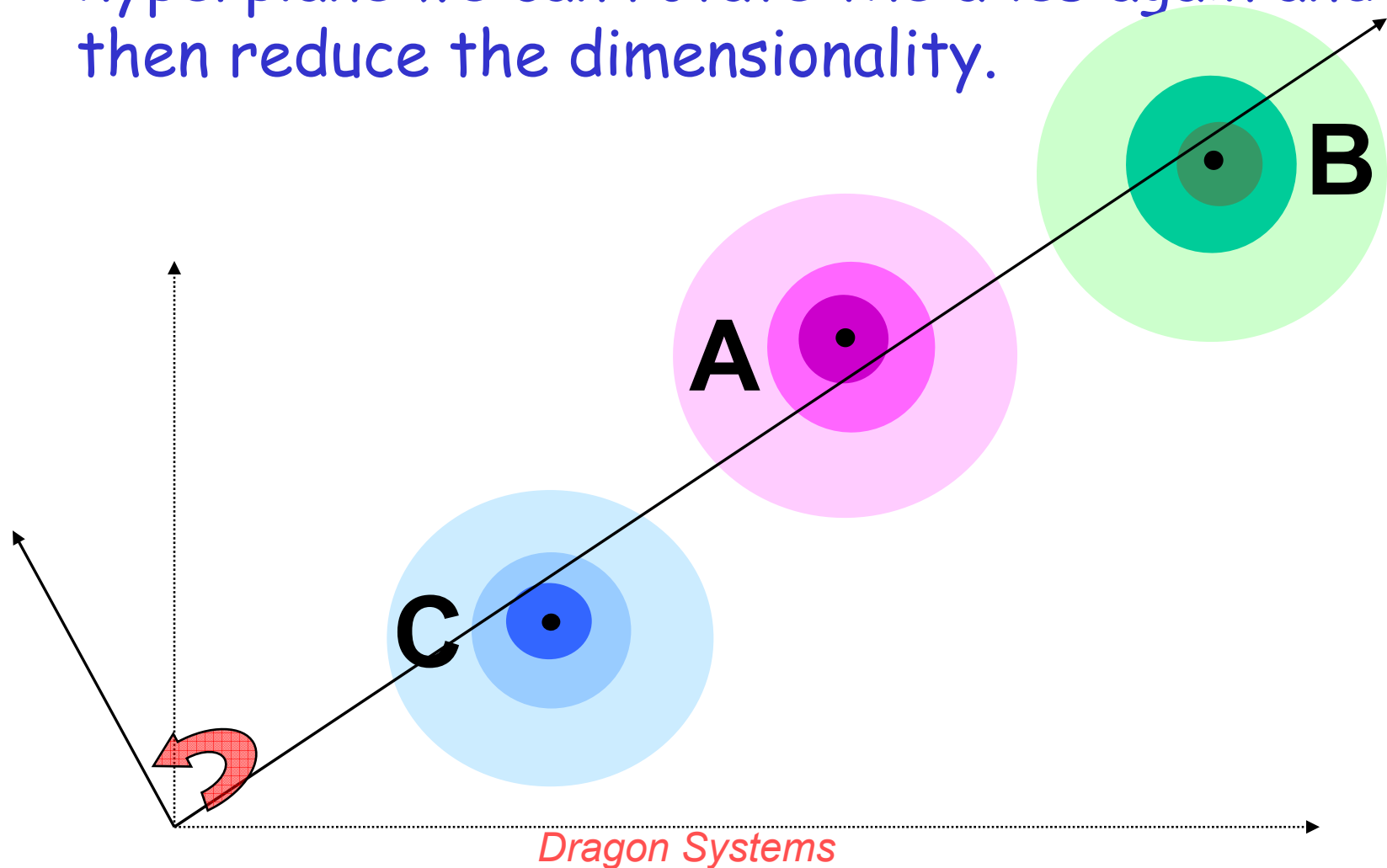
Derivation of LDA:

 The axes can then be scaled to make log probability correspond to Euclidean distance to the centroid




Derivation of LDA:


 If all the centroids lie close to a hyperplane we can rotate the axes again and then reduce the dimensionality.



Dimensionality reduction in LDA

-  Reducing dimensionality reduces storage and computational requirements
 - it sometimes improves recognition accuracy even though the theory says it shouldn't!

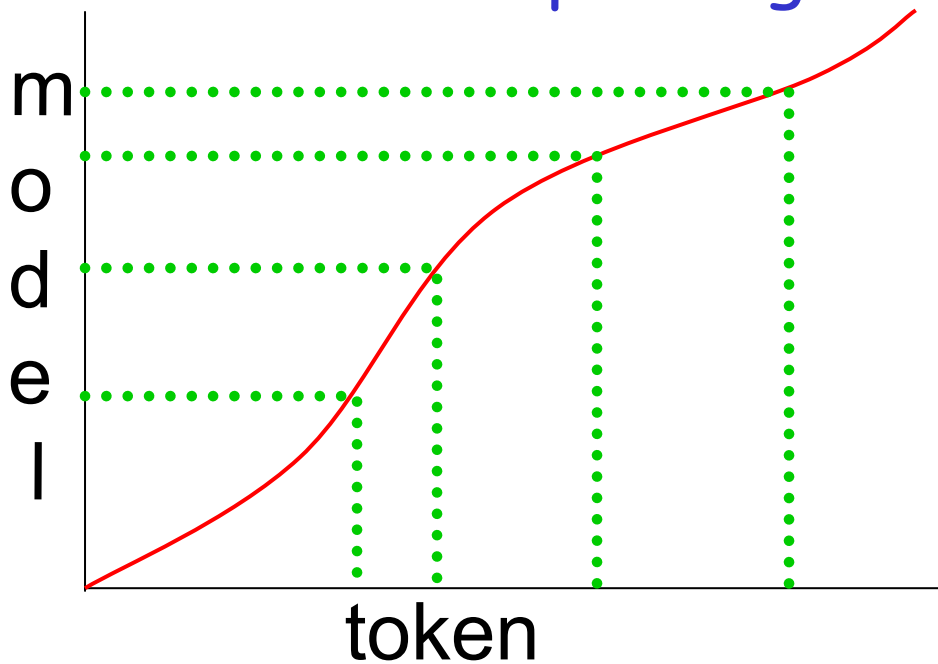
Mathematical Operations in LDA

-  Mathematically, the processes we have just seen correspond to:
1. Diagonalizing the within-class covariance matrix, W
 2. Scaling it to an identity matrix, I .
 3. Diagonalizing the transformed between-class covariance matrix, B .

 This is equivalent to finding the eigenvectors of $W^{-1}B$

In Speech Recognition, You Don't Need to Know the Classes to Derive an LDA Transform!

 You can simply align frames of speech tokens to corresponding models.



Variants on the Standard Analysis: PLP



PLP (*Perceptual Linear Prediction*) (Hermansky)

- spectrum represented on the mel-scale (using Makhoul's "selective LPC")
- loudness weighting applied
- cube-root of spectral energies taken
- all-pole (LPC) fit applied to resulting "perceptual" spectral representation.

Why Does PLP Help?

- 📁 Loudness weighting and cube-root power don't seem to help.
- 📁 Advantage apparently stems from all-pole fit to mel-scale smoothed spectrum (achieved by "selective LPC")
 - is it the mel-scale representation...
 - or the band-pass smoothing that helps?
- 📁 Either way, it's not "perceptual".

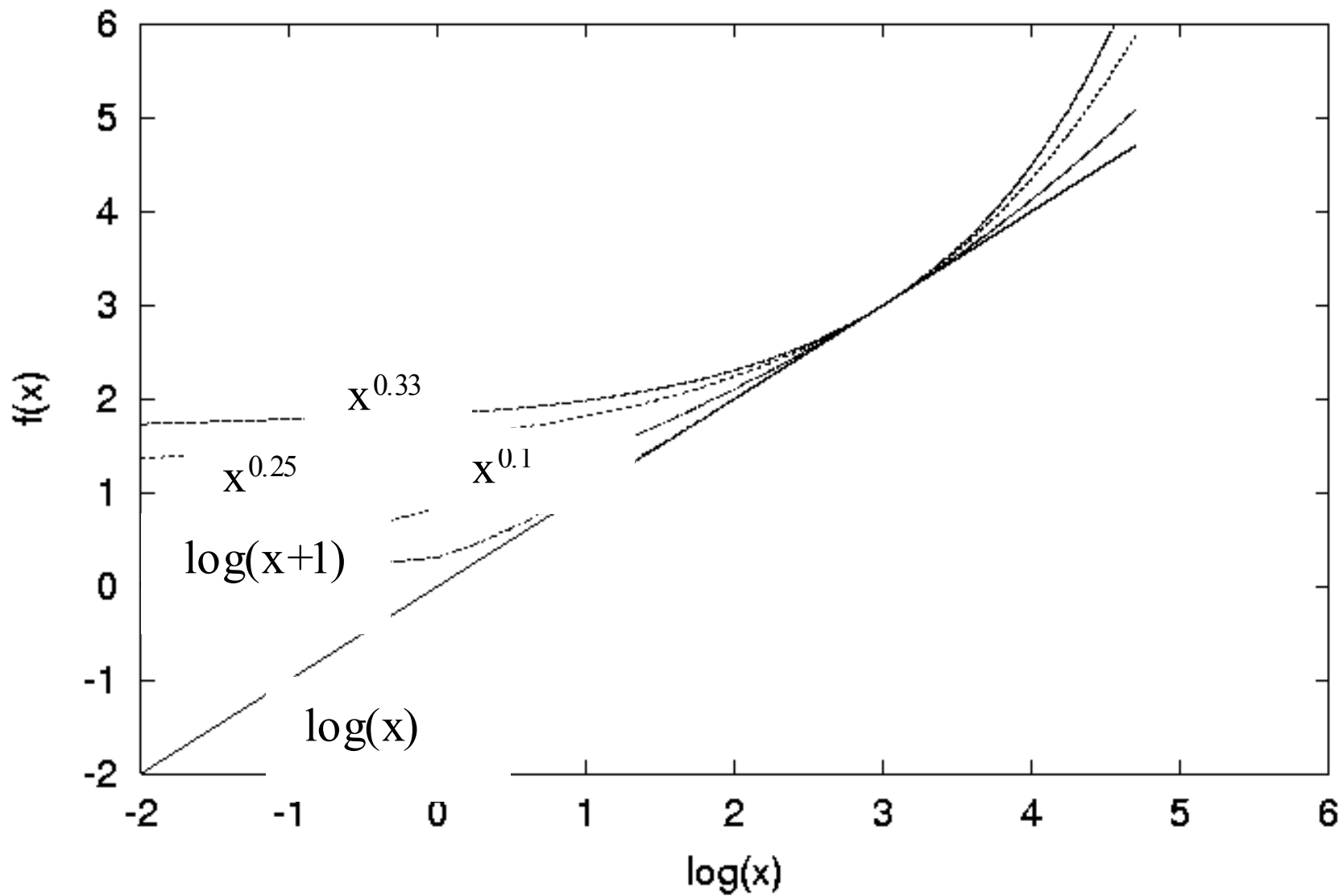
(Cube) Root Power Representations

-  The ear apparently has a cube-root loudness response
-  Evidence that a cube-root power representation is better for ASR than log power in low SNR (Lockwood & Alexandre)


Root Power Representations contd.

- 📖 Actually, root power representations form a continuum, with $\sqrt[n]{y} \rightarrow \sim \log(y)$ as $n \rightarrow \infty$
- 📖 Tian & Viikki found that optimum value of n increased continuously with SNR:
 - $n = .33$ optimal at low SNR;
but $n = \infty$ (*i.e.* log) at high SNR.
- 📖 Lockwood *et al.* found that when the SNR of test and training material is different, the respective n values should be different (higher for noisier material).


Figure 3. Aligned log and power functions plotted on a log scale.



So root power representations resemble logs with “soft” masking

 Given more recent results, cube root doesn't look very special

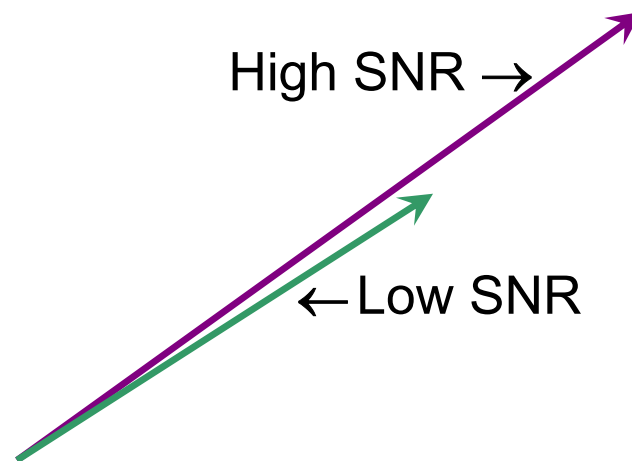
—resemblance to auditory property may be a coincidence.

 Root power results can be explained in terms of signal processing.

Cepstrum Correlation Methods

📖 Originally formulated as a metric (Mansour & Juang)

- a representation can retain the spirit
- cepstrum norm decreases in noise, but direction of the vector is unchanged



Cepstrum Correlation as Spectral Contrast Normalization

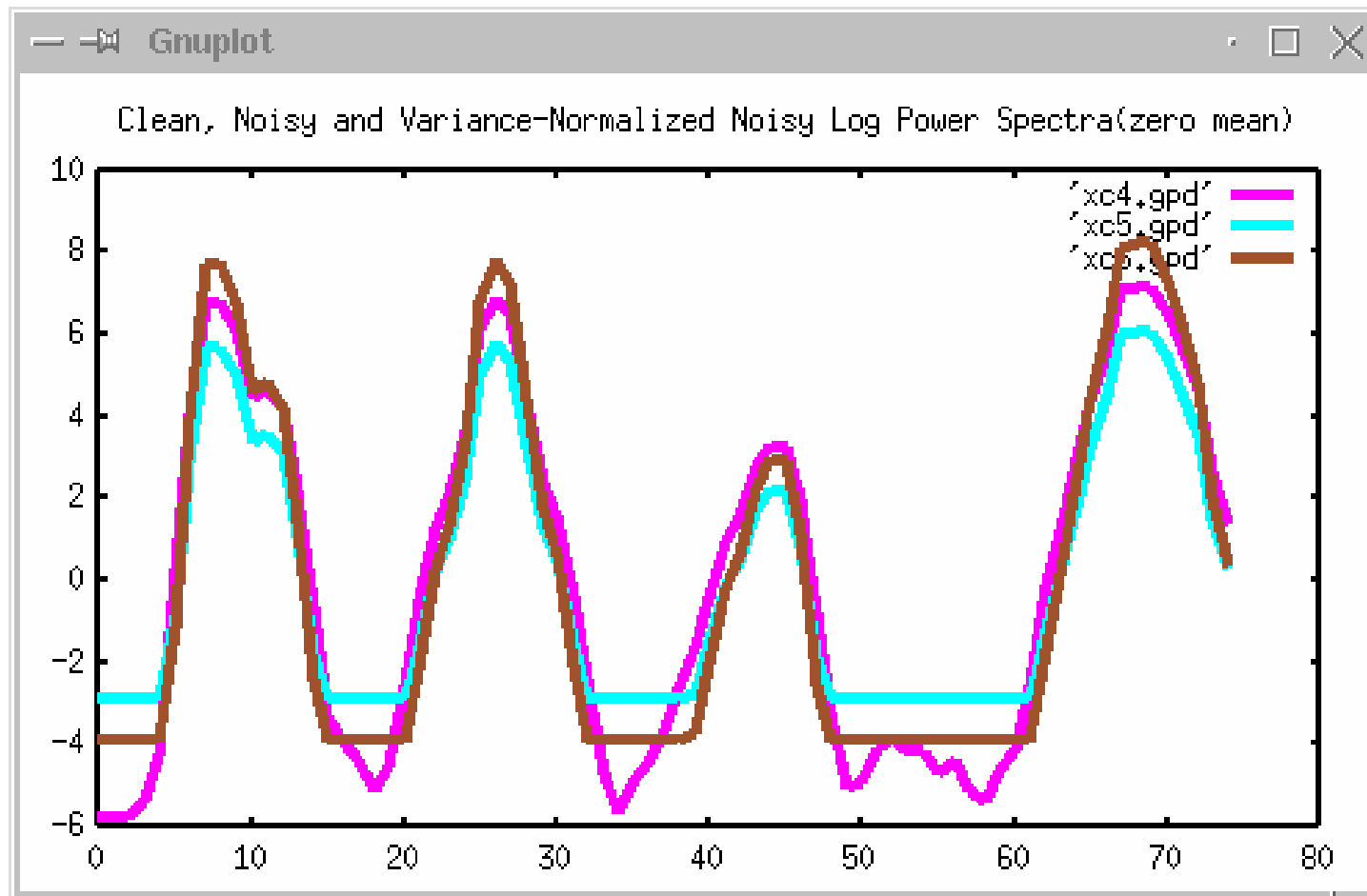
- Smooth noise reduces the dynamic range in the spectrum.
- This is why the cepstrum coefficients are shrunk.
- Normalizing the cepstrum norms preserves their spectral dynamic range (the "spectral contrast").
- But does it work with real-world noise?

Normalizing spectral contrast with artificial data:

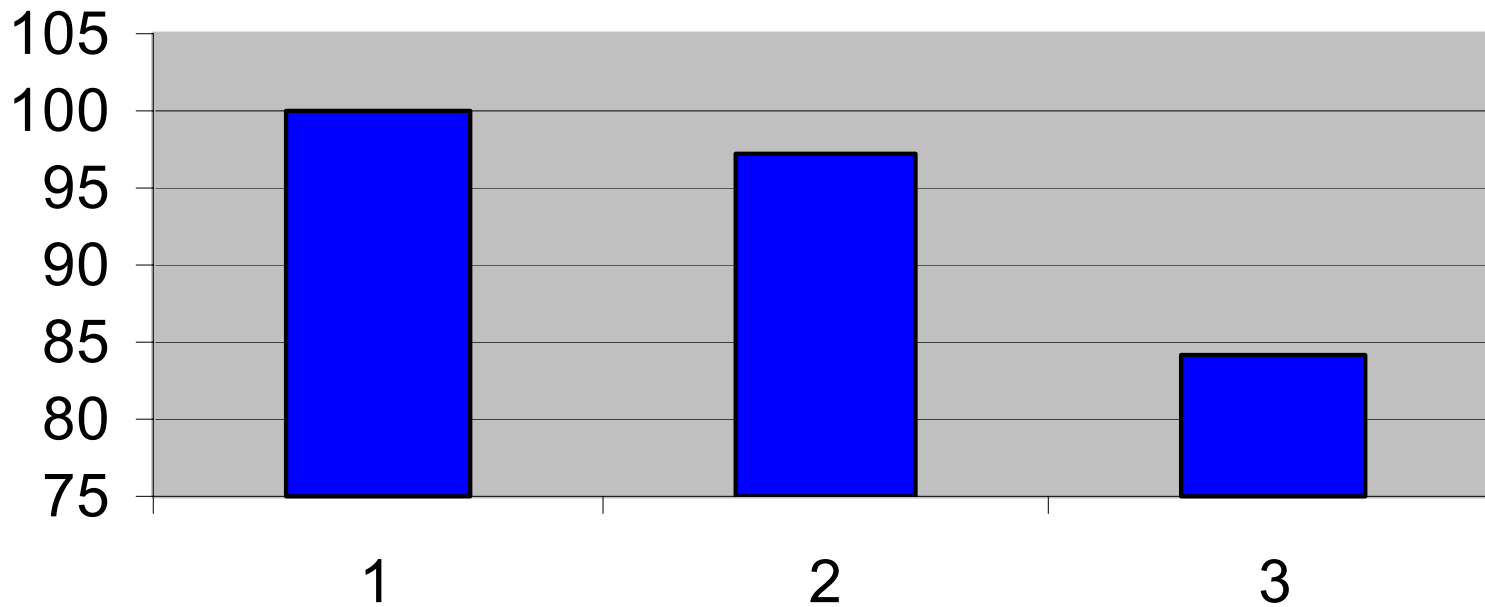
Magenta -- original spectrum

Cyan -- spectrum with noise floor added

Brown -- contrast normalized version



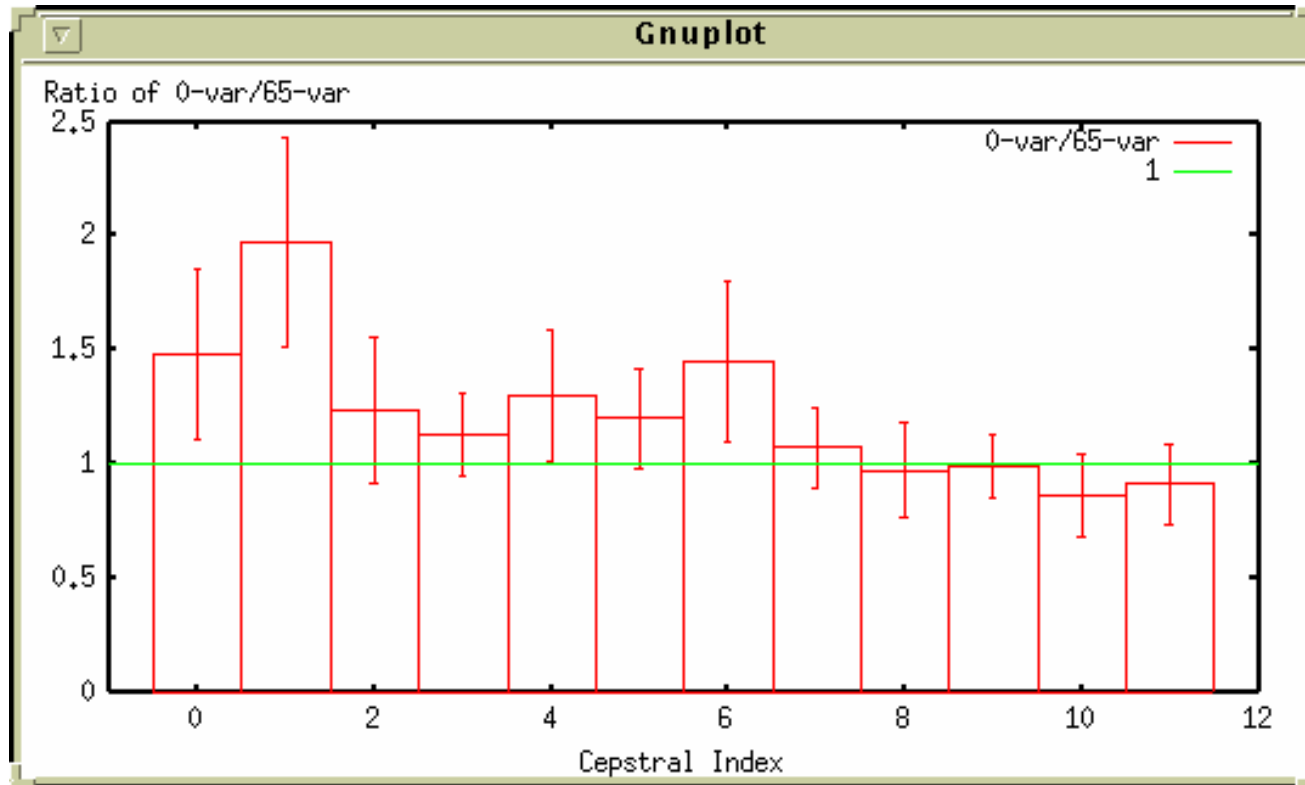
Cepstrum norm reduction with increasing car noise



Average cepstrum norm for:


1 stationary, 2 low-speed, and 3 high-speed conditions

... But in car noise the reduction isn't uniform over the coefs.



Ratios of cepstrum coef. variances from stationary cond. to those at 65 mph, with std. deviations over the 14 speakers.

Summarizing...

 Mel-scale filterbanks, PLP and cube-root power representations don't owe their effectiveness to reproducing human auditory properties.

– and LDA and cepstrum correlation methods don't even claim to.

 In fact, with a conservative criterion for progress, there's no evidence that copying auditory properties has helped ASR.

Where does the promise of future progress lie?

1 Modelling human speech perception

— *e.g.* human ability to perceive speech in noise largely independently of F_0 value.

- human error rates on noisy female speech are about 25% higher than on male speech at same SNR (Steeneken)
- ASR error rates on noisy female speech with standard acoustic analysis are ~3 times higher!

2 Modelling human speech production

— recent promising results from Li Deng