

Lec 33 11/12/01

①

Last week, we used speech processing as a context for talking about autoregressive models and linear prediction, and then made an excursion into Kalman filtering.

Today, we'll use speech processing as a context for discussing pattern recognition and short-time Fourier analysis

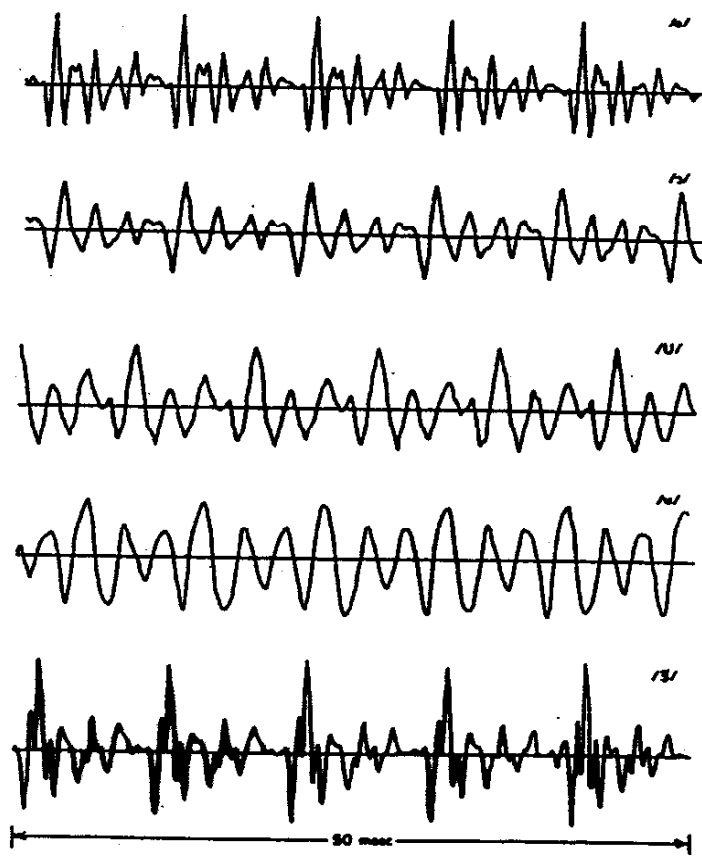


Fig. 3.6 (Continued)

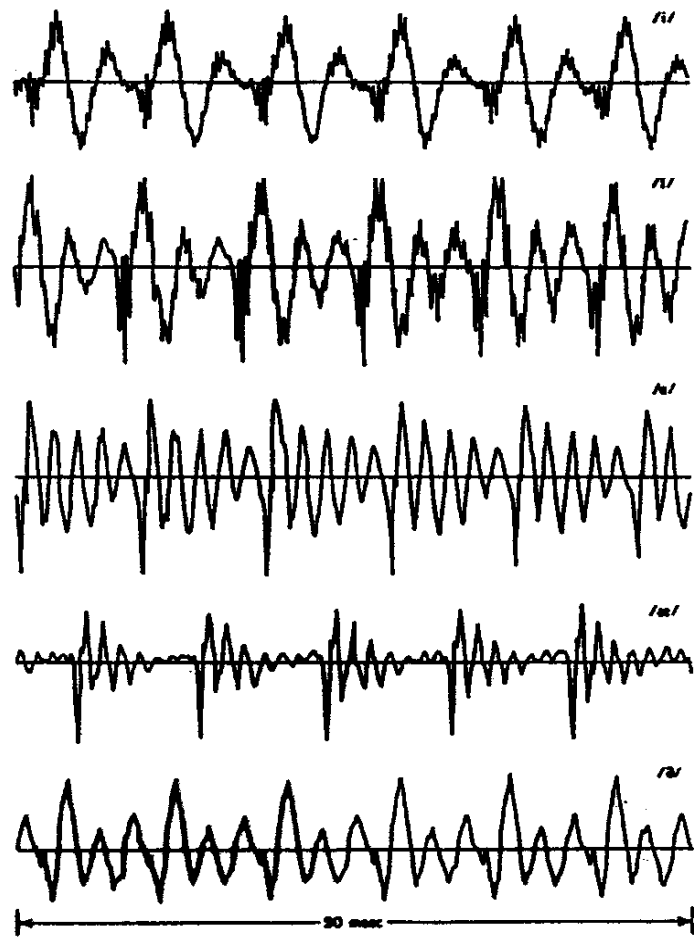
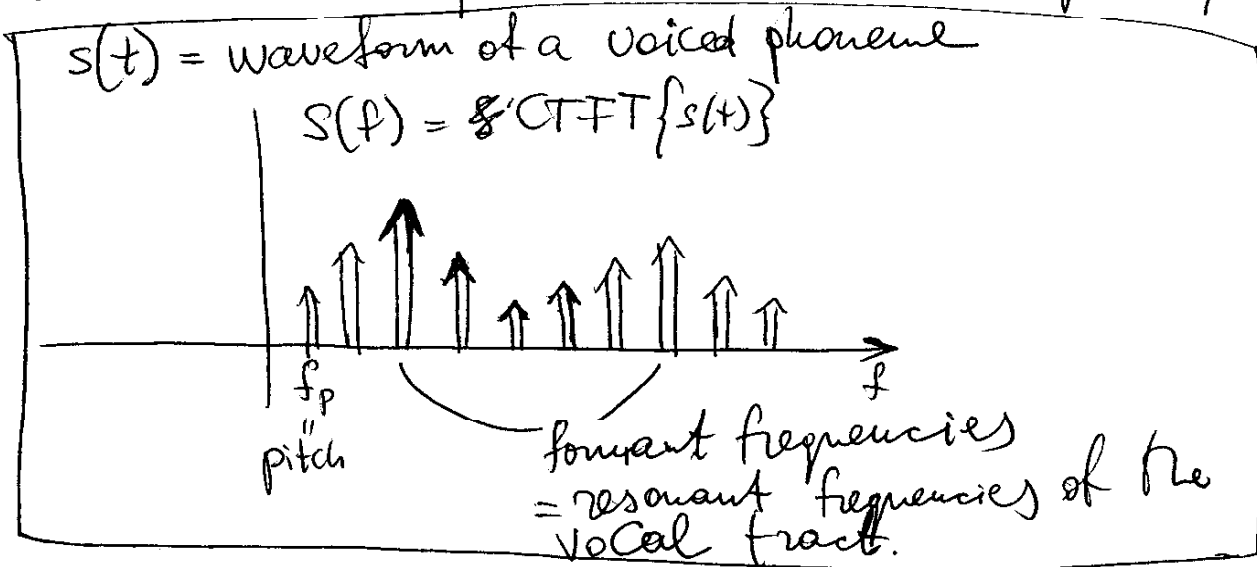


Fig. 1.6 The acoustic waveforms for several American English vowels and corresponding spectrograms.

Formant Analysis (Lab 9 WK 2, Sect. 2.3)

Recall that voiced speech is approximately periodic. The fundamental period is called the pitch period.



~~Waveforms for several vowels~~
You may also recall that these ^{periodic} waveforms are different for different voiced sounds. So, it should be possible to distinguish different sounds somehow, by looking at these waveforms. Two things that help us.

- There are only about 42 different phonemes in American English.
- Vowels are mostly determined by formant frequencies. Different vowels have different sets of formants.

A simple vowel recognition algorithm:

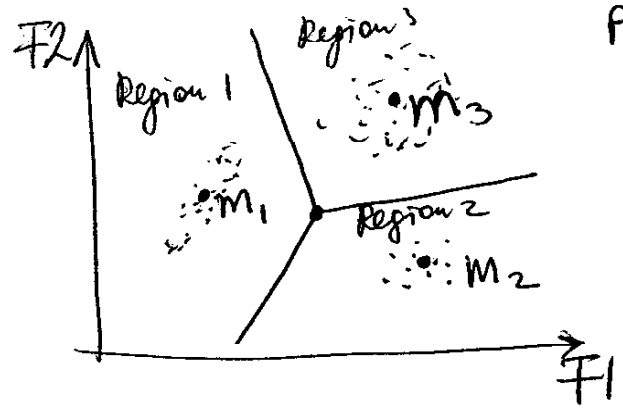
- extract first two formant frequencies
- compare them to a table of formants, and choose the vowel with the "closest" formants

Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F ₁	F ₂	F ₃
IY	i	(beer)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hat)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	u	(foot)	440	1020	2240
OO	ʊ	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

Pattern recognition:

- train the classifier = determine the decision regions (build the table of ^{the table of} _{of formants})
- E.g., assume that $\begin{pmatrix} F1 \\ F2 \end{pmatrix}$ for each vowel is a Gaussian random vector, with mean \underline{m}_k and covariance Λ_k , $k=1, \dots, p$, where p = number of vowels



- Estimate \underline{m}_k and Λ_k from data for class k
- Find decision regions, so as to maximize, e.g., Prob(correct decision)

- recognition/classification: given a vowel sound, find its $F1$ and $F2$. If it's in Region k , say that it's vowel k .

Note that, in this whole story, we sidestepped an important issue, namely, how do we find these formant frequencies?


Problem. Given a speech signal, how do we find the formant frequencies?

If we had an ~~signal for one so~~ infinite-duration signal for one sound, the answer would be simple: take the Fourier transform, and look at its maxima. Unfortunately, if we do that for ~~any~~ a speech signal, we will mix together all the different sounds, and will not get anything meaningful.

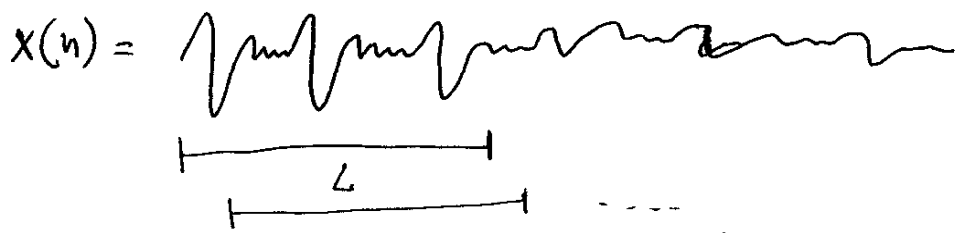
Solution: take Fourier transform over short time windows.

2.4 Short-Time ~~Fourier Transform~~ (Windowed) Fourier Transform

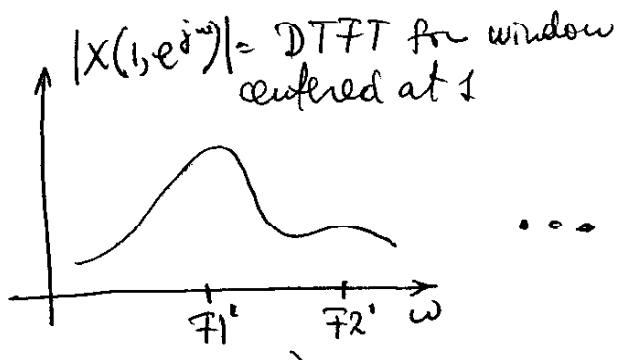
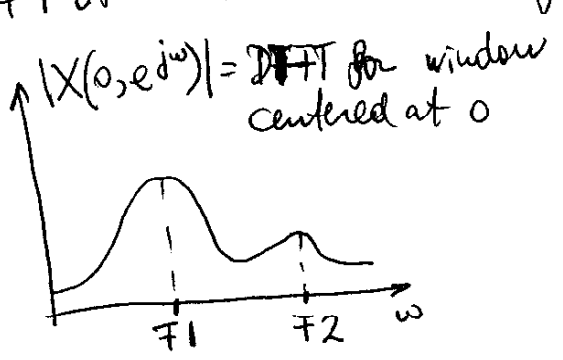
$$X(m, e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n) w(n-m) e^{-j\omega n}$$

- $X(n)$ is a DT signal
- $w(n)$ is a window, e.g. 
- for fixed m , $X(m, e^{j\omega})$ is the DTFT of $x(n)w(n-m)$, ~~the~~ a windowed version of $x(n)$.

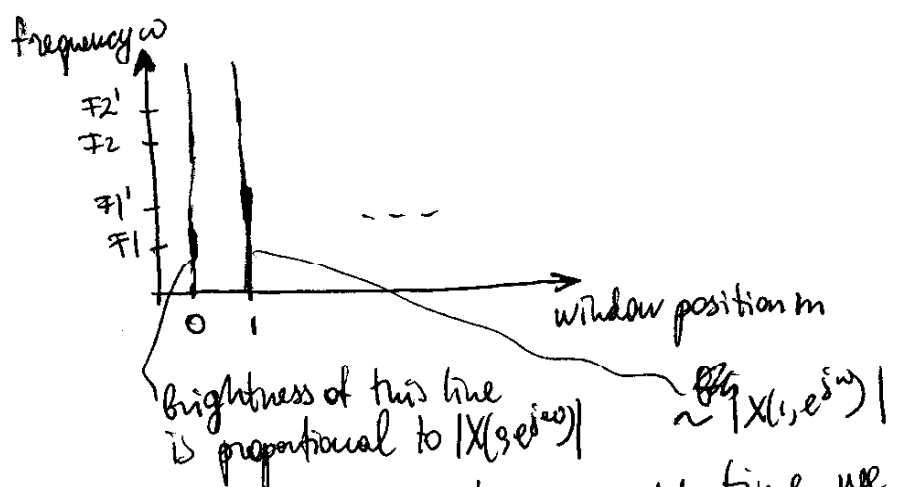
We are trying to window the data, to obtain a segment for which the vocal tract does not change much. Let's see what window would be appropriate.



Chop up the signal into small segments, and take the DFT of each windowed piece:



Spectrogram: $|X(m, e^{j\omega})|$ (or $|X(m, e^{j\omega})|^2$)
 often depicted as an image in the $m-\omega$ plane, with large values of $|X(m, e^{j\omega})|$ corresp. to dark colors, and small " " " " light colors



By tracing the peaks across time, we can estimate the formant frequencies and how they change.

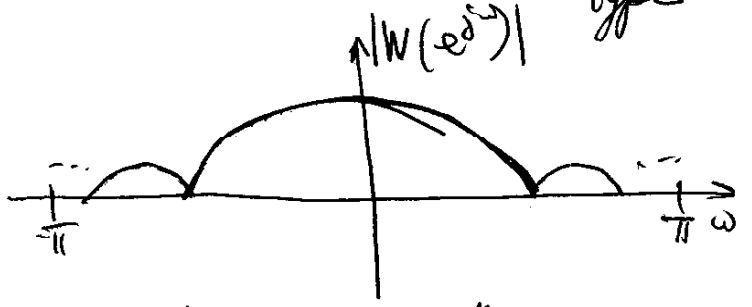
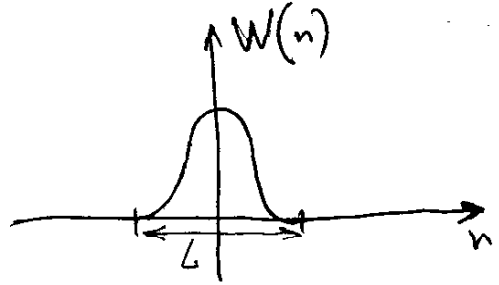
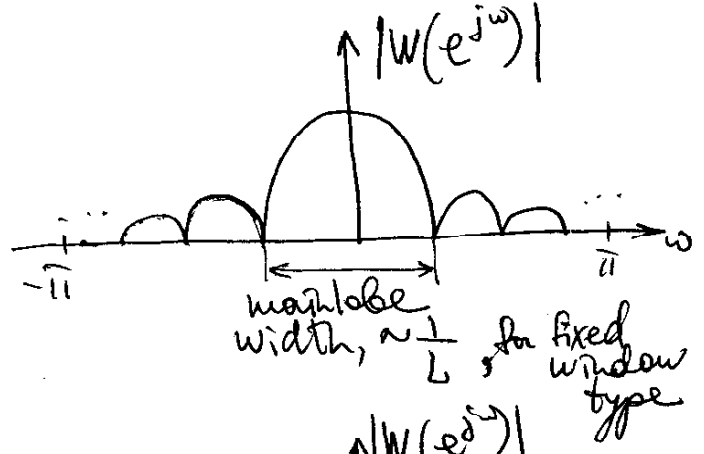
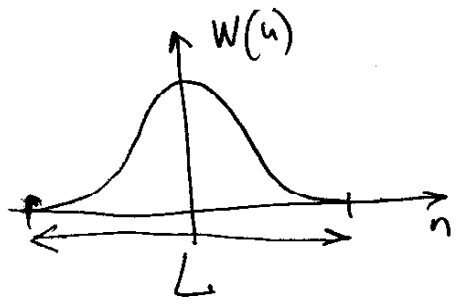
What's a good window width L ?

Let $\tilde{X}(n) = X(n)w(n)$.

Multiplication in the domain corresponds to convolution in frequency domain, in particular, for DTFT,

$$\tilde{X}(e^{j\omega}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\mu}) W(e^{j(\omega-\mu)}) d\mu$$

Recall from the lab on filter design that the Fourier transform of a window is a sinc-like function whose width is inversely proportional to the width of the window:



- Large $L \Rightarrow$ poor resolution in time (mix together many different frequencies),
 But good resolution in frequency: $|W(e^{j\omega})|$ is close to $\delta(\omega) \Rightarrow \tilde{X}(e^{j\omega}) \approx X(e^{j\omega})$.
- Small $L \Rightarrow$ good resolution in time
 But poor resolution in frequency: $|W(e^{j\omega})|$ is wide $\Rightarrow \tilde{X}(e^{j\omega})$ mixes together many frequency components of $X(e^{j\omega})$.

For speech,

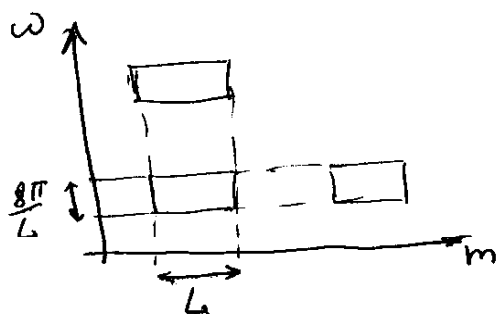
- "Wideband spectrogram"
window length ≈ 1 pitch period
- "Narrowband spectrogram"
window length $\approx 5-6$ pitch periods

Remarks.

1. Windowed CT Fourier transform:

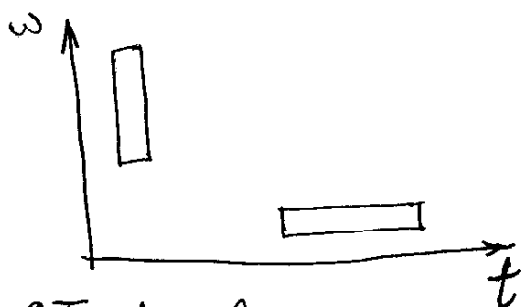
$$X(f, t) = \int_{-\infty}^{\infty} x(u) w(u-t) e^{-j2\pi fu} du$$

2. Time-frequency tiling:
(e.g., take Hamming windows whose mainlobe width is $\frac{8\pi}{L}$)



A window with time ~~resolution~~ size L has frequency size $\sim \frac{8\pi}{L}$
i.e., if we mix together $\sim L$ samples in time, we mix together $\sim \frac{8\pi}{L}$ frequencies

A different time-frequency tiling:

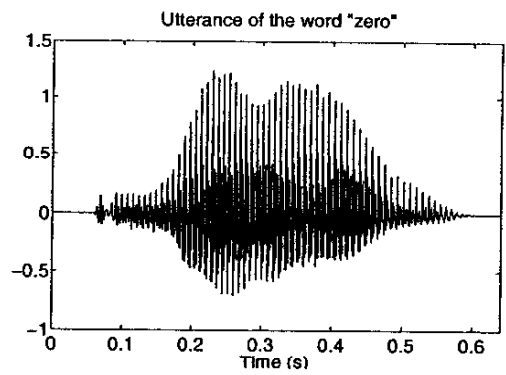


(low resolution at low freq., high resolution at high freq.)

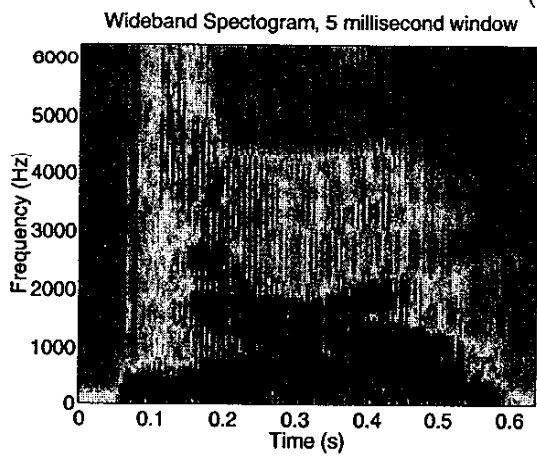
CT Wavelet transform:

$$Wf(t, s) = \int_{-\infty}^{\infty} f(u) \frac{1}{\sqrt{s}} \psi^*\left(\frac{u-t}{s}\right) du$$

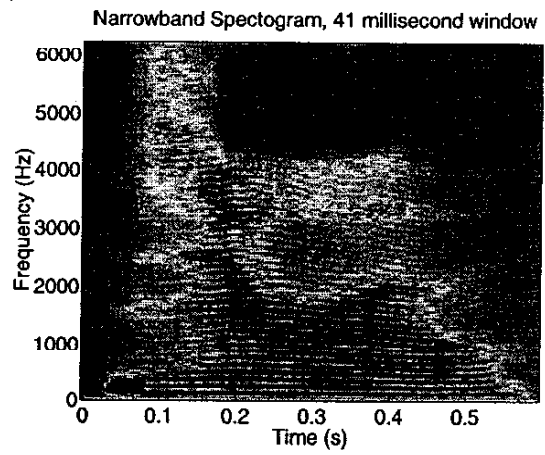
(To analyze structures of different time-frequency sizes, use time-freq. atoms of different supports.)



(a)



(b)



(c)

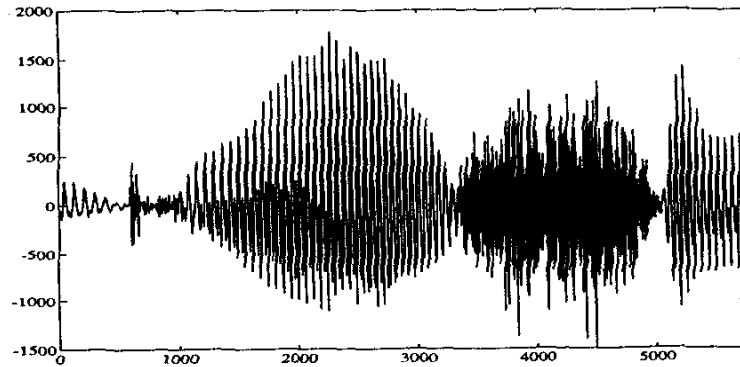


Figure 2(a): speech recording of the word “greasy”, sampled at 16 kHz.

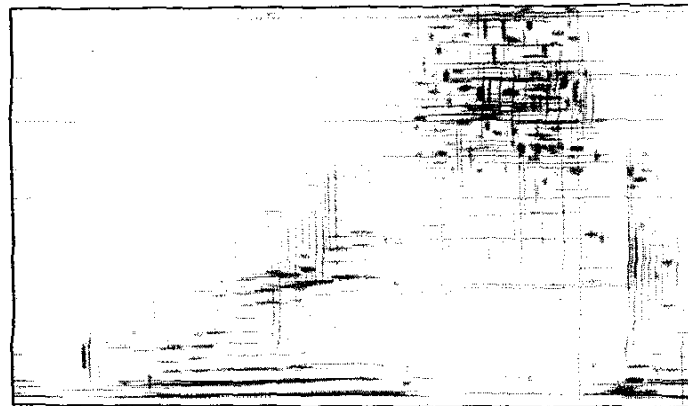


Figure 2(b): time-frequency energy distribution of the speech recording shown in (a). We see the low-frequency component of the “g”, the quick burst transition to the “ea” and the harmonics of the “ea”. The “s” has energy spread over high frequencies.

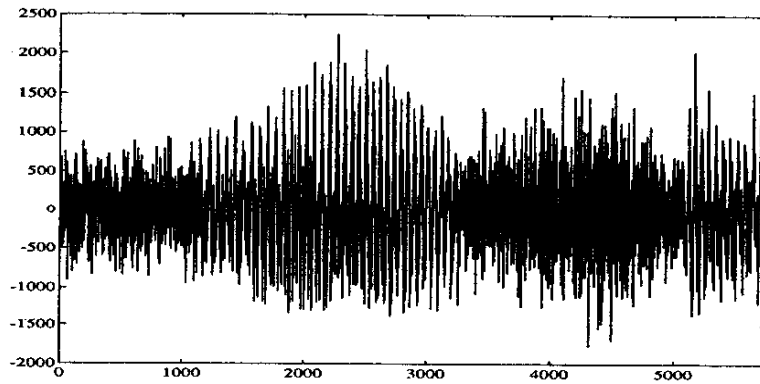


Figure 4(a): signal obtained by adding a Gaussian white noise to the speech recording shown in Fig.2(a). The signal to noise ratio is 1.5db.

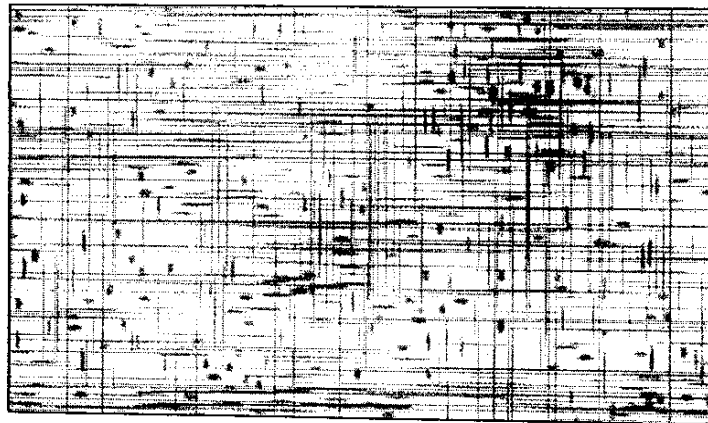


Figure 4(b): time-frequency energy distribution of the noisy speech signal. The energy distribution of the white noise is spread across the whole time-frequency plane.

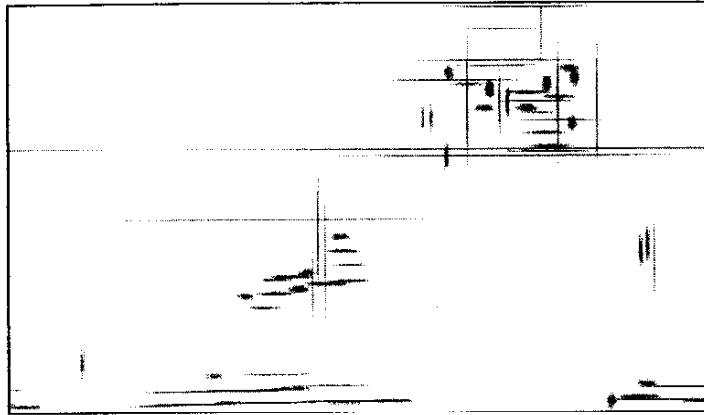


Figure 5(a): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig.3(a).

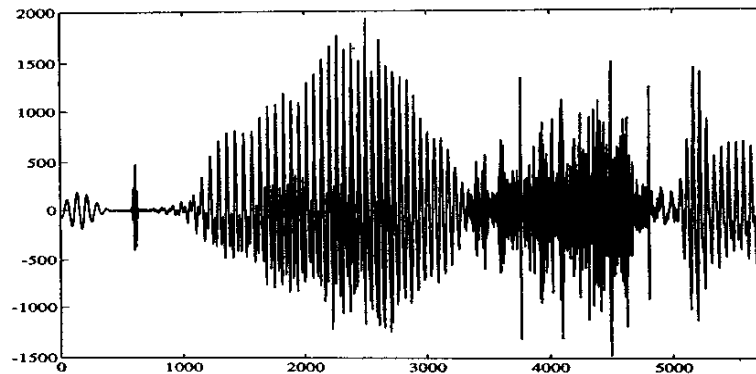


Figure 5(b): time-frequency energy distribution of the $m = 76$ coherent structures of the noisy speech signal shown in Fig. 3. (b): signal reconstructed from the 76 coherent structures shown in (a). The white noise has been removed.

A Wavelet Tour of Signal Processing
Stéphane Mallat, Academic Press 1999 (2nd edition)

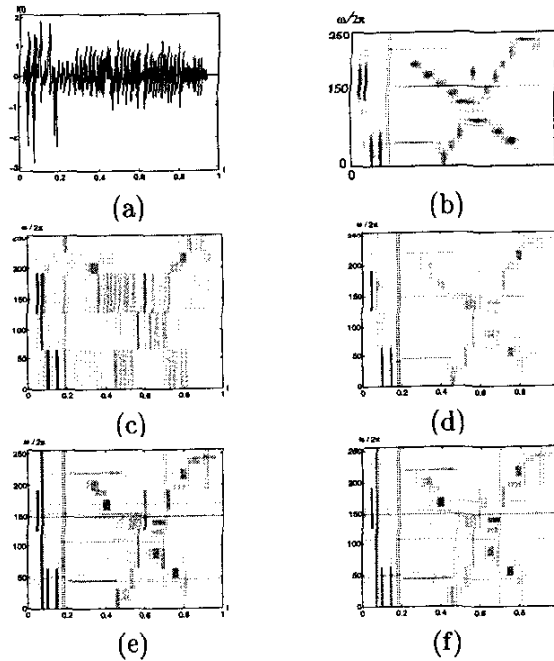


Figure 9.11: (a): Signal synthesized with a sum of chirps, truncated sinusoids, short time transients and Diracs. The time-frequency images display the atoms selected by different adaptive time-frequency transforms. The darkness is proportional to the coefficient amplitude. (b): Gabor matching pursuit. Each dark blob is the Wigner-Ville distribution of a selected Gabor atom. (c): Heisenberg boxes of a best wavelet packet basis calculated with Daubechies 8 filter. (d): Wavelet packet basis pursuit. (e): Wavelet packet matching pursuit. (f): Wavelet packet orthogonal matching pursuit.