A Phonologically-Calibrated Acoustic Dissimilarity Measure

Greg Kochanski, Ladan Baghai-Ravary, and John Coleman
University of Oxford Phonetics Laboratory

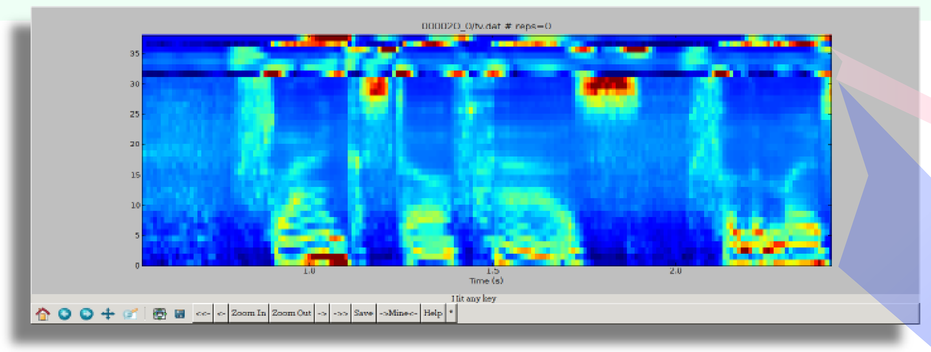
We thank the ESRC for grant number RES-062-23-1323.

What is it?
* A way of measuring differences between utterances.
  * Large differences imply phonologically different
  * Small differences imply phonologically identical
* Can resolve a fraction of a minimal pair distance..

Why use it?
* You want to talk about "large" or "small" differences.
* Measure fine phonetic detail.
* Measure coarticulation.
* Compare synthesized speech to natural speech.

What does it tell us?
* What aspects of the signal carry phonological distinctions.

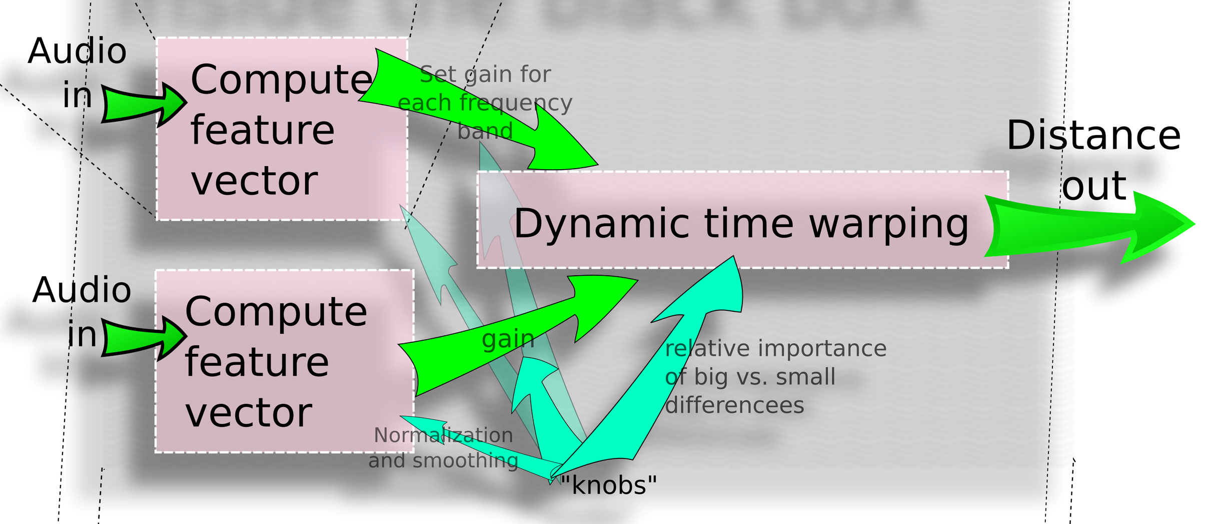
The data for were a set of 2578 phonetically rich English utterances that ten subjects, aged 19-62, read from randomised lists. The corpus included a total of 828 different texts with a mean sentence length of 6.5 words.



the feature vector

Spectrum + first derivative
* Spectrum = 4th order monotone filter bank
  * ~1 erb frequency bins
  * 20 ms time window
  * cube-root of power
* Spectrum is normalized
  * amplitude is relatively umimportant
  * spectrum is divided by:
    * local spectrum + 0.05*utterance average
* First derivative
  * 5 broad bands
  * 40 ms smoothing
  * looks at spectral change over 40 ms interval

Inside the black box



Training data

Pair the utterances: either different texts, or the same text.

Black box distance estimator: Give it data and set its "knobs", and it returns a distance

The "knobs" define which acoustic properties are an important part of the distance.

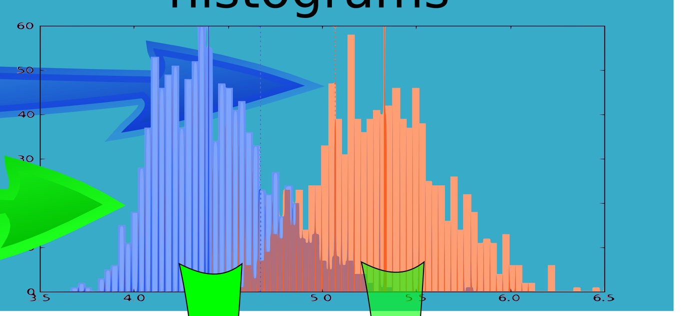
Adjusts "knobs"

Bootstrap Markov-Chain Monte-Carlo optimization code.

Different texts

Same texts

Accumulate distances for utterances into two histograms



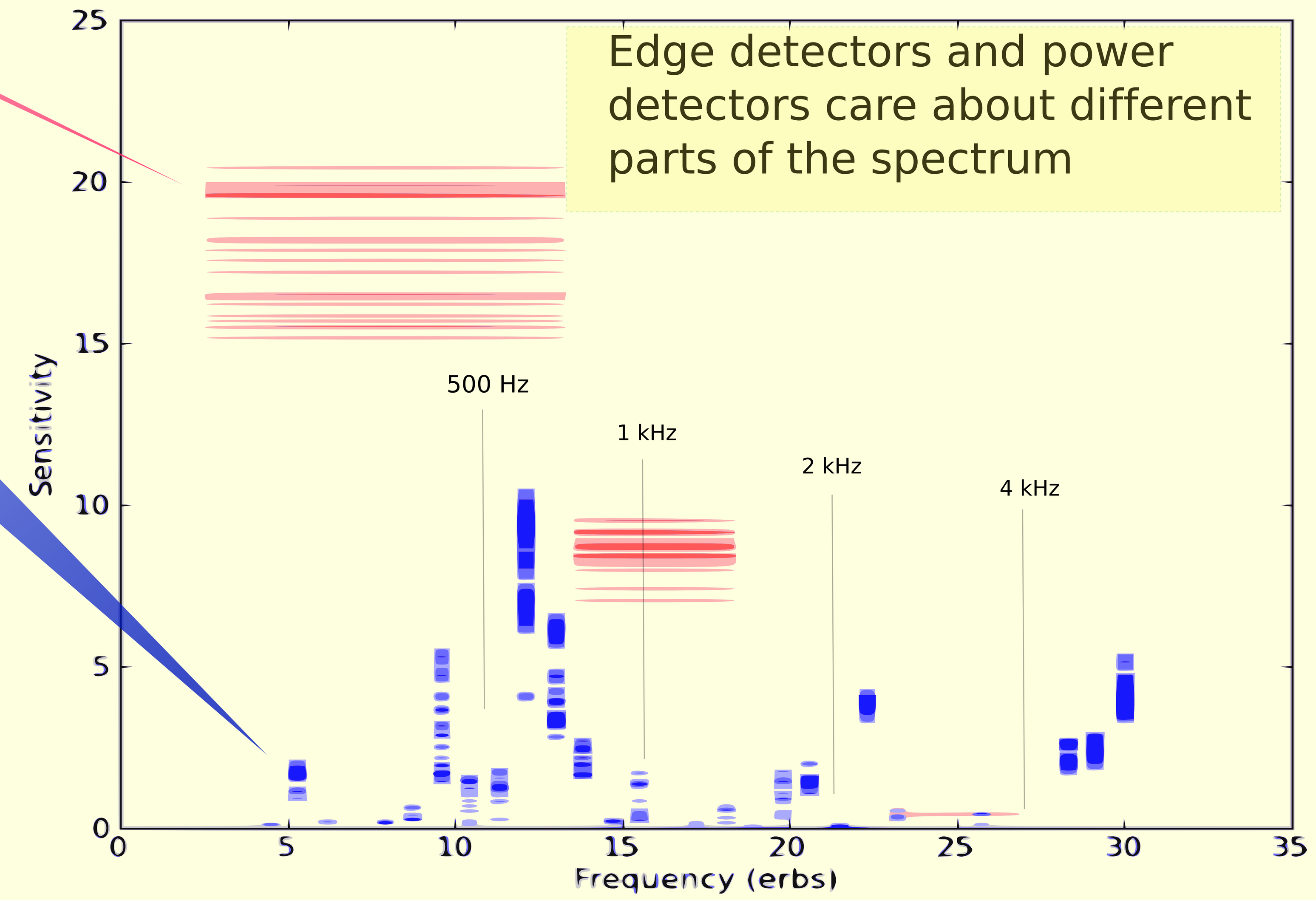
Same texts

Different texts

Compute how well separated the two histograms are: t-statistic

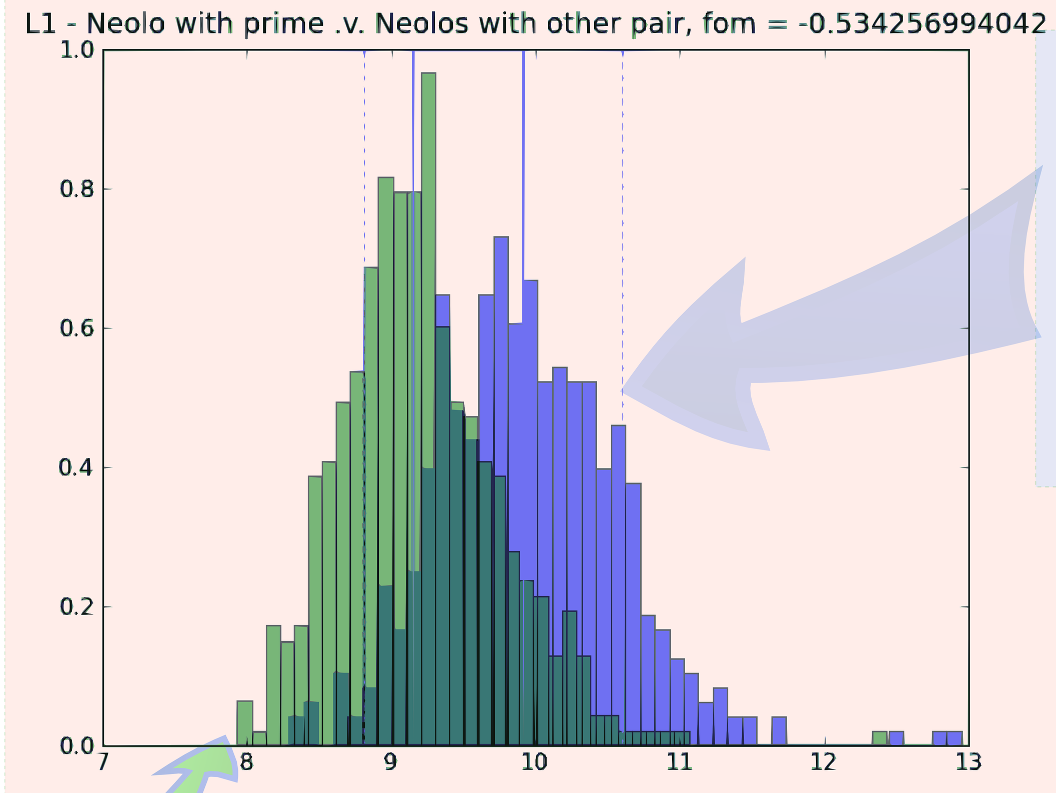
Did the last change improve things?

Interpreting the optimized distance metric



* The most important component is the low-frequency derivative (edge detector).
* 450-700 Hz is a very important region
  * as spectrum and for edge detector
* 4500-6200 Hz important
* 1700-2500 Hz

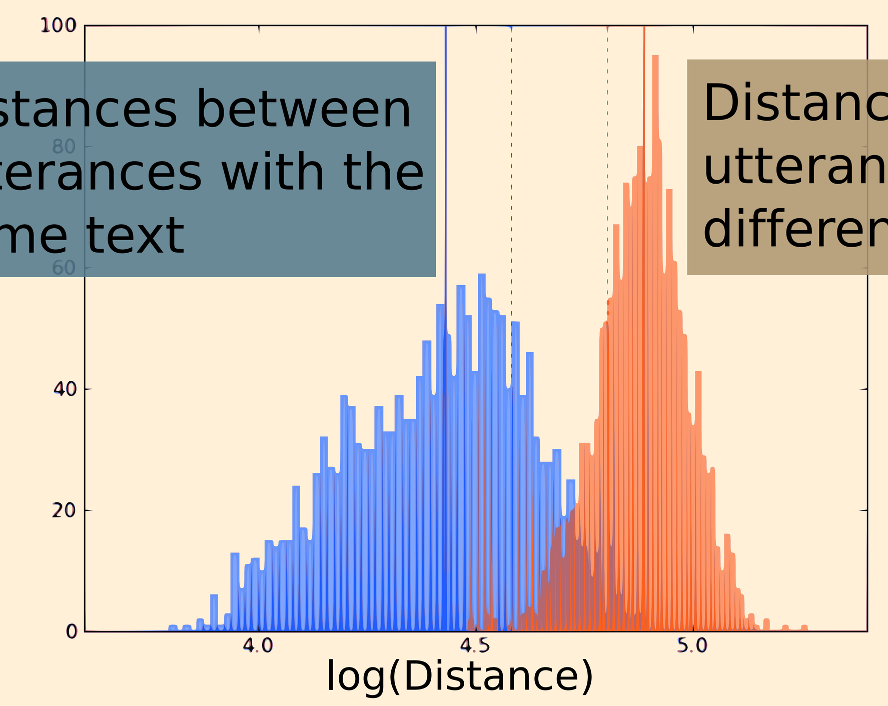
Single-vowel differences



Histograms of difference due to a change in a single vowel. The difference is typically 1 or 2 phonological features, and the region is typically 2 phones long.

Acoustic differences between the performance of phonologically identical regions

Optimized Performance



Distances between utterances with the same text

Distances between utterances with different texts

Separation: t-statistic = 2.7 versus ~1 for unoptimized distance metric.

* Quantitative measurements of phonological similarity and difference are possible.

* These techniques can resolve minimal pairs, and may be able to measure fine phonetic detail.

* Can be customized to a particular dialect, language, or reording conditions.

* The optimization procedure can be applied to other distance metrics (we have achieved substantial improvements in Itakura-Saito divergence with similar techniques).