COMP ENG 4TL4:

Digital Signal Processing

Notes for Lecture #27

Tuesday, November 11, 2003

6. SPECTRAL ANALYSIS AND ESTIMATION

6.1 Introduction to Spectral Analysis and Estimation

The discrete-time Fourier transform (DTFT) decomposes infinite discrete-time signals into infinite-duration complex exponentials with infinite frequency resolution.

In Lecture #15 we saw that if we limit the duration of discrete-time sequences by windowing, we limit the effective frequency resolution of the DTFT, and consequently of the discrete Fourier transform (DFT).

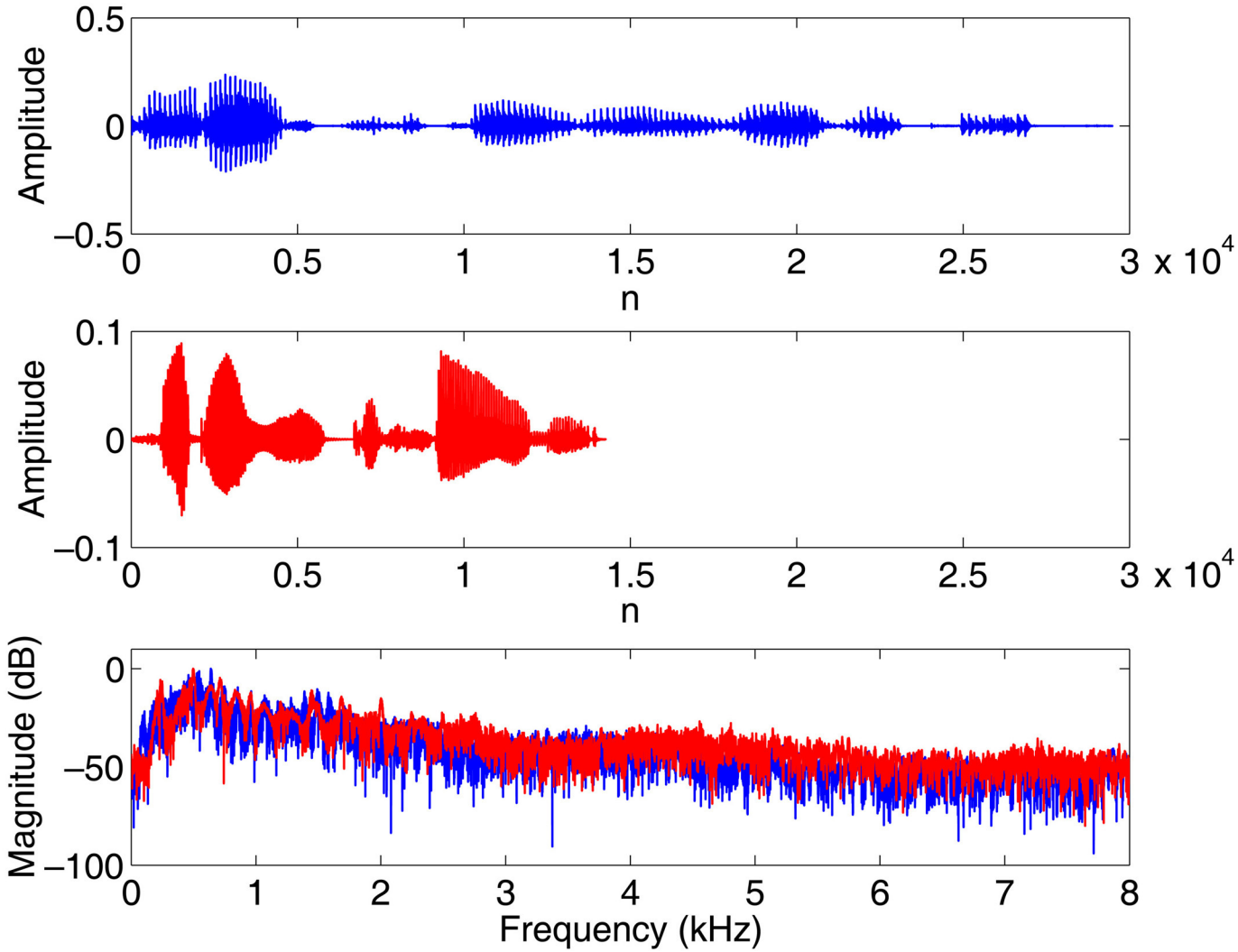
Question: Should our aim then always be to use the longest window length that is computationally feasible when analyzing the spectrum of a signal with the DFT?

Answer: No! Two important cases in which we may wish to use shorter window lengths are:

1. spectral analysis of *time-varying signals* (e.g., speech), and
2. spectral estimation of *stationary random signals*.

The reason for the former should be self evident (see the next slide); the reason for the latter will become apparent later.

Long-term spectra of two different sentences:



6.2 Spectral Analysis of Time-Varying Signals

Short-Time Fourier Transform (STFT):

The STFT (sometimes referred to as the *time-dependent* Fourier transform) of a signal $x[n]$ is defined as:

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

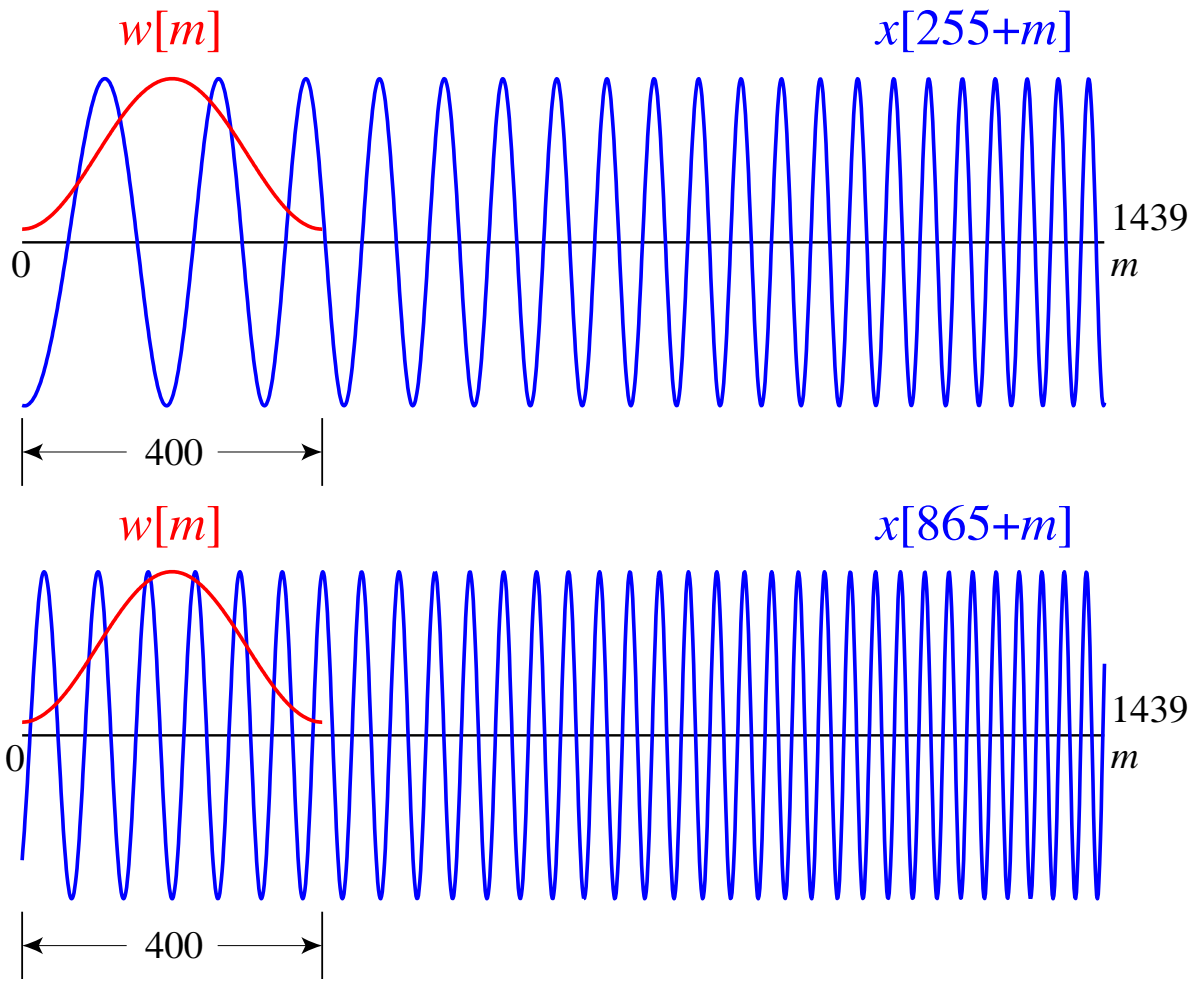
where $w[n]$ is a window sequence of length L.

- Note that a one-dimensional sequence $x[n]$ is transformed into a two dimensional function of the time variable n, which is discrete, and the frequency variable ω, which is continuous.
- Like in the DTFT, the frequency variable ω is periodic with 2π, so we need only consider values of ω for $0 \leq \omega < 2\pi$.
- The STFT can be interpreted as the DTFT of the shifted signal $x[n+m]$ as it moves past the stationary window $w[m]$.

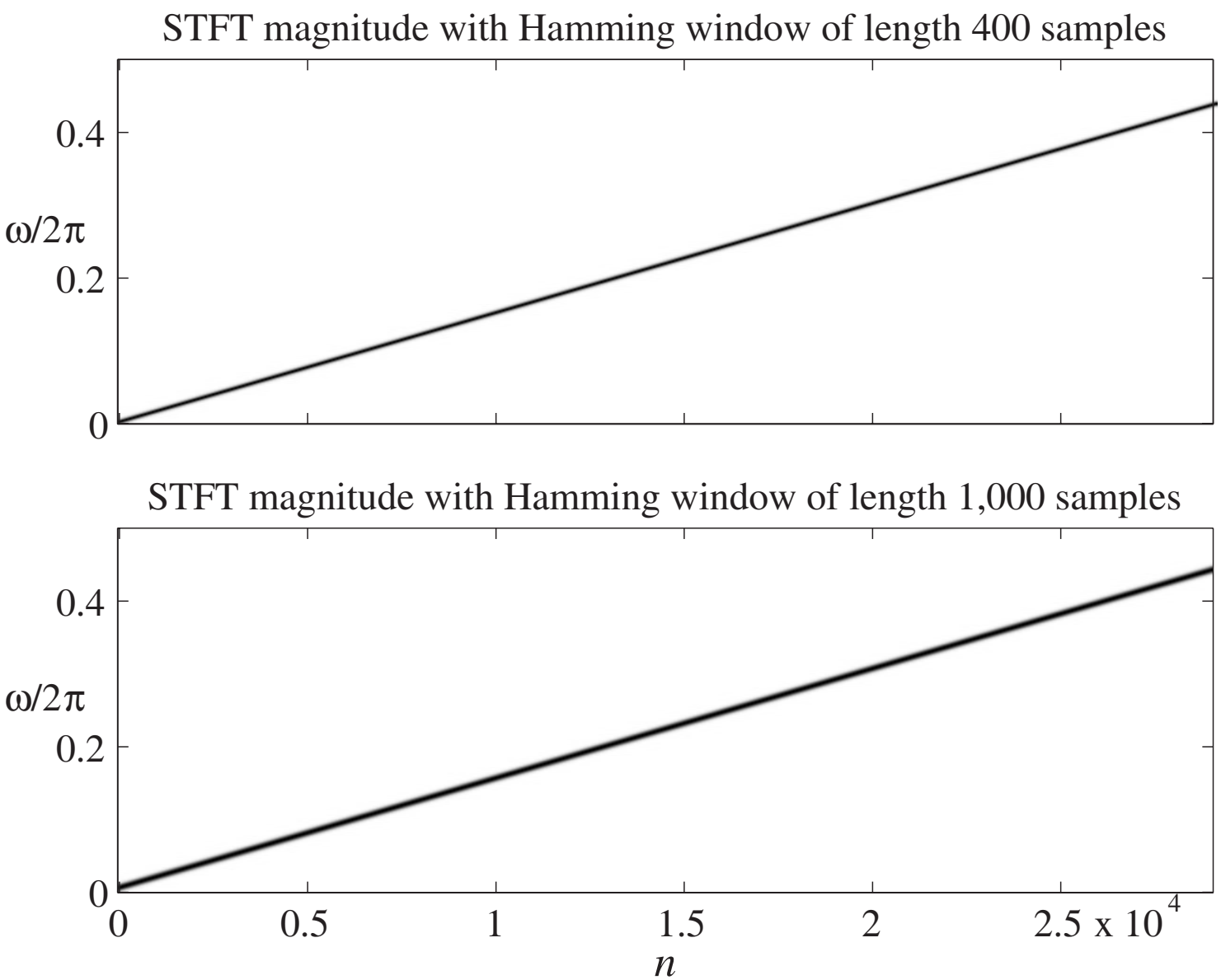
Example: Consider the discrete-time signal:

$$x[n] = \cos\left(\omega_0 n^2\right), \quad \omega_0 = 2\pi \times 7.5 \times 10^{-6},$$

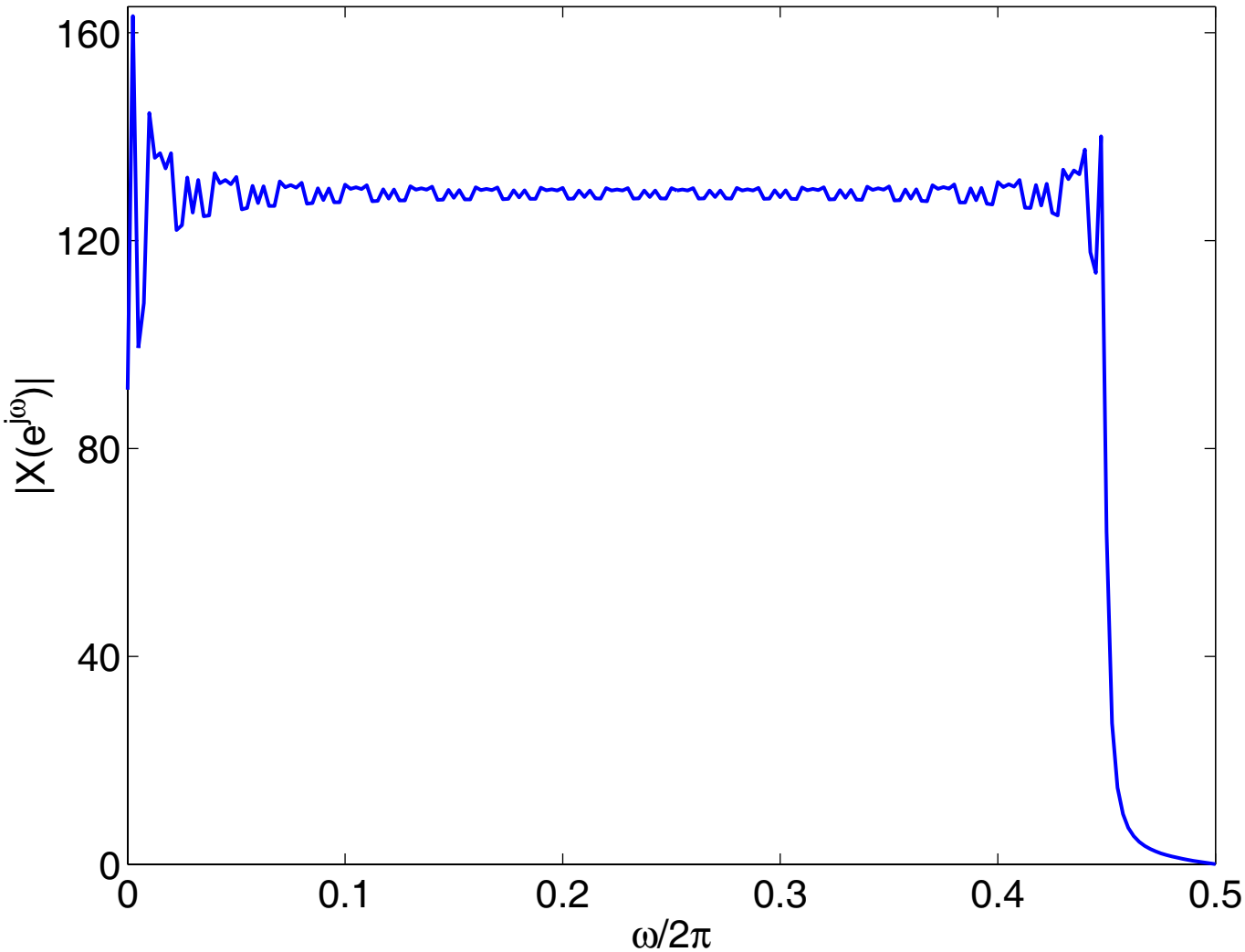
referred to as a linear chirp.



STFT magnitude of the linear chirp signal:



DTFT of the whole chirp signal:



The inverse STFT is given by:

$$x[n] = \frac{1}{2\pi w[0]} \int_0^{2\pi} X[n,\omega)\, d\omega,$$

if $w[0] \neq 0$.

Note that if we sample $X[n,\omega)$ at N equally spaced frequencies $\omega_k = 2\pi k/N$, with $N \geq L$, then we can still recover the original sequence $x[n]$.

This gives us the discrete STFT:

$$\begin{aligned} X[n,k] &= X[n, 2\pi k/N) \\ &= \sum_{m=0}^{L-1} x[n+m]\, w[m]\, e^{-j(2\pi/N)km}, \end{aligned}$$

which is the DFT of the windowed sequence $x[n+m]w[m]$.

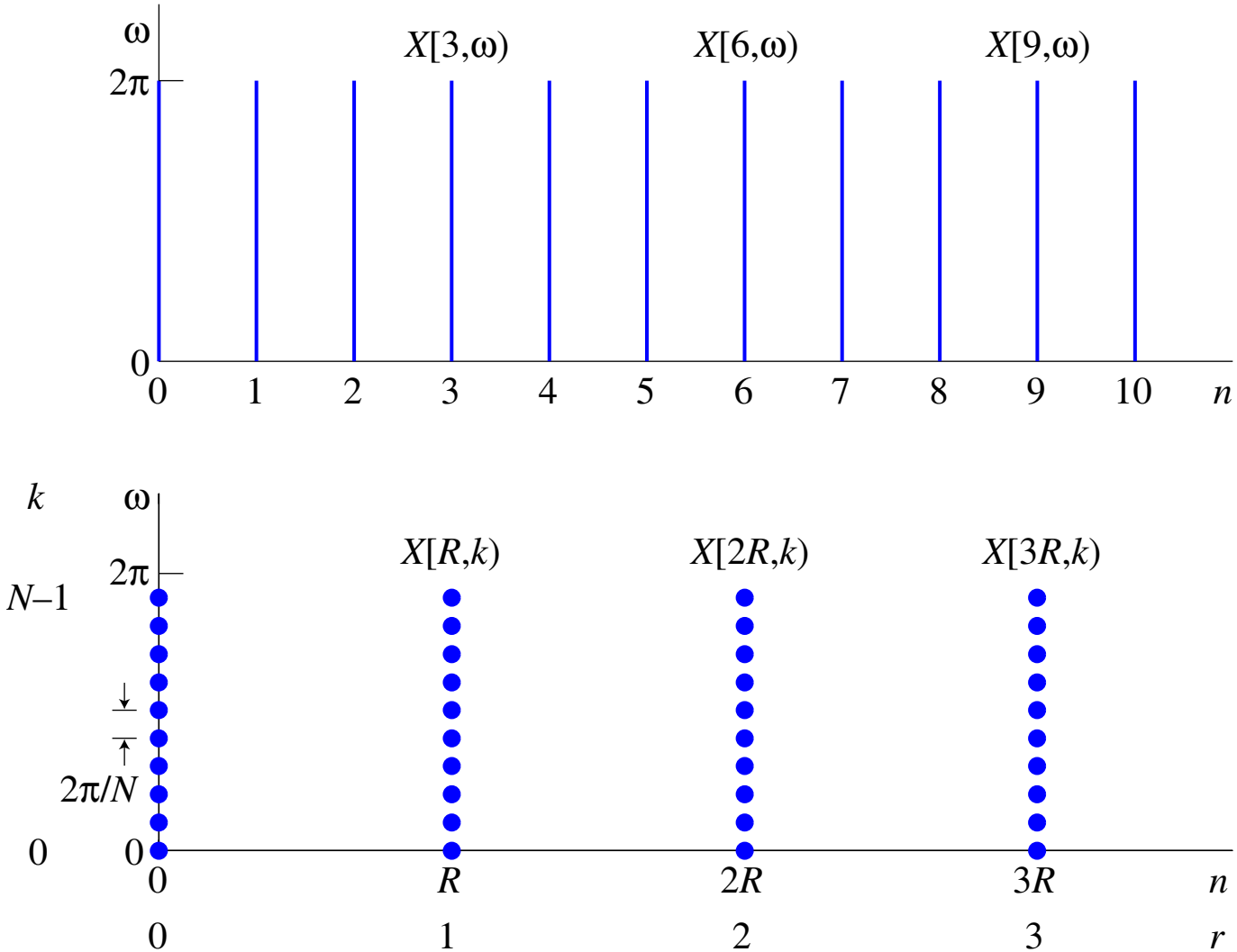
It is also unnecessary to evaluate the STFT or discrete STFT at every time sample n; we can still reconstruct the original sequence if $X[n,\omega)$ or $X[n,k]$ is sampled every R time samples:

$$\begin{aligned} X[rR,k] &= X[rR,2\pi k/N) \\ &= \sum_{m=0}^{L-1} x[rR+m]\, w[m]\, e^{-j(2\pi/N)km}, \end{aligned}$$

where r and k are integers such that $-\infty < r < \infty$ and $0 \leq k \leq N-1$, if $N \geq L \geq R$.

The condition $R \leq L$ ensures that all samples $x[n]$ are included in the discrete STFT for some r. If $R = L$, then the signal will be broken up into non-overlapping contiguous *frames* indexed by r. If $R < L$, then the frames will overlap.

Region of support for $X[n,\omega)$ (top panel) and grid of sampling points (bottom panel) for $X[rR,k]$ with $N = 10$ and $R = 3$:



Discrete STFT Analysis of Speech Signals:

Speech is produced by excitation of the *vocal tract*, which extends from the *glottis* in the larynx to the *lips*.

One way of classifying speech sounds is according to the excitation source:

- *Voiced sounds* (e.g., a, e, i, o, u, m, n) are produced by quasi-periodic pulsing of the glottis.
- *Fricative sounds* (e.g., f, s, sh, ch) are produced by noise-like turbulence created at a constriction of the vocal tract.
- *Plosive sounds* (e.g., p, k, t) are produced by completing closing the vocal tract to build up air pressure behind the closure, and then abruptly releasing the pressure to generate a single impulse-like airflow.

It is also possible to combine voicing with the other two sound sources $\Rightarrow$ *voiced fricatives* (e.g, v, z) and *voiced plosives* (e.g., b, g, d).

Conceptual model of speech production:

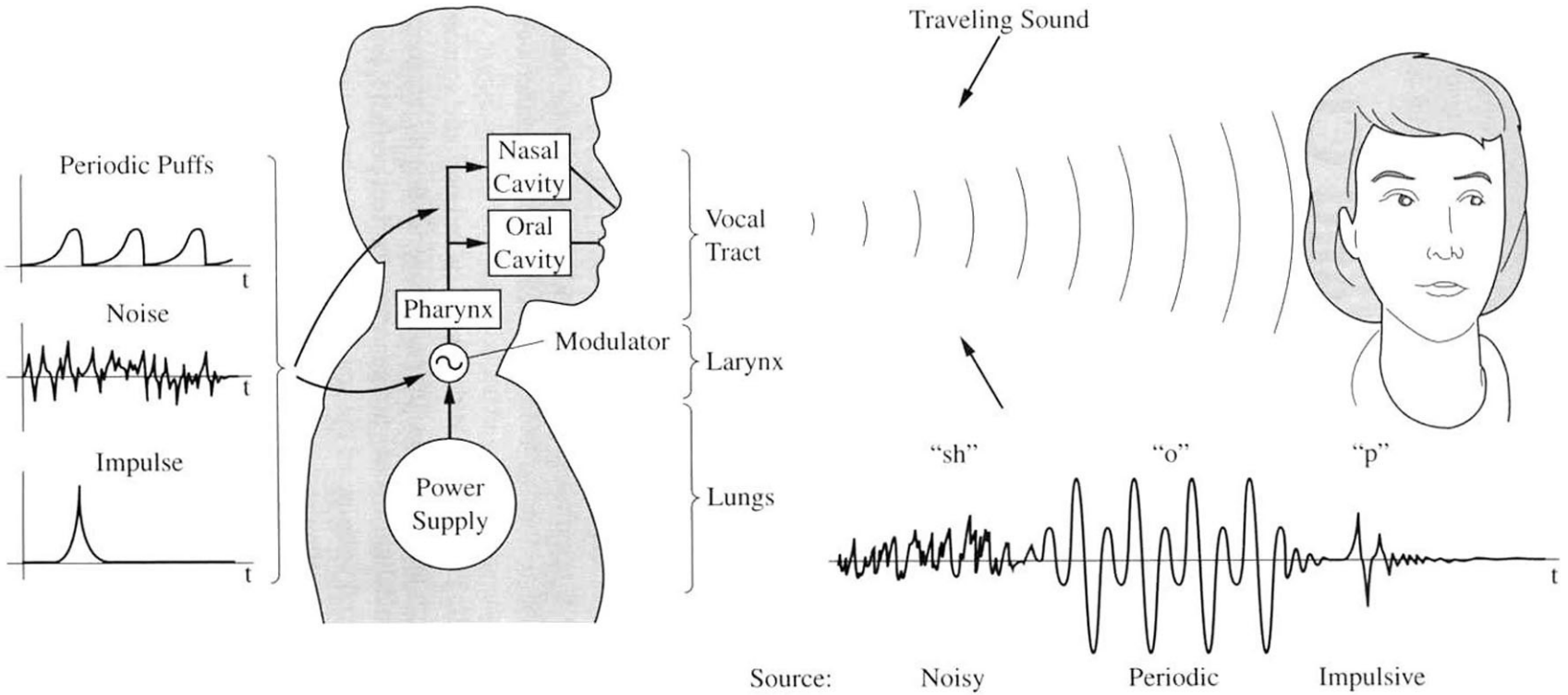


Figure 3.1 Simple view of speech production. The sound sources are idealized as periodic, impulsive, or (white) noise and can occur in the larynx or vocal tract.

(Quatieri)

Examples of speech sounds with different excitation sources:

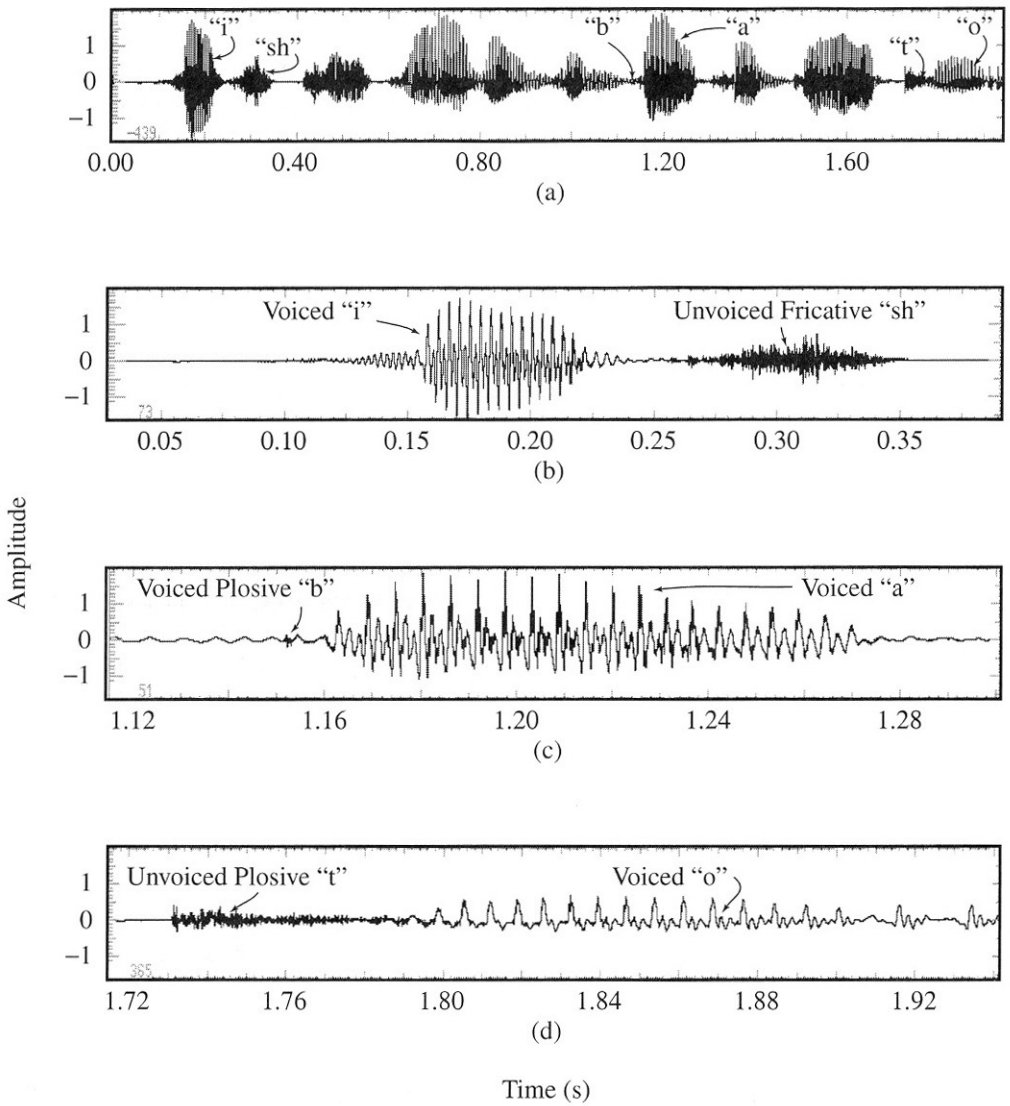


Figure 3.13 Examples of voiced, fricative, and plosive sounds in the sentence, "Which tea party did Baker go to?": (a) speech waveform; (b)–(d) magnified voiced, fricative, and plosive sounds from (a). (Note the "sh" is a component of an affricate to be studied in Section 3.4.6.)

(Quatieri)

With a constant vocal tract shape, speech can be modeled as the response of an LTI filter (the vocal tract) to one of the particular excitation sources.

In natural speech, the vocal tract changes shape relatively slowly over time as the throat, tongue and lips perform the gestures of speech, and consequently it can be viewed as a slowly time-varying filter that imposes its frequency response properties on the spectrum of the excitation source.

The spectrogram, a graphical display of the *magnitude* of the time-varying discrete STFT, is given by:

$$S[n,k] = |X[n,k]|^2 ,$$

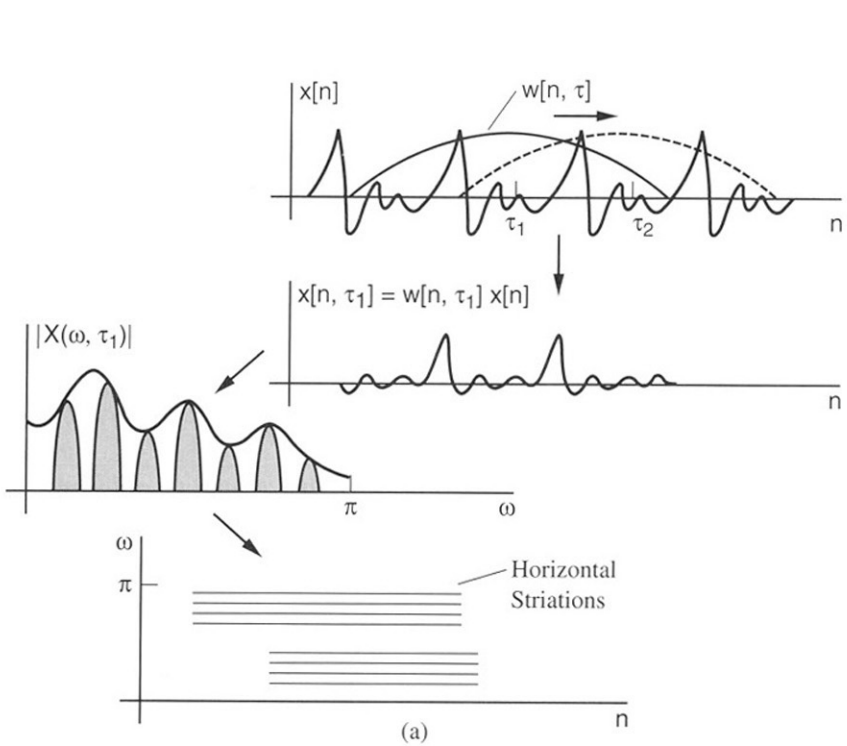
or

$$S_{\text{dB}}[n,k] = 20 \log_{10} |X[n,k]| \quad \text{in dB.}$$

The wideband spectrogram has a “short” window with a duration less than one pitch period of voiced speech (i.e., < 10 ms for male speakers). Consequently, it has very good temporal resolution, such that the temporal dynamics of short speech sounds (e.g., unvoiced plosives) are well defined, but poor frequency resolution, such that the harmonics in voiced sounds are unresolved. However, the periodicity in voicing appears as vertical striations and the vocal tract resonances (formants) appear as greater-magnitude (e.g., darker) regions on the spectrogram.

The narrowband spectrogram has a “long” window with a duration of several pitch periods of voiced speech (typically 20–40 ms). Consequently, it has very good frequency resolution, such that the harmonics in voiced sounds are resolved and appear as horizontal striations in the spectrogram, but poor temporal resolution, such that the spectra of transient speech sounds are smeared over time.

Formation of the narrowband and wideband spectrograms:



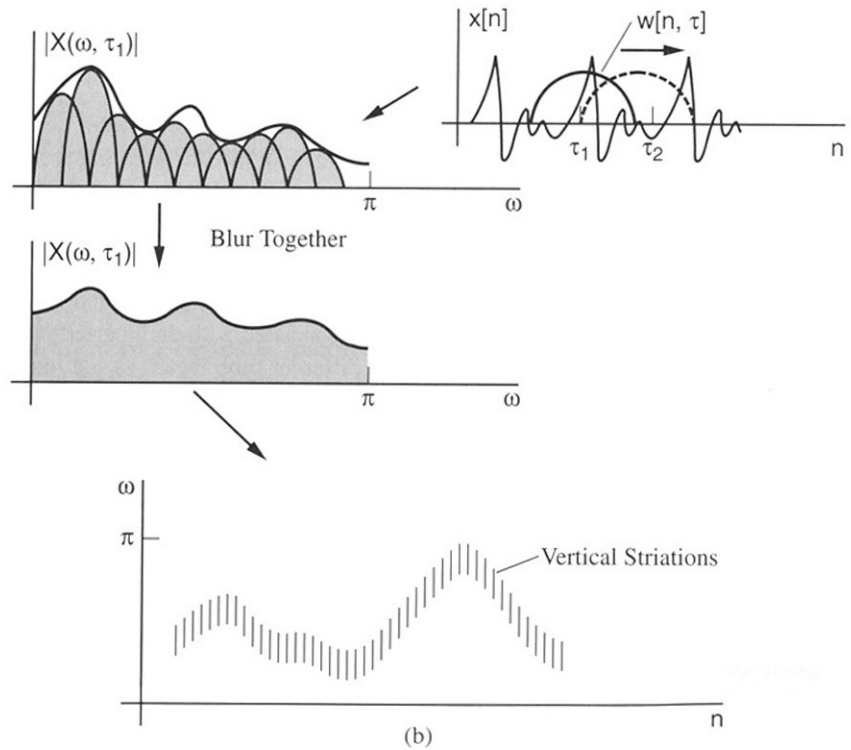
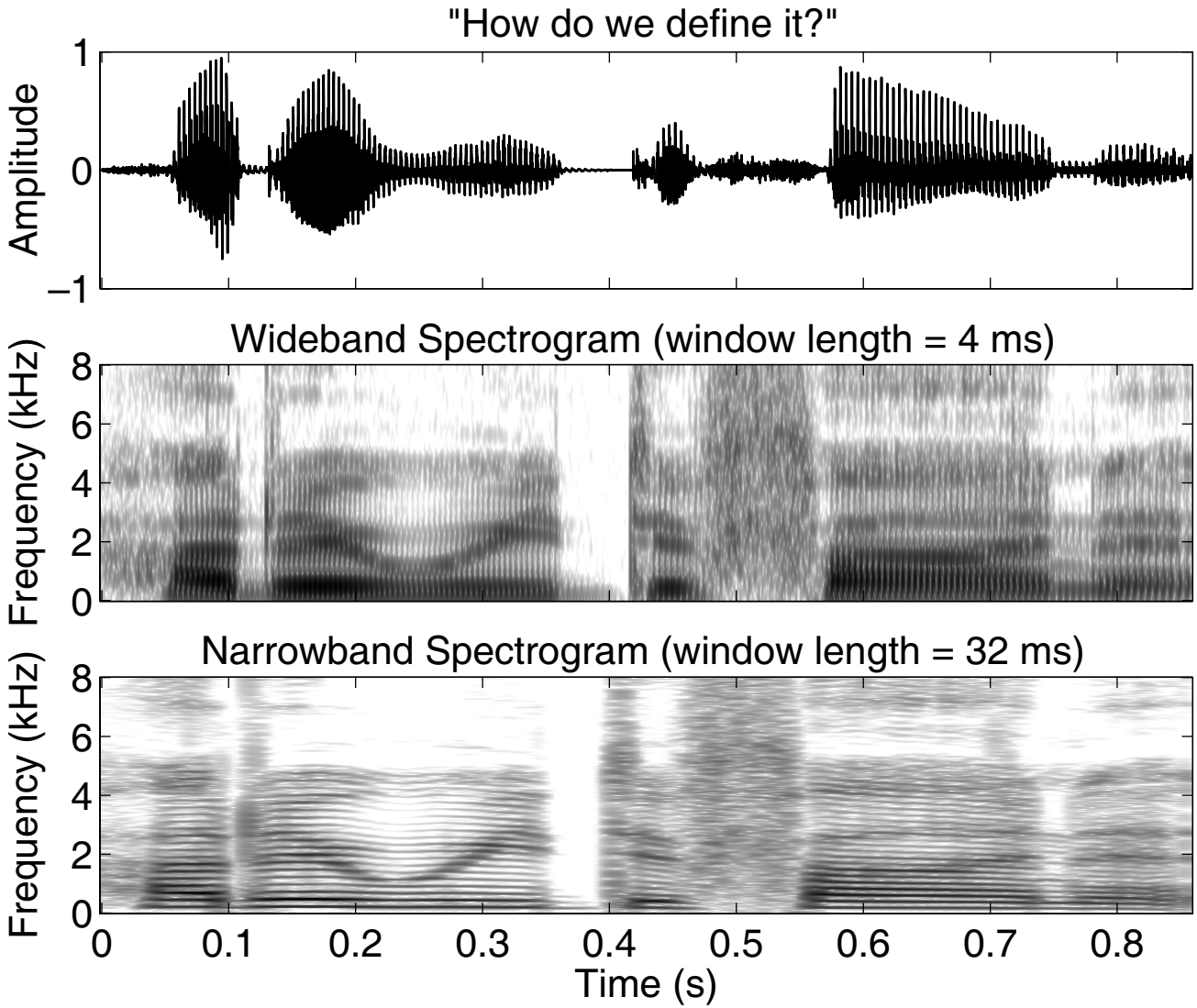


Figure 3.14 Formation of (a) the narrowband and (b) the wideband spectrograms.

(Quatieri)

Example:


"How do we define it?"
Amplitude
1
0
−1
Wideband Spectrogram (window length = 4 ms)
Frequency (kHz)
8
6
4
2
0
Narrowband Spectrogram (window length = 32 ms)
Frequency (kHz)
8
6
4
2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Time (s)