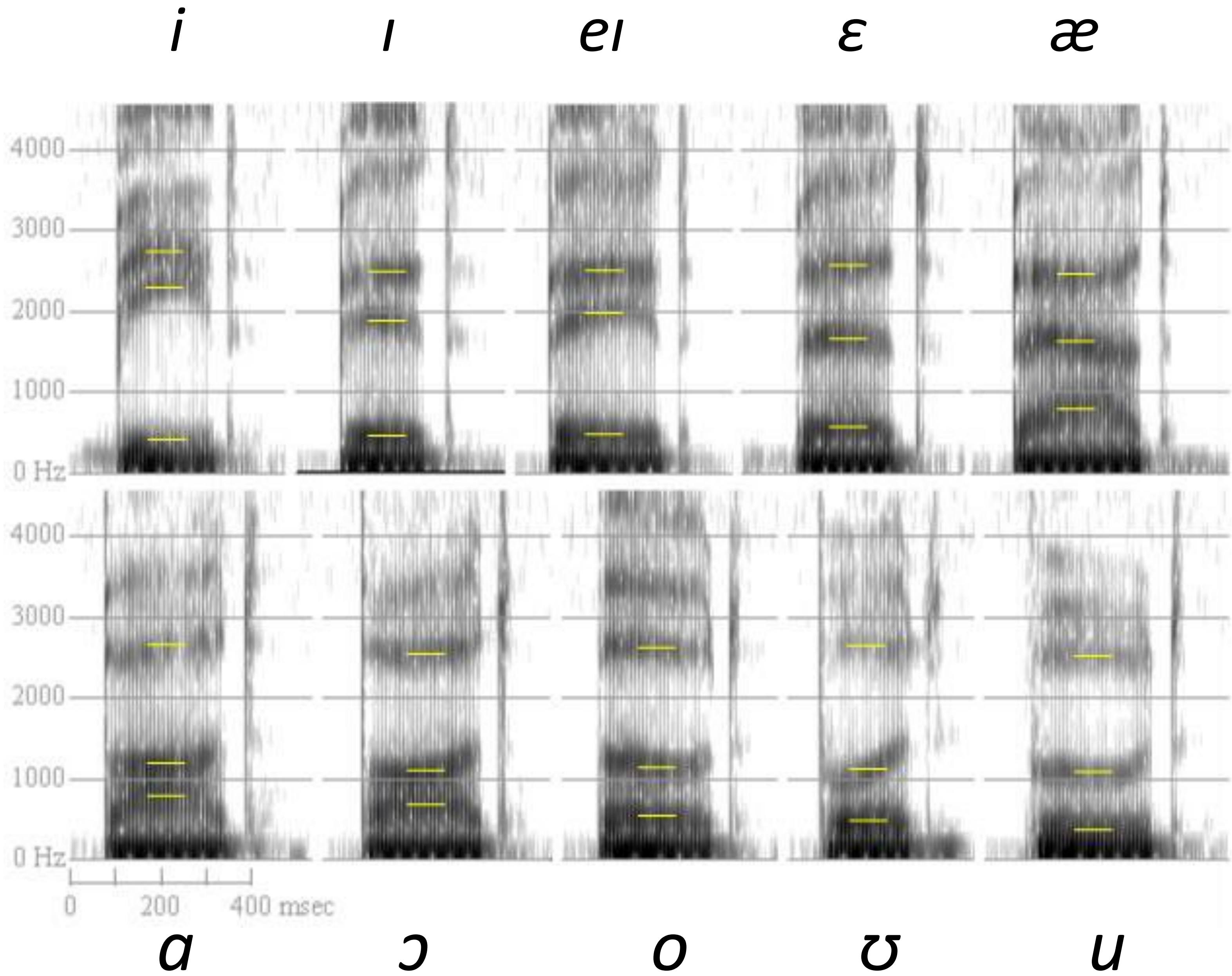
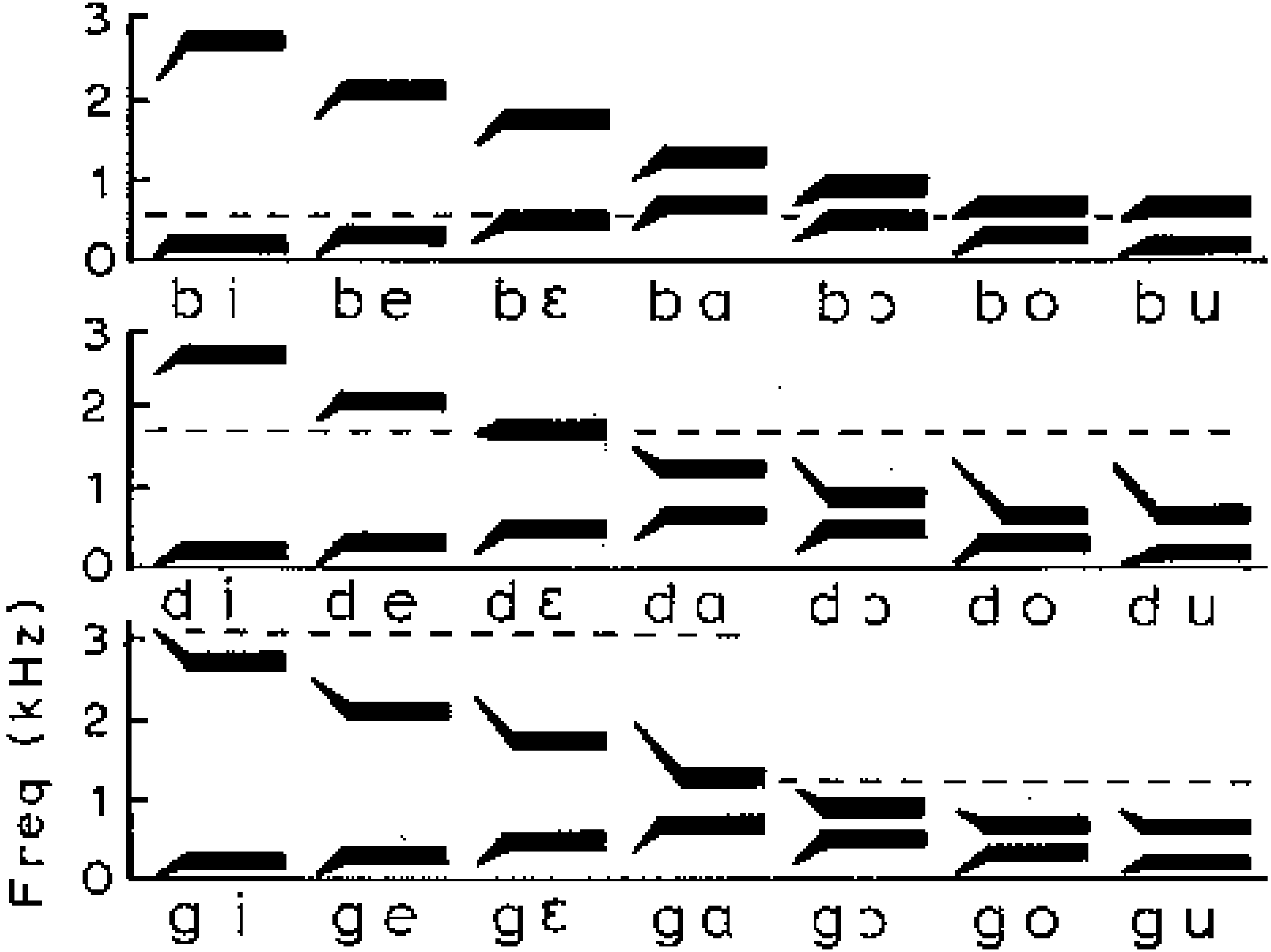
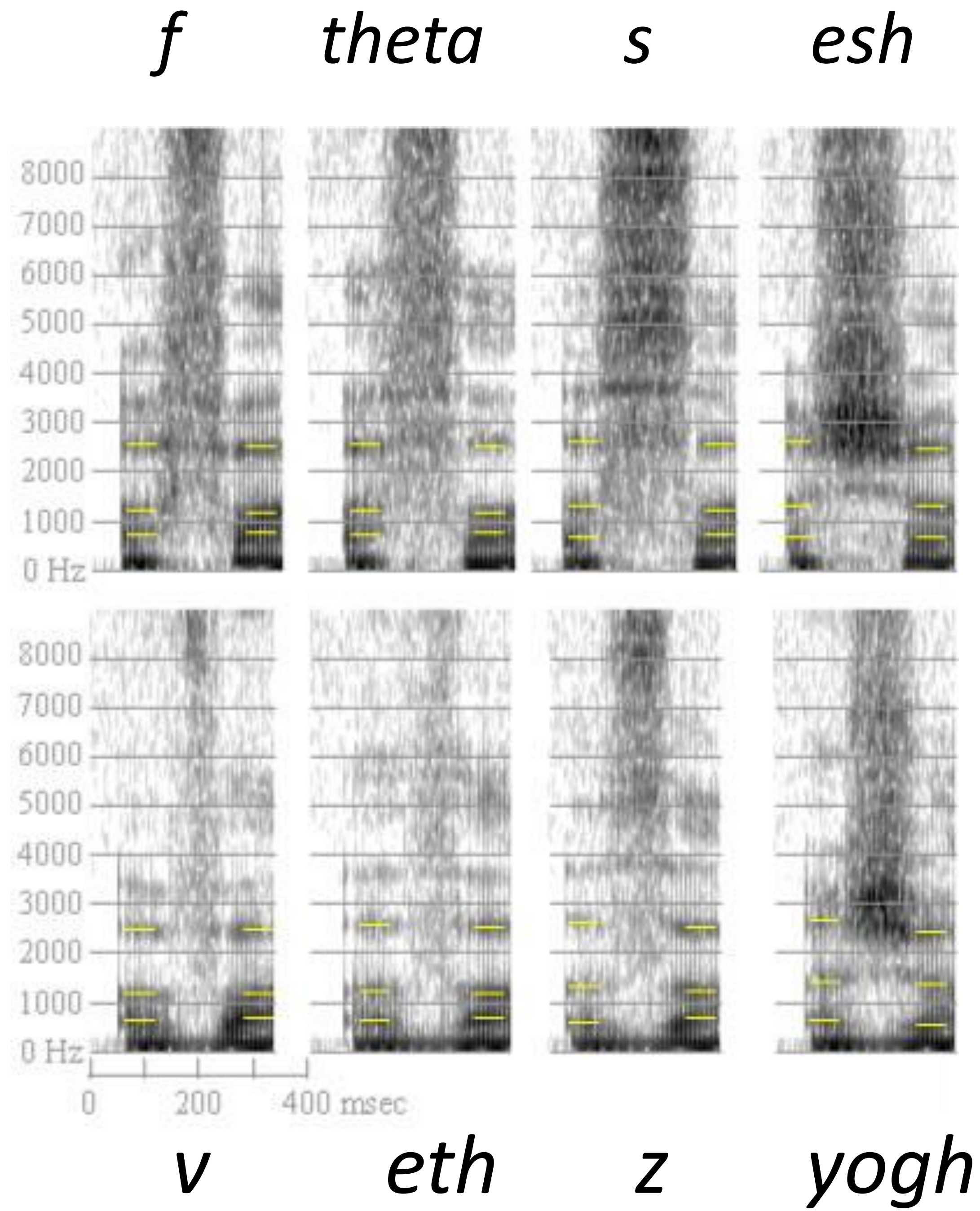
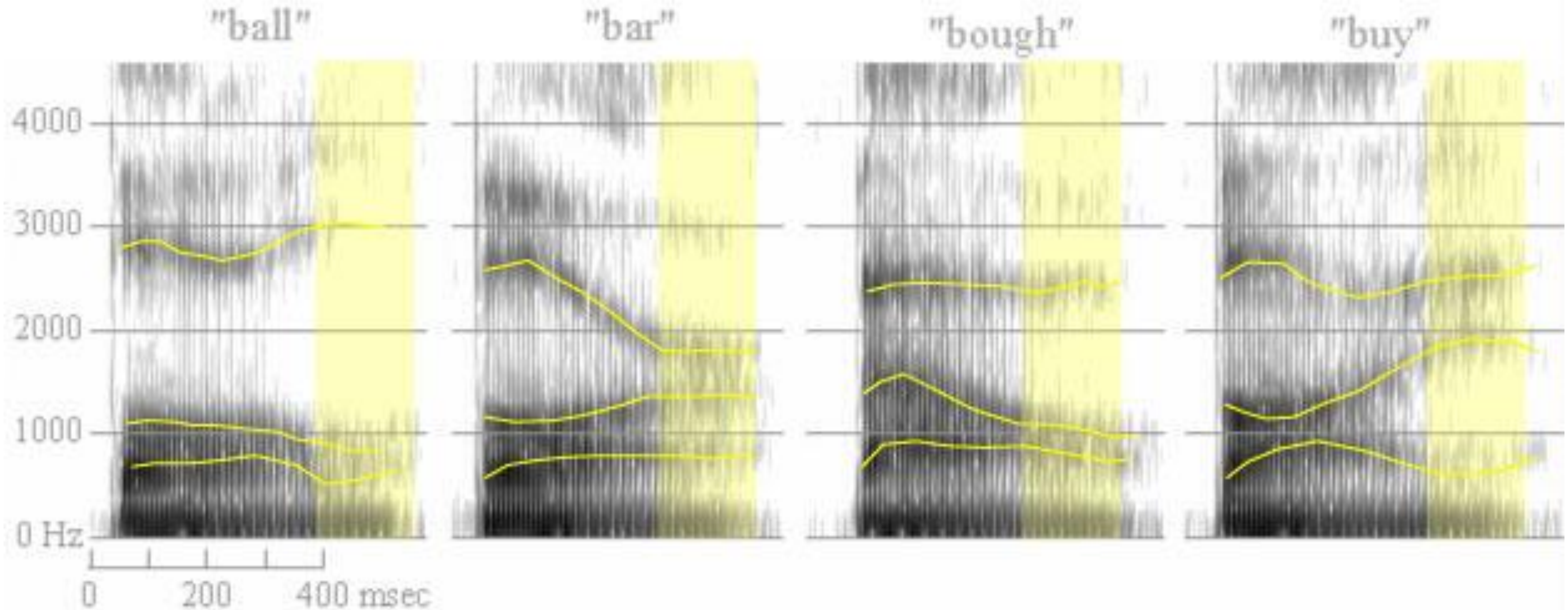


i
ɪ
eɪ
ɛ
æ
4000
3000
2000
1000
0 Hz
0
200
400 msec
ɑ
ɔ
o
ʊ
u



Freq (kHz)
0
1
2
3
bi
be
bɛ
ba
bɔ
bo
bu
di
de
dɛ
da
dɔ
do
du
gi
ge
gɛ
ga
gɔ
go
gu

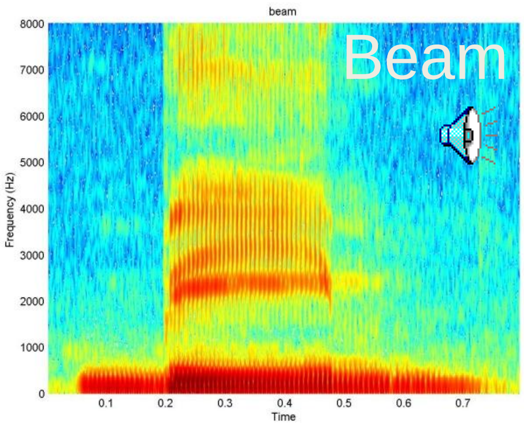

f
theta
s
esh
8000
7000
6000
5000
4000
3000
2000
1000
0 Hz
0
200
400 msec
v
eth
z
yogh

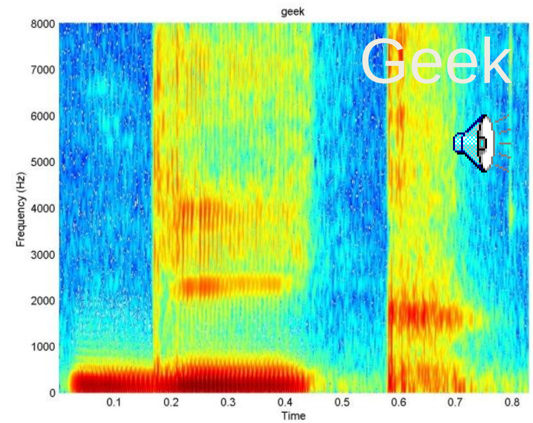

"ball"
"bar"
"bough"
"buy"
4000
3000
2000
1000
0 Hz
0
200
400 msec

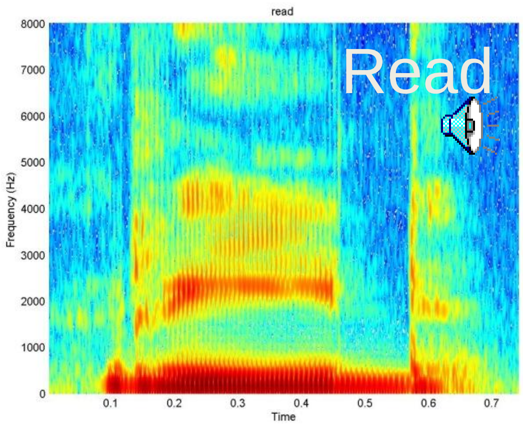
Patterns in spectrogram
Signature of phonemes in spectrogram
Typical signatures
Changes in variety of conditions – speaker, noise, context, language, emotion

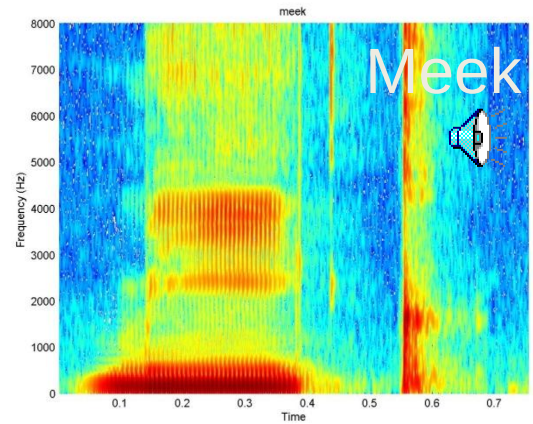
Spectrogram reading

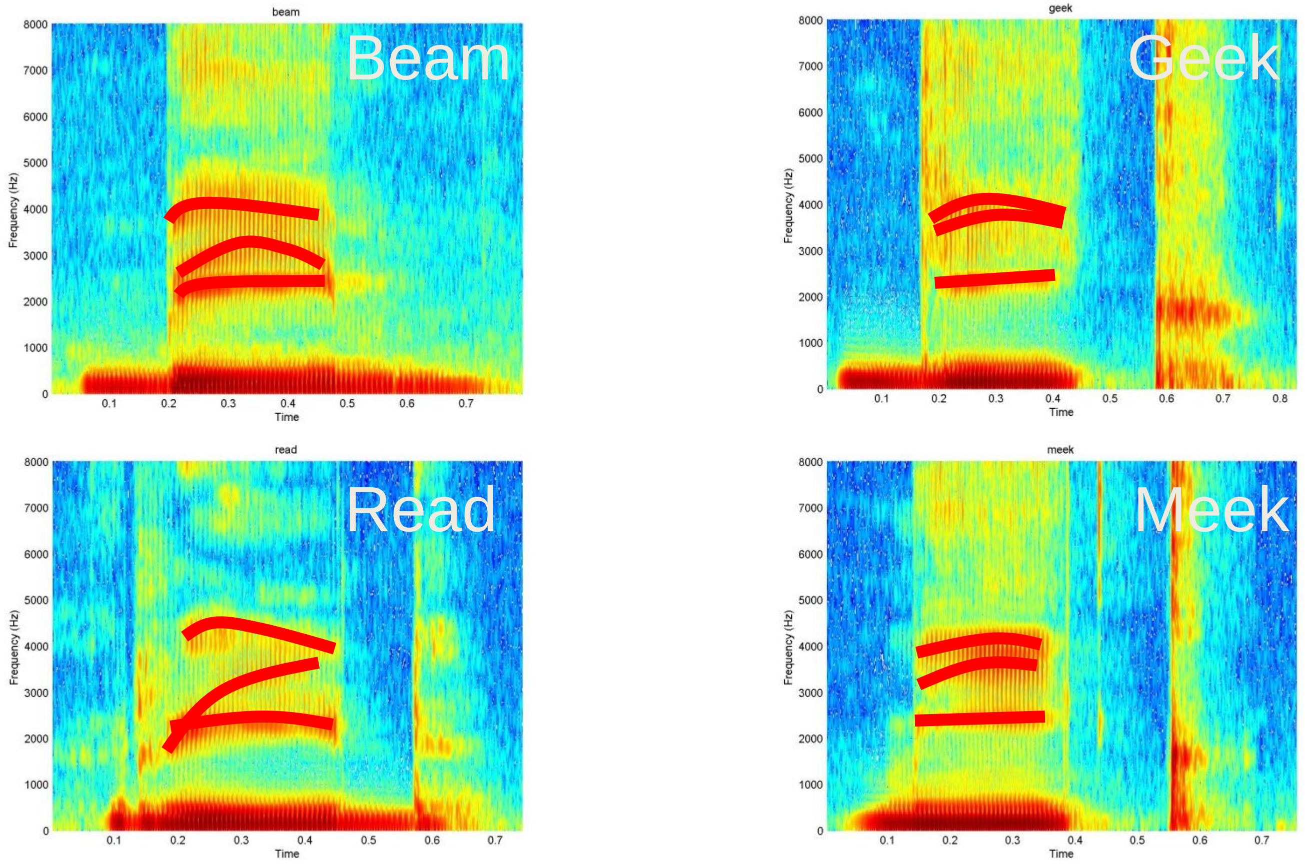
http://www.cslu.ogi.edu/tutordemos/SpectrogramReading/spectrogram_reading.html


beam
Beam
Frequency (Hz)
8000
7000
6000
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time

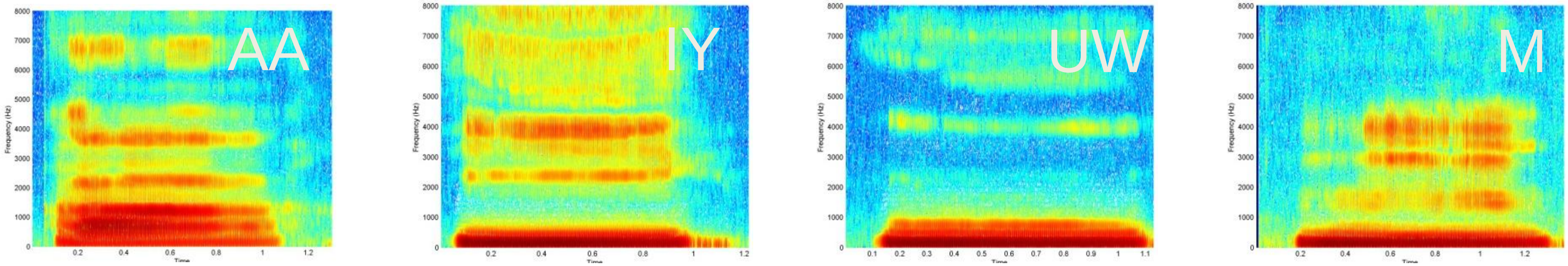

geek
Geek
Frequency (Hz)
8000
7000
6000
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Time


read
Read
Frequency (Hz)
8000
7000
6000
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time

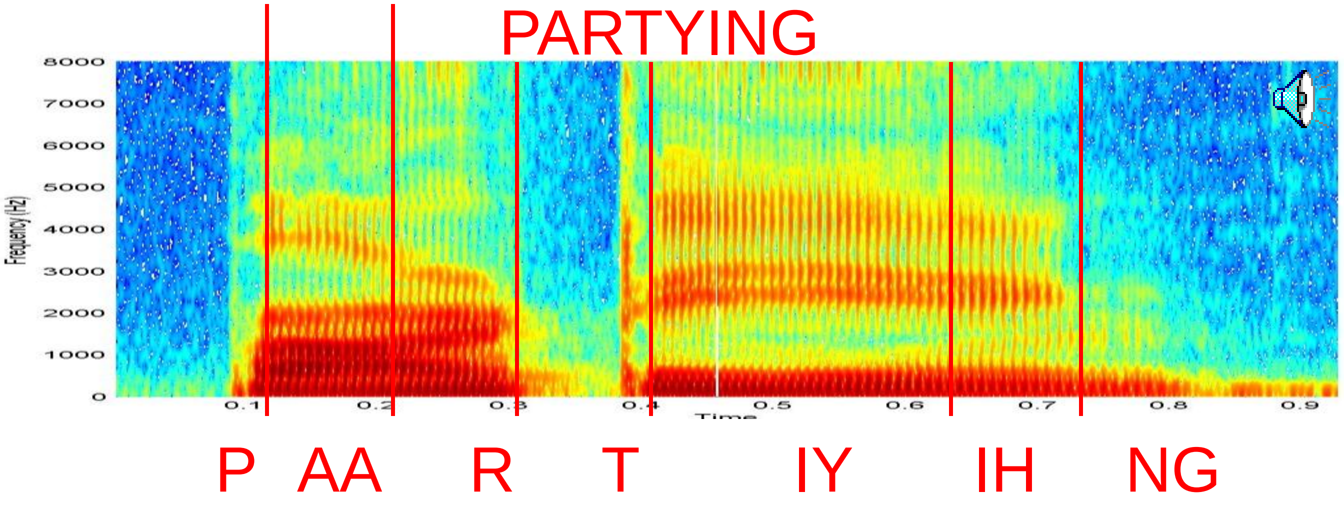

meek
Meek
Frequency (Hz)
8000
7000
6000
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Time


beam
Beam
geek
Geek
read
Read
meek
Meek
Frequency (Hz)
Time

- Every phoneme has a locus
  - The spectral shape that would be observed if the phoneme were uttered in isolation, for a long time



  - In continuous speech, the spectrum attempts to arrive at locus of the current sound



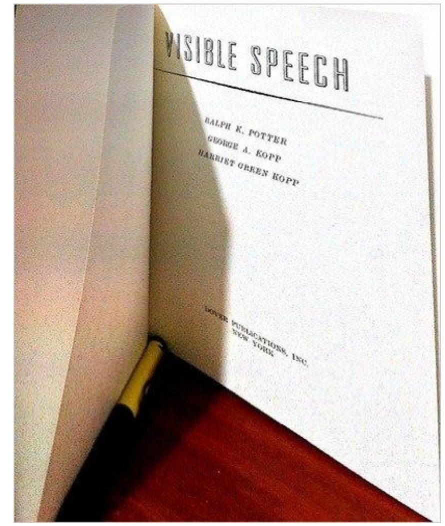
COARTICULATION AND THE LOCUS THEORY

Pierre Delattre

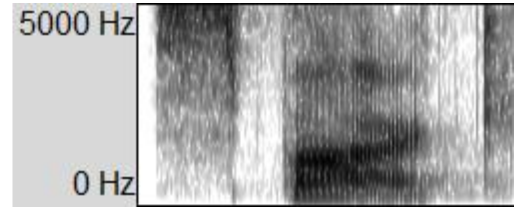
Studia Linguistica
Volume 23, Issue 1, pages 1–26, June 1969

The visible speech

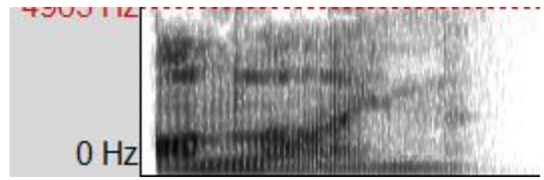
by Ralph Kimball Potter , George A. Kopp , Harriet Green Kopp



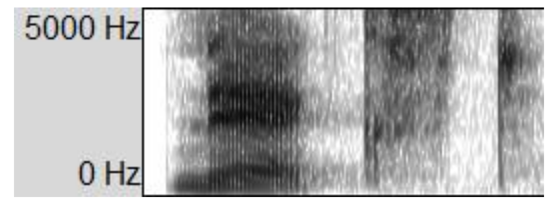
Online



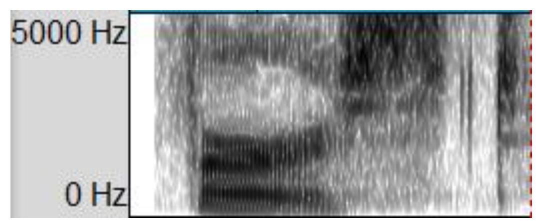
Next



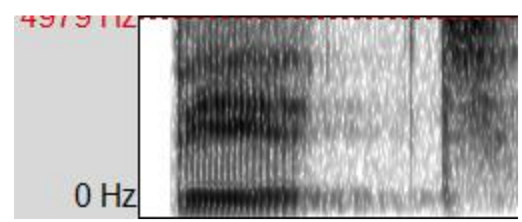
First



Bit



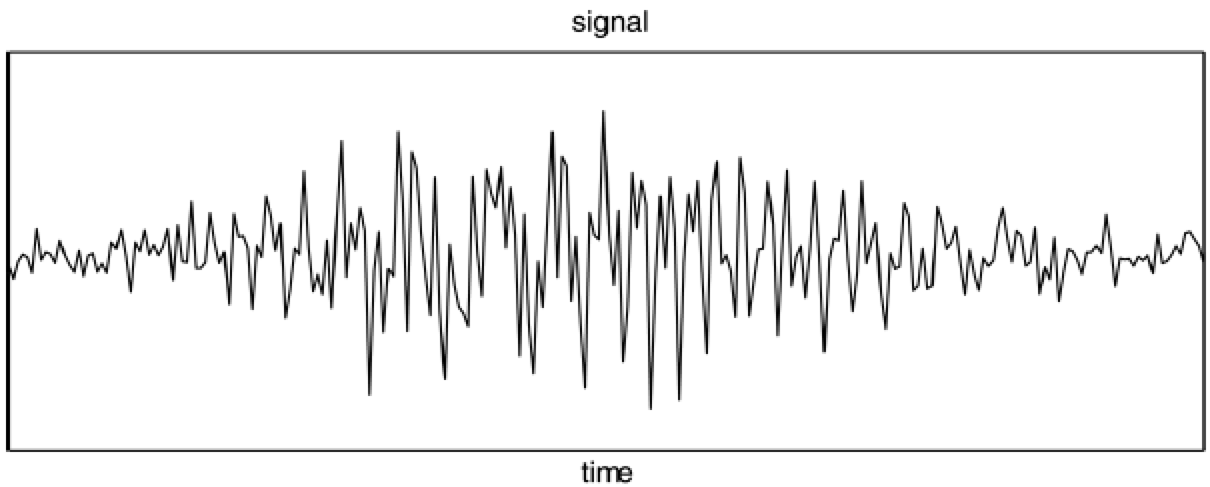
spark

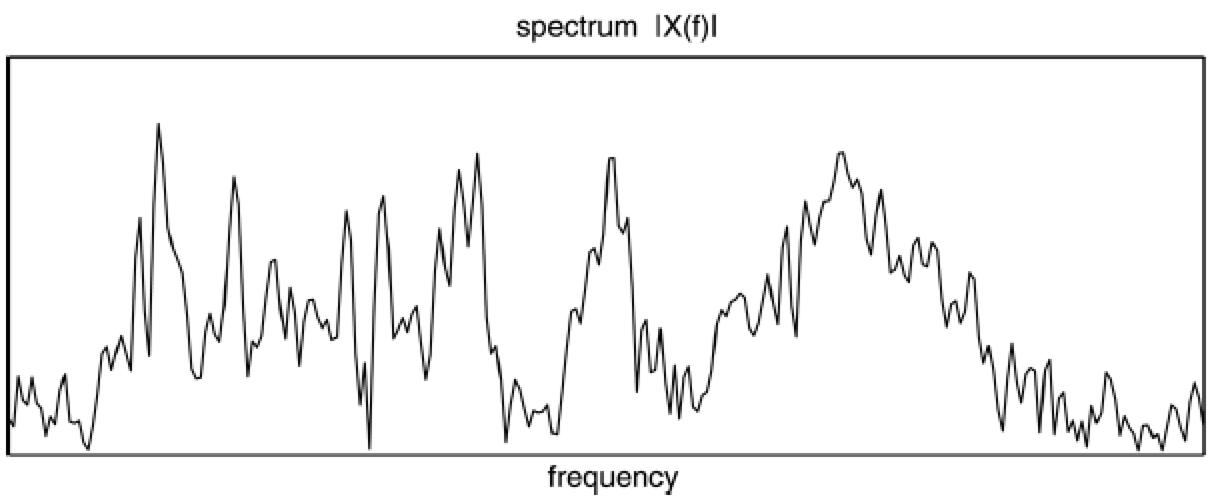


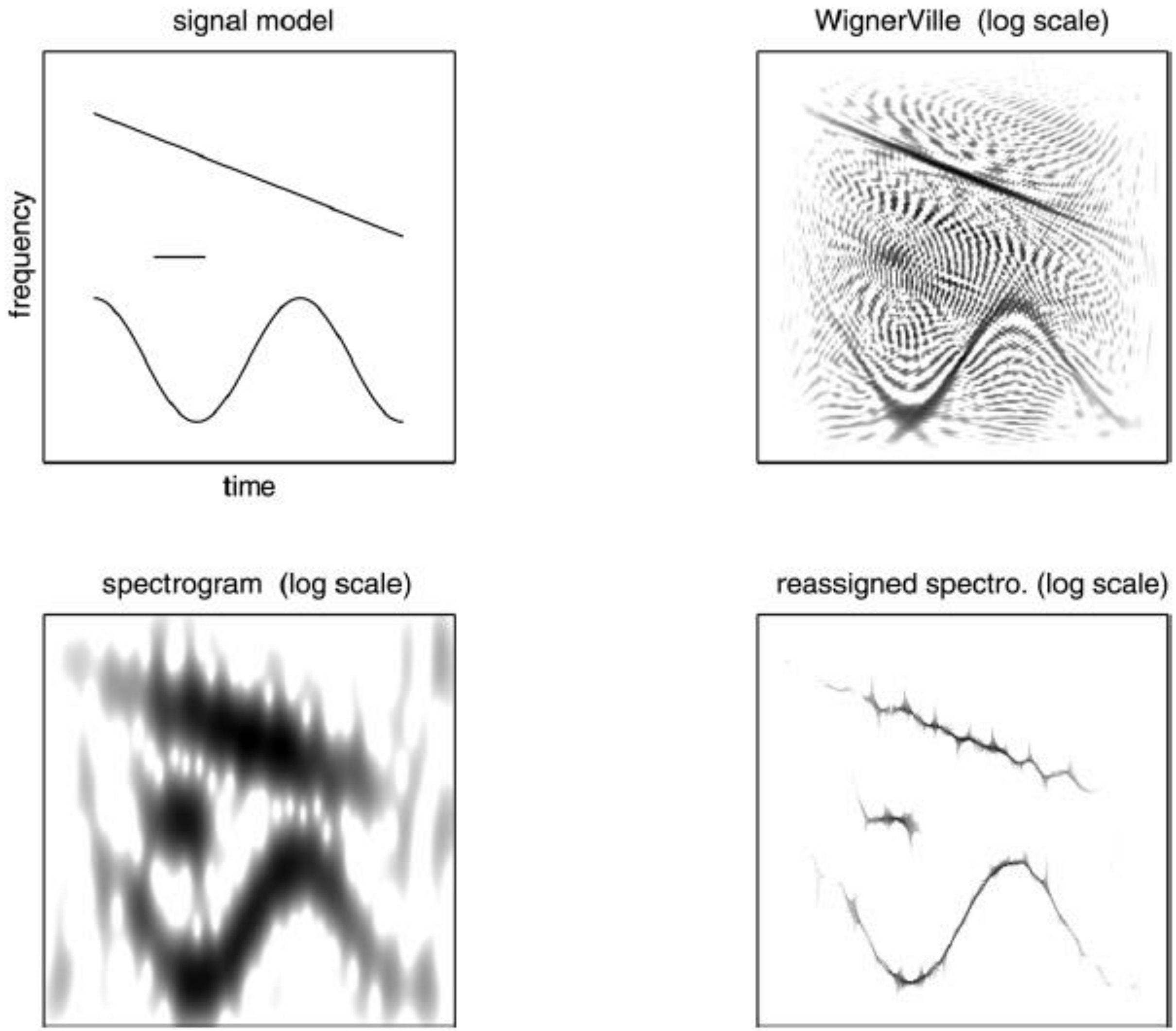
weighted overlap-add

[8] Portnoff, M.R. 1980. "Time-Frequency Representation of Digital Signals and Systems Based on Short Time Fourier Analysis," IEEE Trans on ASSP, vol 28(1), pp. 55-69.

[9] Crochiere, R.E. 1980. "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis," IEEE Trans on ASSP, vol 28(1), pp. 99-102.


signal
time


spectrum |X(f)|
frequency



signal model
frequency
time
WignerVille (log scale)
spectrogram (log scale)
reassigned spectro. (log scale)

WVD

$$W_x(t,\omega) = \int_{-\infty}^{\infty} x(t+s/2)\, x^*(t-s/2)\, e^{-jsw}\, ds.$$

Spectrogram (Mag. sqd. of STFT)

$$S_x^h(t,\omega) = \iint_{-\infty}^{+\infty} W_x(t,\omega)\, W_h(s-t, \xi-\omega)\, \frac{ds\, d\xi}{2\pi}$$

$\varphi(t,\omega)$ – phase of STFT

Reassignment

$$\hat{t}_x(t,\omega) \triangleq \frac{1}{S_x^h(t,\omega)} \iint_{-\infty}^{+\infty} s\, W_x(s,\xi)\, W_h(s-t, \xi-\omega)\, \frac{ds\, d\xi}{2\pi}$$

$$\hat{\omega}_x(t,\omega) \triangleq \frac{1}{S_x^h(t,\omega)} \iint_{-\infty}^{+\infty} \xi\, W_x(s,\xi)\, W_h(s-t, \xi-\omega)\, \frac{ds\, d\xi}{2\pi}$$

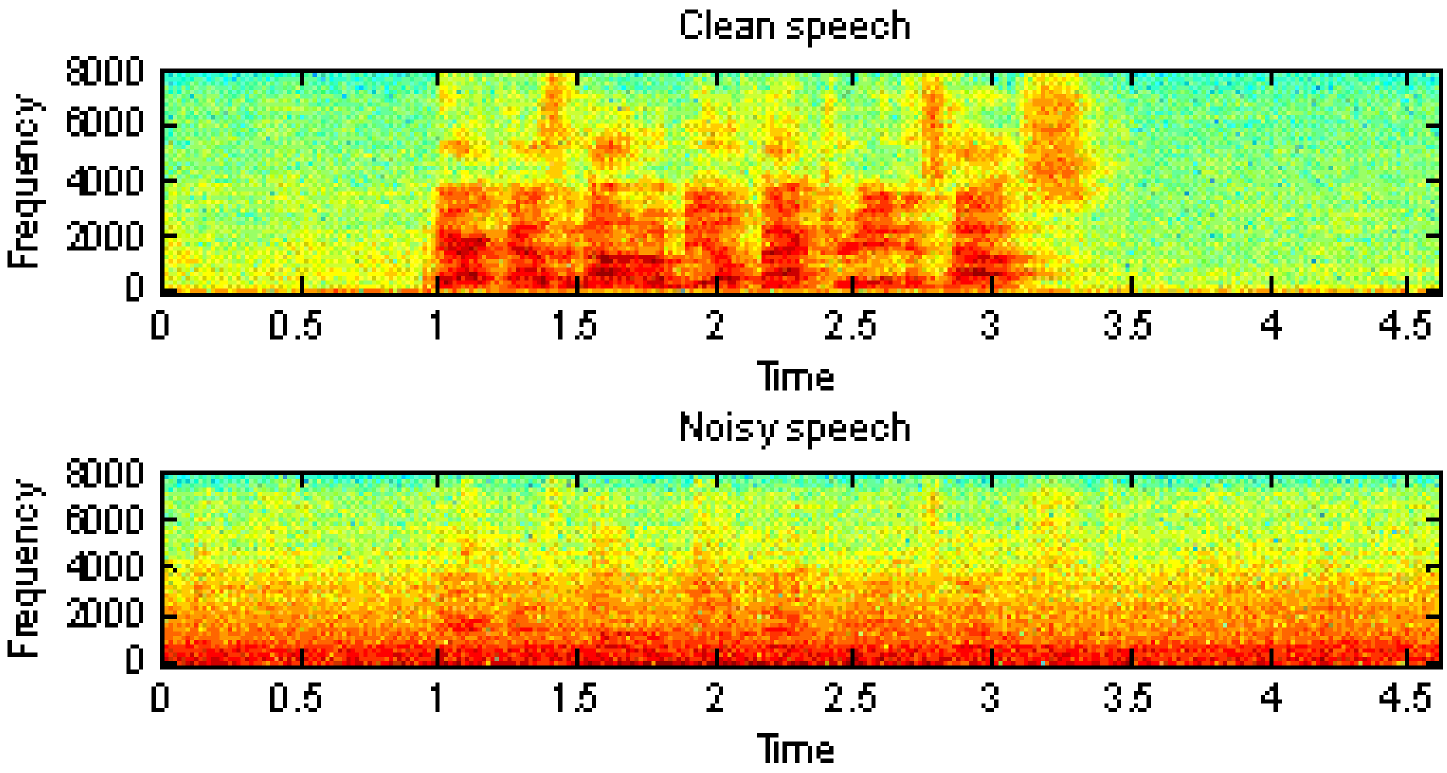
Reassigned Spectrogram

$$\tilde{S}_x^h(t,\omega) \triangleq \iint_{-\infty}^{\infty} S_x^h(t,\omega)\, \delta\left(t-\hat{t}_x(t,\omega), \omega-\hat{\omega}_x(s,\xi)\right) \frac{ds\, d\xi}{2\pi}$$

It can be shown

$$\hat{t}_x(t,\omega) = \frac{t}{2} - \frac{\partial}{\partial\omega}\varphi(t,\omega)$$

$$\hat{\omega}_x(t,\omega) = \frac{\omega}{2} + \frac{\partial}{\partial t}\varphi(t,\omega)$$


Clean speech
Frequency
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Time
Noisy speech
Frequency
8000
6000
4000
2000
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
Time

Morphological Processing of Spectrograms for Speech Enhancement

Joyner Cadore, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno

Universidad Carlos III de Madrid, Escuela Politécnica Superior,
Avda. de la Universidad 30, 28911 Madrid, Spain
{jcadore,gallardo,carmen}@tsc.uc3m.es
http://gpm.tsc.uc3m.es/

Abstract. In this paper a method to remove noise in speech signals improving the quality from the perceptual point of view is presented. It combines spectral subtraction and two dimensional non-linear filtering techniques most usually employed for image processing. In particular, morphological operations like erosion and dilation are applied to a noisy speech spectrogram that has been previously enhanced by a conventional spectral subtraction procedure. Anisotropic structural elements on gray-scale spectrograms have been found to provide a better perceptual quality than isotropic ones and reveal themselves as more appropriate for retaining the speech structure while removing background noise. Our procedure has been evaluated by using a number of perceptual quality estimation measures for several Signal-to-Noise Ratios on the Aurora database.