Short-Time Discrete-Time Fourier Transform (STDTFT)

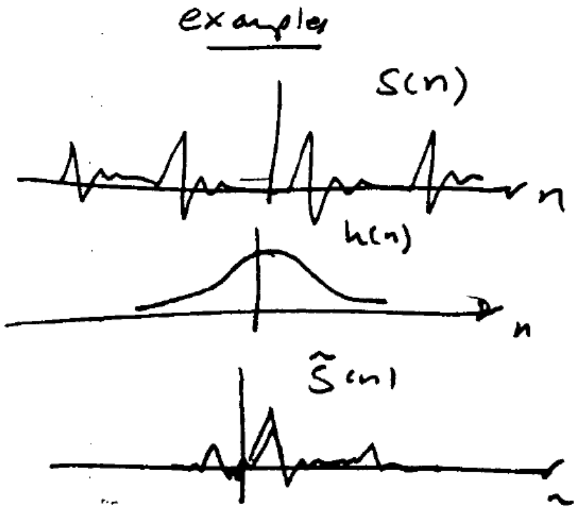
Two interpretations for the STDFT

1. *Spectrogram* – Fix the displacement n_0 of the window, and vary the frequency ω at which the transform is evaluated
2. *Filter bank* – Fix the frequency ω at which the transform is evaluated, and vary the displacement n_0 of the window

Time-Domain Definitions

3. Speech waveform – $s[n]$
4. Window function – $w[n]$
5. Windowed speech waveform – $\tilde{s}\left[n,n_0\right]=s[n]w[n_0-n]$
    a. The parameter n_0 is the displacement of the window, and is considered to be fixed (This is consistent with the first interpretation of the STDTFT)
    b. A more natural way to define $\tilde{s}\left[n,n_0\right]$ would be $\tilde{s}\left[n,n_0\right]=s[n]w[n-n_0]$ since n_0 is a fixed displacement
    c. However, we shall see later that it will be convenient to define $\tilde{s}\left[n,n_0\right]$ as in part (a) above
    d. Note that we can write $\tilde{s}\left[n,n_0\right]=s[n]w[-(n-n_0)]$
    e. If the window is even, i.e. $w[-n]=w[n]$, then $\tilde{s}\left[n,n_0\right]=s[n]w[n-n_0]$

Example Time-Domain Waveforms



Note that the displacement is zero for this example; so the window is not shifted

STDTFT

1. Now let us consider the DTFT of the windowed speech waveform $\tilde{s}[n,n_0]$

$$\tilde{S}(\omega,n_0)=\sum_{n=-\infty}^{\infty}\tilde{s}[n,n_0]e^{-jwn}$$

$$=\sum_{n=-\infty}^{\infty}s[n]w[-(n-n_0)]e^{-jwn}$$

   a. Note that for convenience, we let the summation range from $n=-\infty$ to $n=\infty$, keeping in mind that the summand will only be non-zero within the support of the shifted window $w[-(n-n_0)]$

   b. Note that $\tilde{S}(\omega,n_0)$ depends on two parameters:
      i. The frequency ω
      ii. The displacement of the window n_0

2. According the product theorem we have that

$$\tilde{S}(\omega,n_0)=\frac{1}{2\pi}\int_{-\pi}^{\pi}S(\mu)DTFT\left\{w[-(n-n_0)]\right\}\Big|_{(\omega-\mu)}d\mu$$

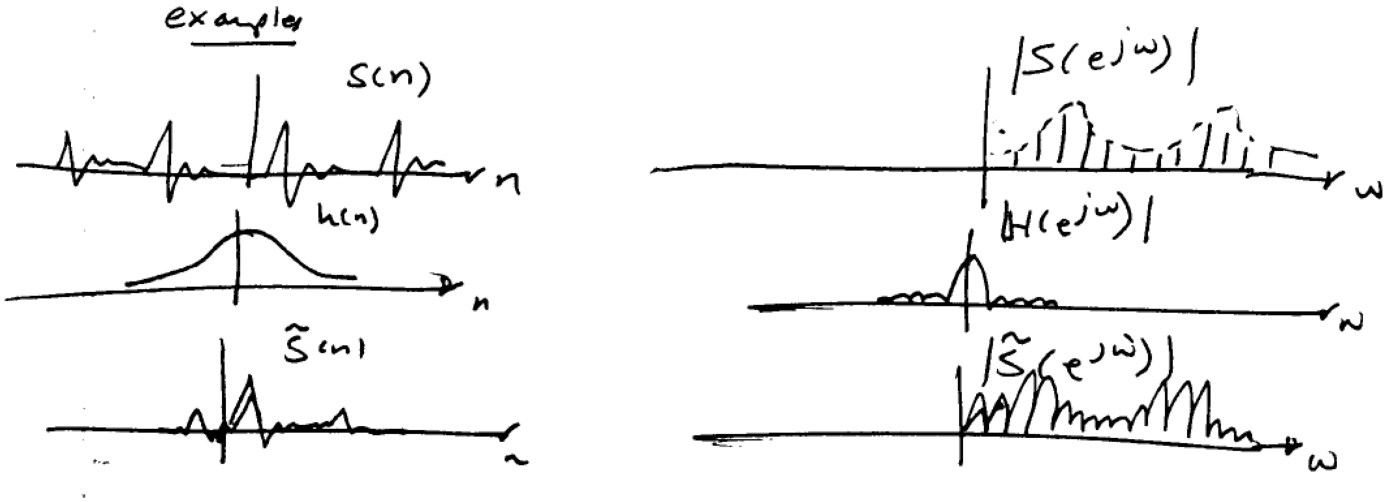
$$=\frac{1}{2\pi}\int_{-\pi}^{\pi}S(\mu)W\left(-(\omega-\mu)\right)e^{-j(\omega-\mu)n_0}d\mu$$

3. Here, we have used the facts that

$$w[-n]\overset{DTFT}{\leftrightarrow}W(-\omega)$$

$$w[-(n-n_0)]\overset{DTFT}{\leftrightarrow}W(-\omega)e^{-j\omega n_0}$$

Example Time-Domain Waveforms and their Spectra



Note again that the displacement of the window is zero in this example.

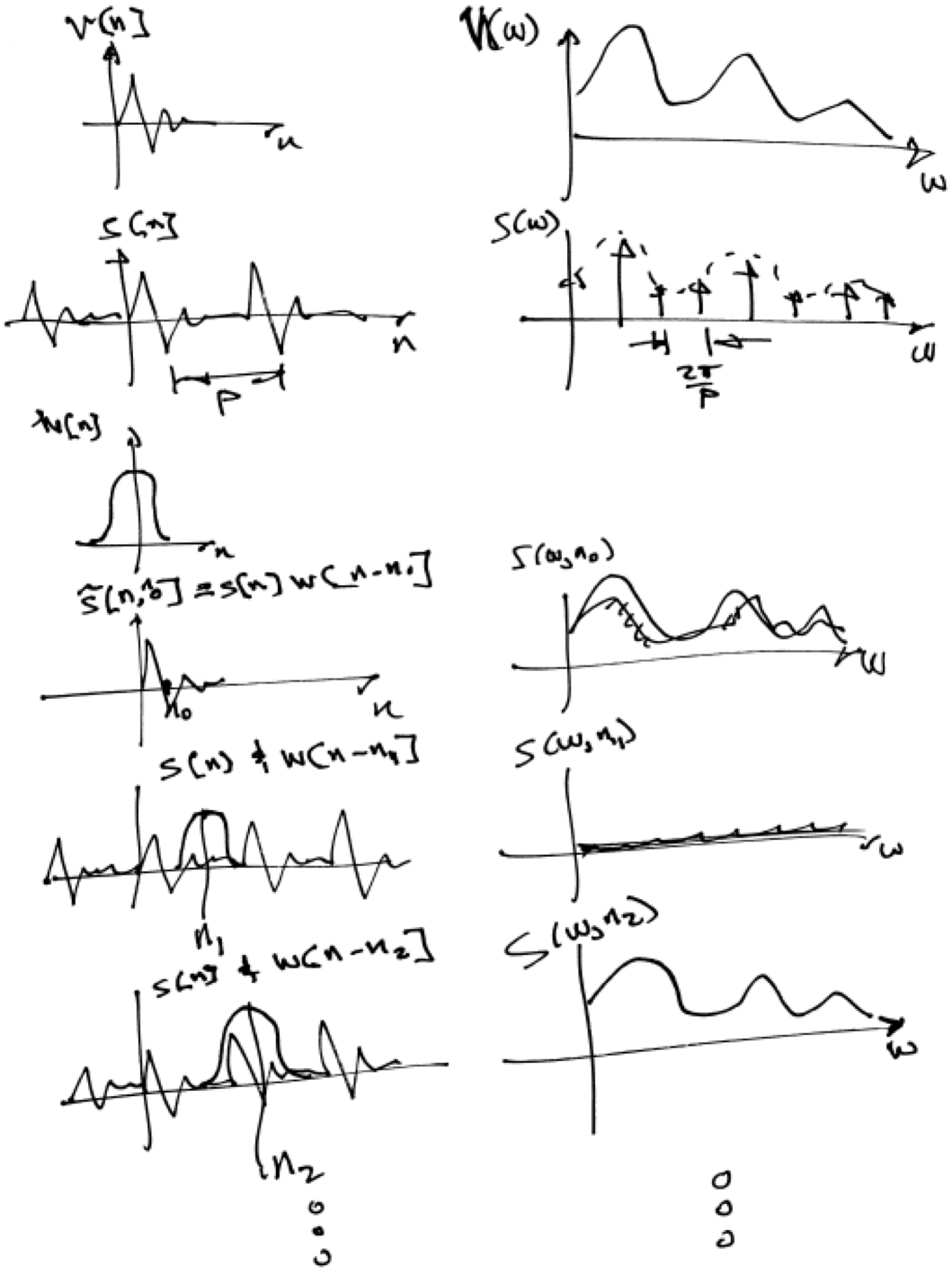
Wideband and Narrowband Spectrograms

1. The example we are showing here is that of an epoch from a voiced phoneme
2. Assume that the speech waveform $s[n]$ has period P samples.
3. This is an abuse of our notation, since we earlier defined P to be the period in seconds for the continuous-time waveform $s(t)$. However, we will not further consider the continuous-time case. So, hopefully, there will be no confusion.
4. Assume that the window function satisfies the following properties
   a. $w[-n] = w[n]$
   b. $w[n] \neq 0$ only for $-(N-1)/2 \leq n \leq (N-1)/2$, where N is assumed to be odd; so $(N-1)/2$ is integer-valued
5. There are two cases to consider
   a. Wideband – $N < P$
   b. Narrowband – $N \gg P$
6. To visualize the spectra for these two cases, we
   a. Vary the displacement n_0 of the window
   b. Use the fact that since $w[-n] = w[n]$, $W(-\omega) = W(\omega)$
   c. Ignore the phase factor in the integral for $S(\omega, n_0)$

$$\tilde{S}(\omega, n_0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\mu) W(\omega - \mu) e^{-j(\omega-\mu)n_0} d\mu$$

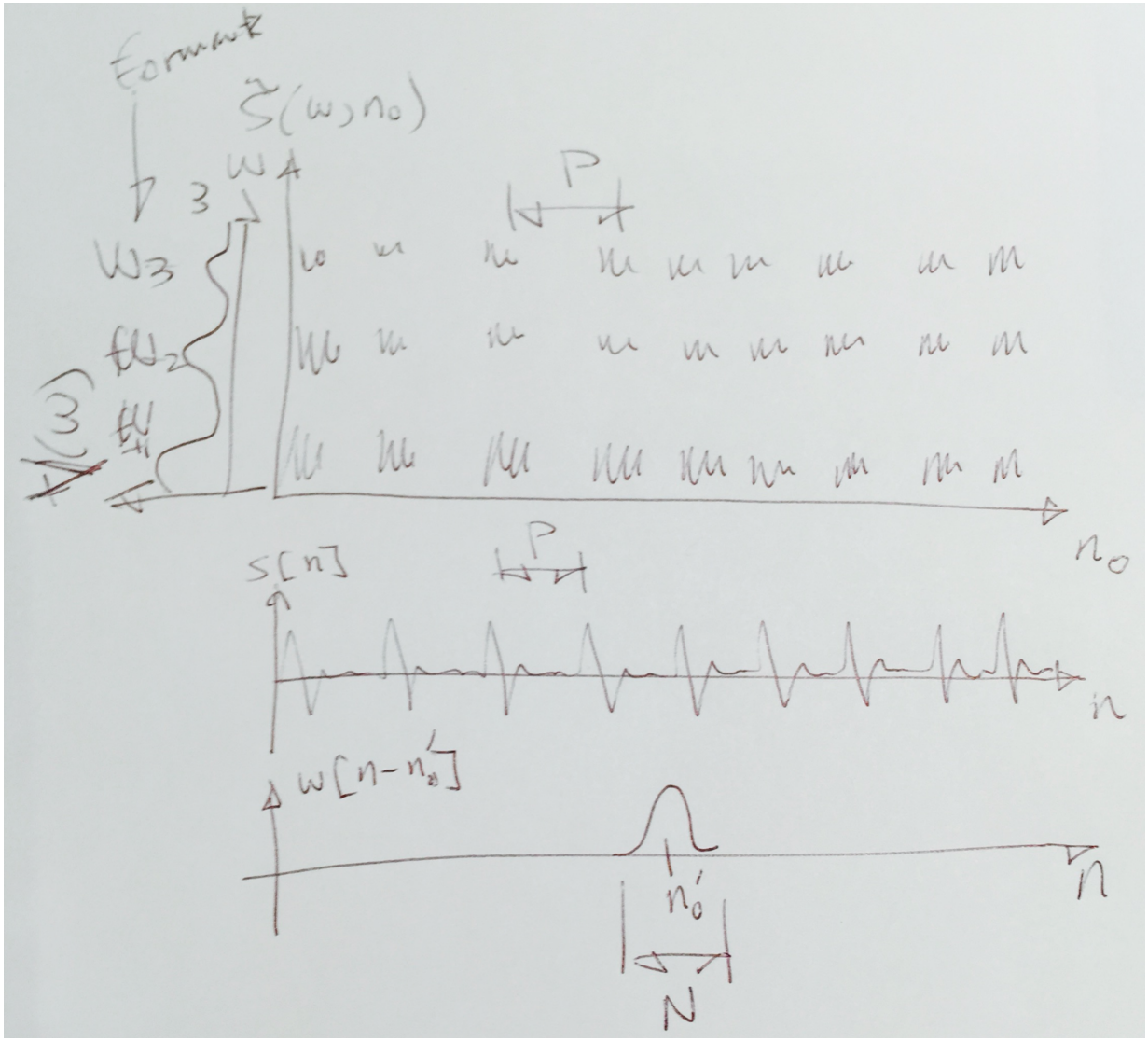
$$\approx \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\mu) W(\omega - \mu) d\mu$$

Visualization of Wideband Spectrum ($N < P$)

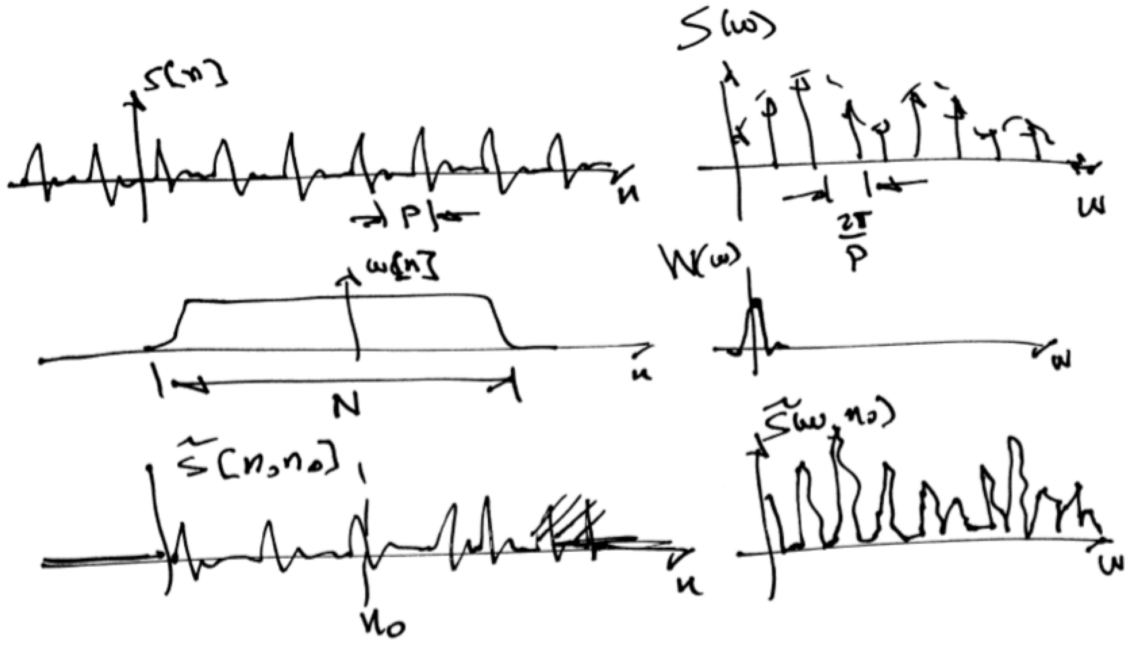


1. We see that as the window slides along the time axis, it alternately:
   a. captures a complete replica of the vocal tract response (when displacements are n_0 and n_2)
   b. picks up only the gap between two replicas of the vocal tract response (when the displacement is n_1)
2. Thus, as a function of the displacement n_0 of the window, $\tilde{S}(\omega, n_0)$ alternates between:
   a. showing the spectrum $V(\omega)$ of the vocal tract response
   b. having very little energy.
3. The alternation occurs with interval given by the pitch period P.
4. Since as a function of the displacement n_0, the STDTFT $\tilde{S}(\omega, n_0)$ resolves the temporally periodic structure of the speech waveform $s[n]$, we call this a *wideband spectrogram*
5. The spectrogram is a plot of the magnitude $\left|\tilde{S}(\omega, n_0)\right|$ as a function of
   a. the displacement n_0 along the horizontal axis
   b. the frequency ω along the vertical axis
   c. the magnitude is depicted by the darkness at each point; so the spectrogram is effectively an image

Wideband Spectrogram



Visualization of Narrowband Spectrogram ($N \gg P$)

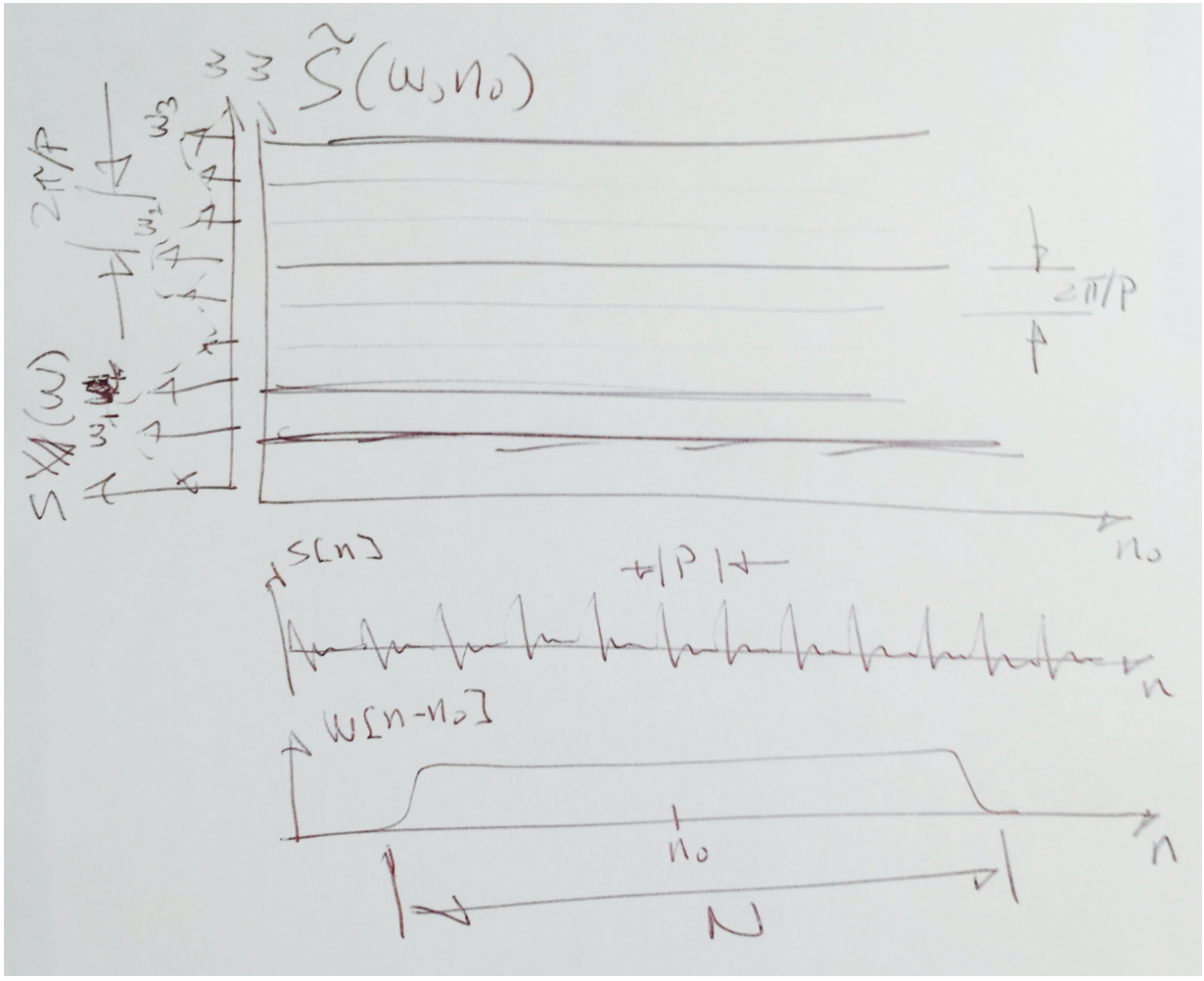


1. Here, we see that the window captures many replications of the vocal tract response for any value of the displacement n_0
2. Thus, for any n_0, the $\left|\tilde{S}(\omega,n_0)\right|$ is essentially the same, and does not change as a function of n_0
3. On the other hand, since the window captures many replications of the vocal tract response $v[n]$, $\tilde{s}\left[n,n_0\right]$ is looks like a signal that is periodic as a function of n
4. So its DTFT shows localizations of spectral energy at multiples of $2\pi / P$ radians/sample along the frequency axis ω.
5. More precisely, convolution of $W(\omega)$ with $S(\omega)$ as indicated by

$$\tilde{S}(\omega, n_0) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\mu) W(\omega - \mu) d\mu$$

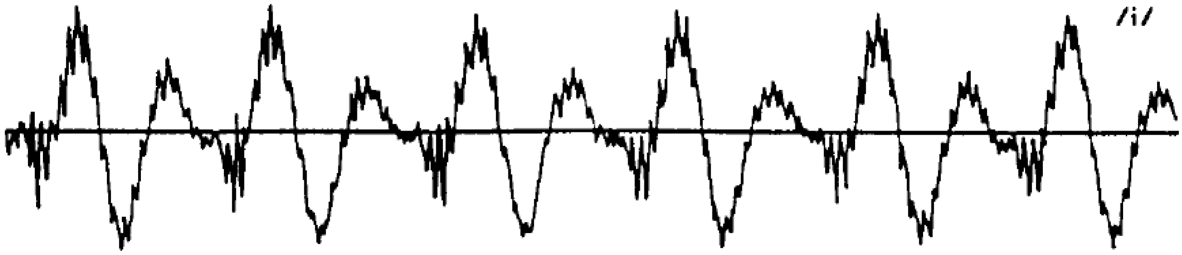
causes the replacement of each impulse in $S(\omega)$ by a shifted replica of $W(\omega)$.

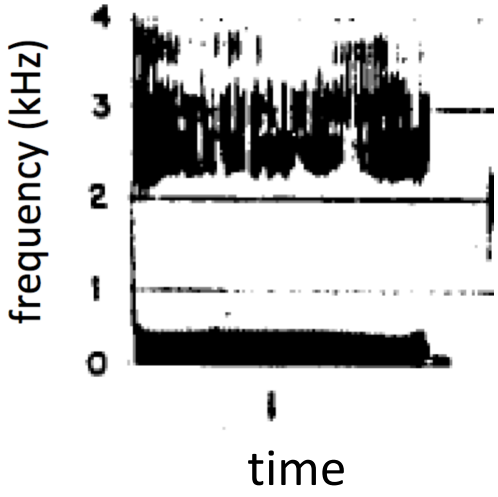
Narrowband Spectrogram



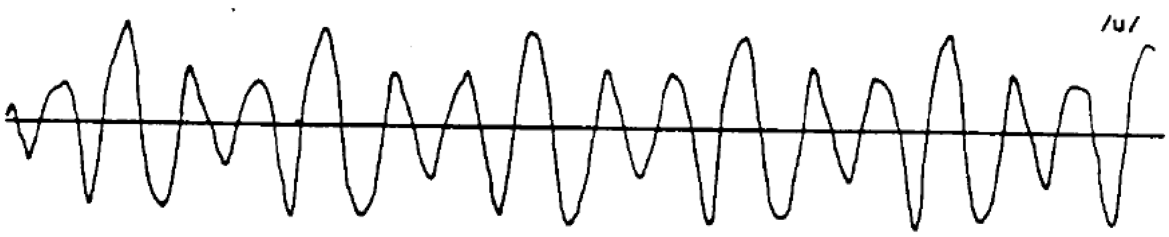
Example waveforms and wideband spectrograms for real speech phonemes

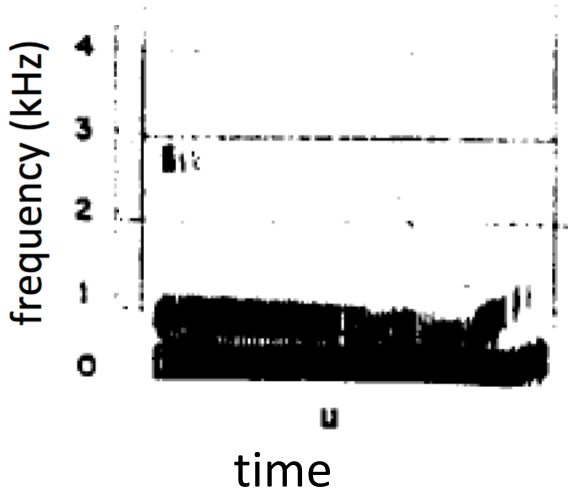
1. /i/ as in “beet”
    a. low first formant (270 Hz), high second formant (2290 Hz)
    b. time waveform shows slow oscillation with superimposed fast oscillation





2. /u/ as in “loom”
    a. low first and second formants (300 Hz and 870 Hz)
    b. time waveform is very smooth





Additional waveforms and wideband spectrograms for real speech phonemes (1/2)

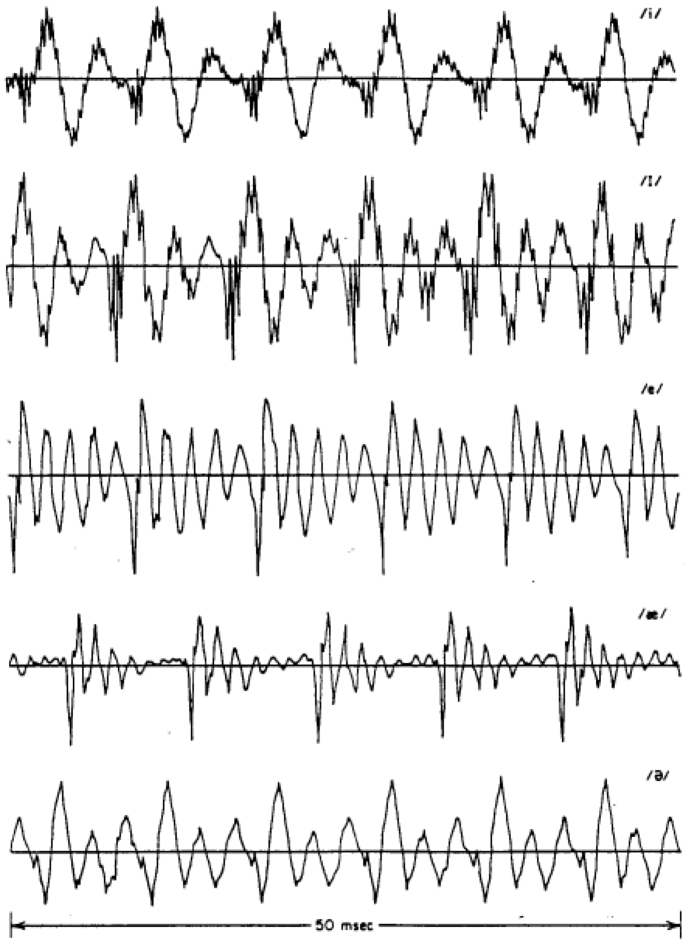
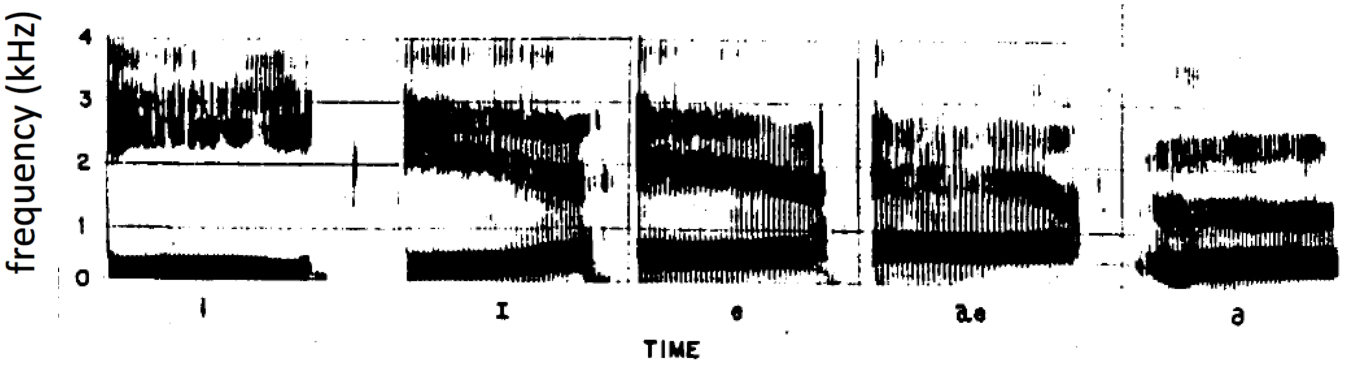


Fig. 3.6 The acoustic waveforms for several American English vowels and corresponding spectrograms.



Additional waveforms and wideband spectrograms for real speech phonemes (2/2)

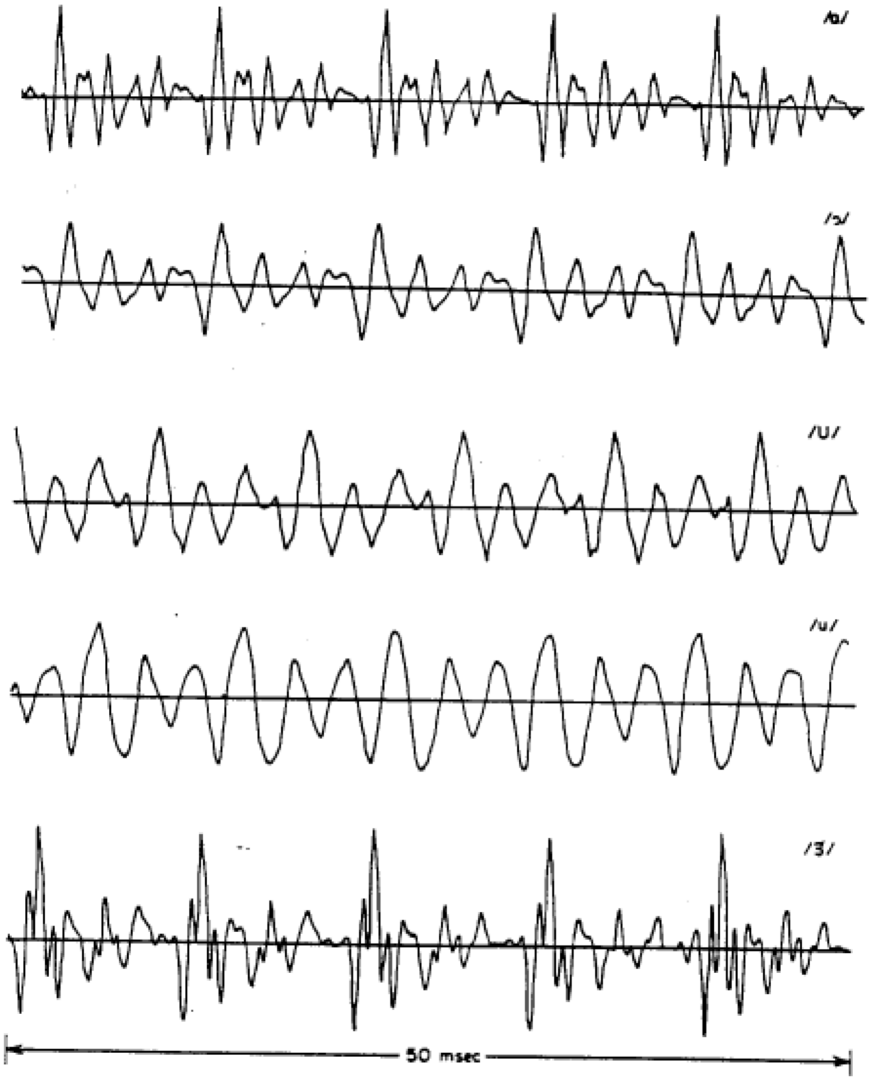
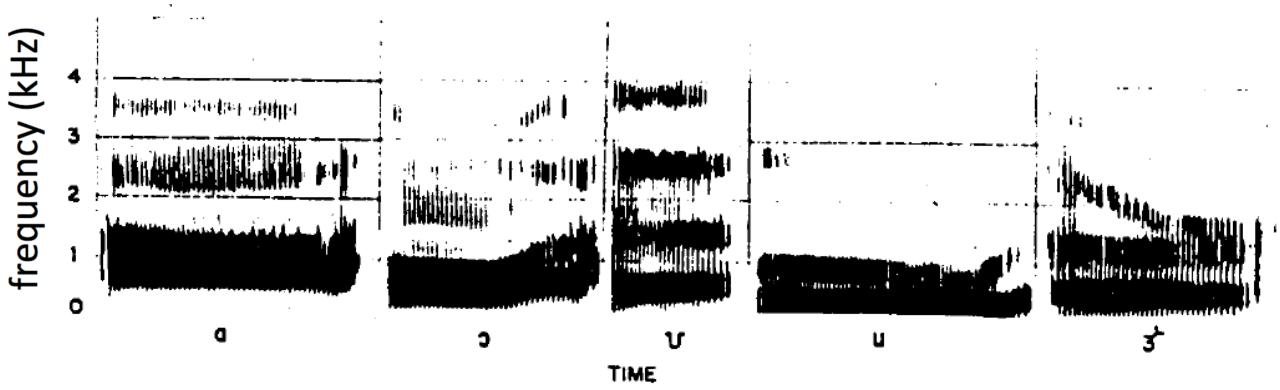


Fig. 3.6 (Continued)



Formants for the English language (to aid interpretation of the plots on the preceding two pages)

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	(beet)	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	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

Time Domain Waveform and Wideband and Narrowband spectrograms for the utterance: "Should we chase"

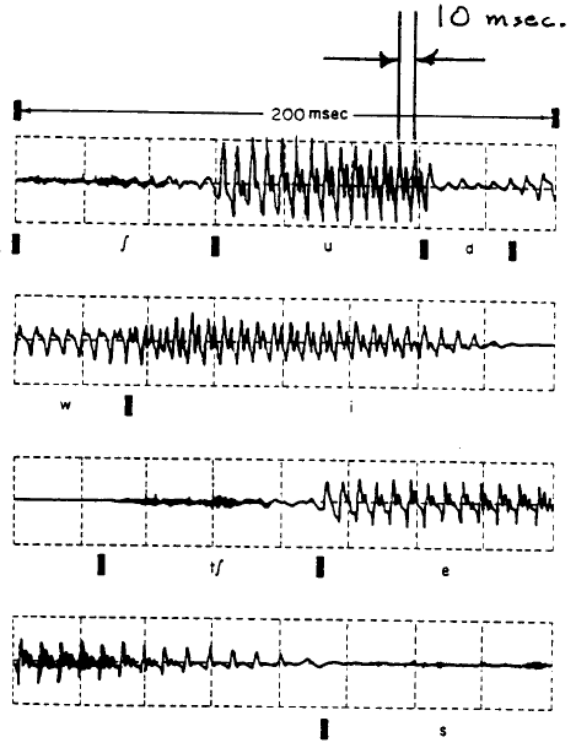
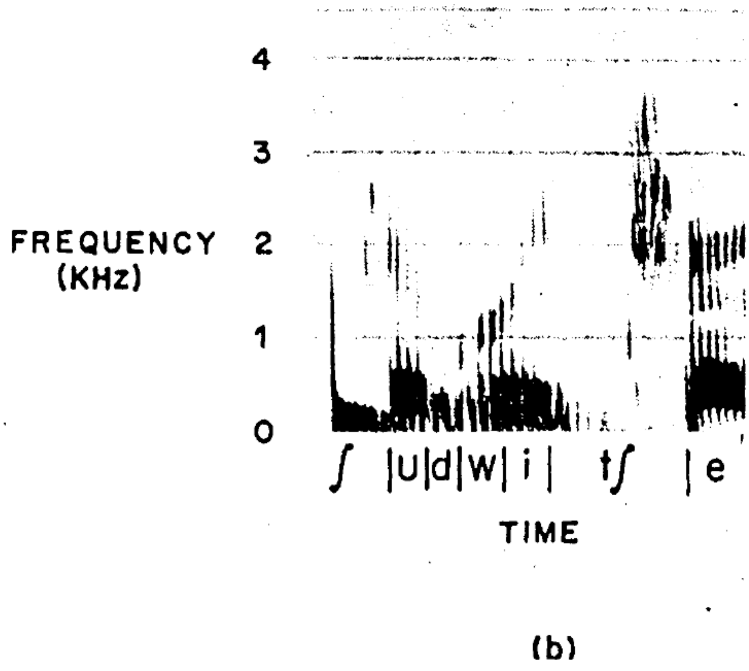


FIGURE 3-2. Example of a speech waveform illustrating different classes of sounds. The utterance is "should we chase . . .".



(b)

Wideband and Narrowband spectrograms for the utterance: "There was some delay on the rayon stockings"

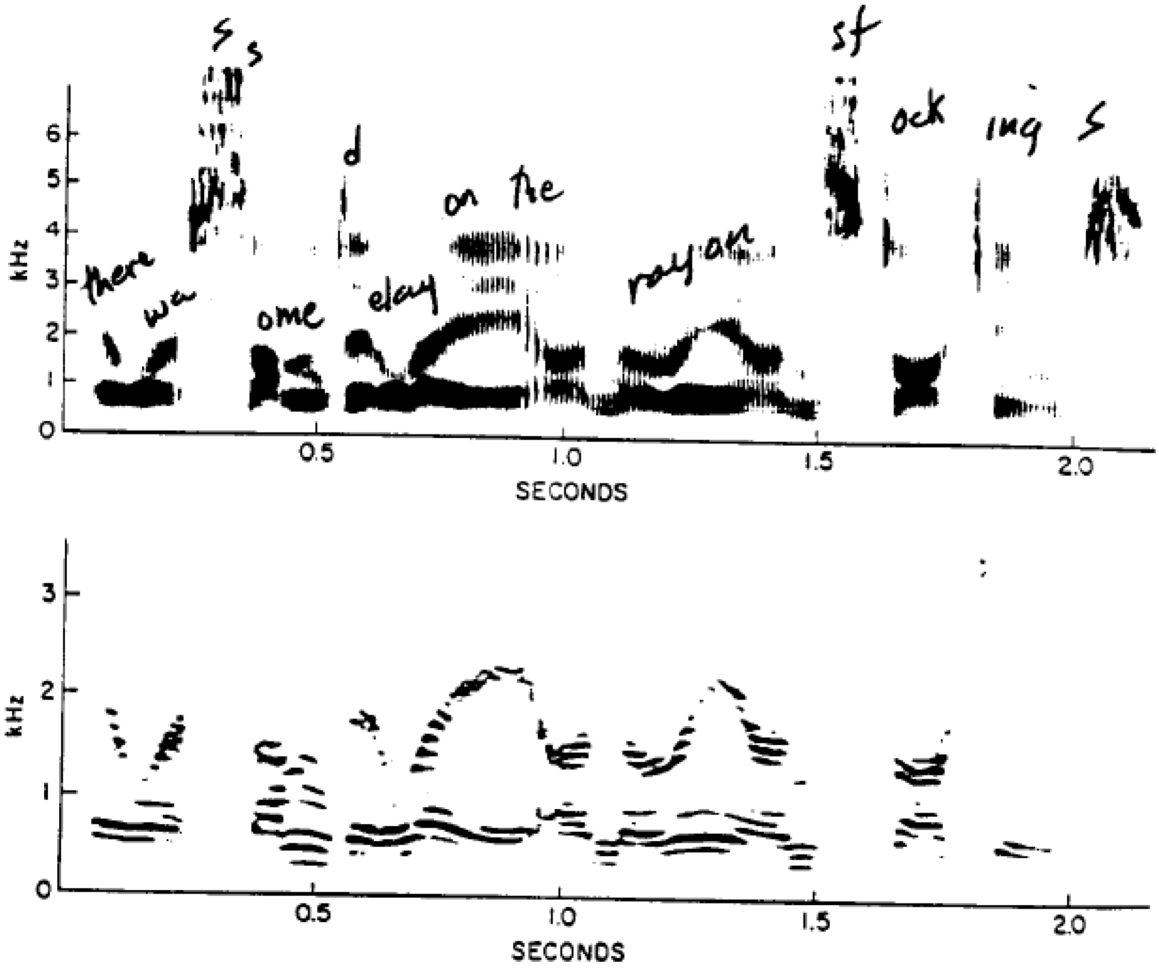


FIGURE 3-4. Speech spectrograms of the utterance "there was some delay on the rayon stockings"; (a) wideband spectrogram; (b) narrowband spectrogram.