ECE 797:

Speech and Audio Processing

Hand-out for Lecture #10

Thursday, March 25, 2004

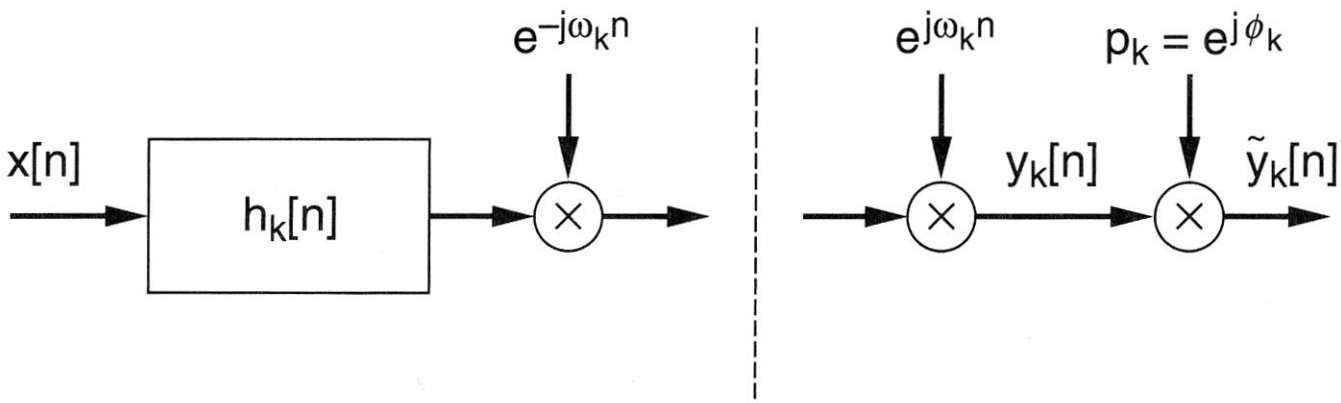


Figure 8.1 Phase adjustment factor of kth channel in FBS synthesis.

SOURCE: L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* [42]. ©1978, Pearson Education, Inc. Used by permission.

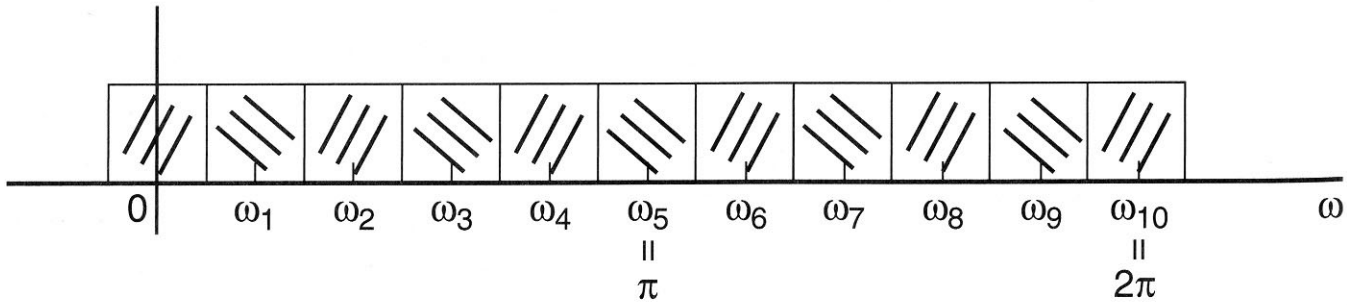


Figure 8.2 Filters whose center frequencies are symmetric about π, i.e., $\omega_{N-k} = 2\pi - \omega_k$ with $\omega_k = \frac{2\pi}{N}k$. In this example $N = 10$.

SOURCE: L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* [42]. ©1978, Pearson Education, Inc. Used by permission.

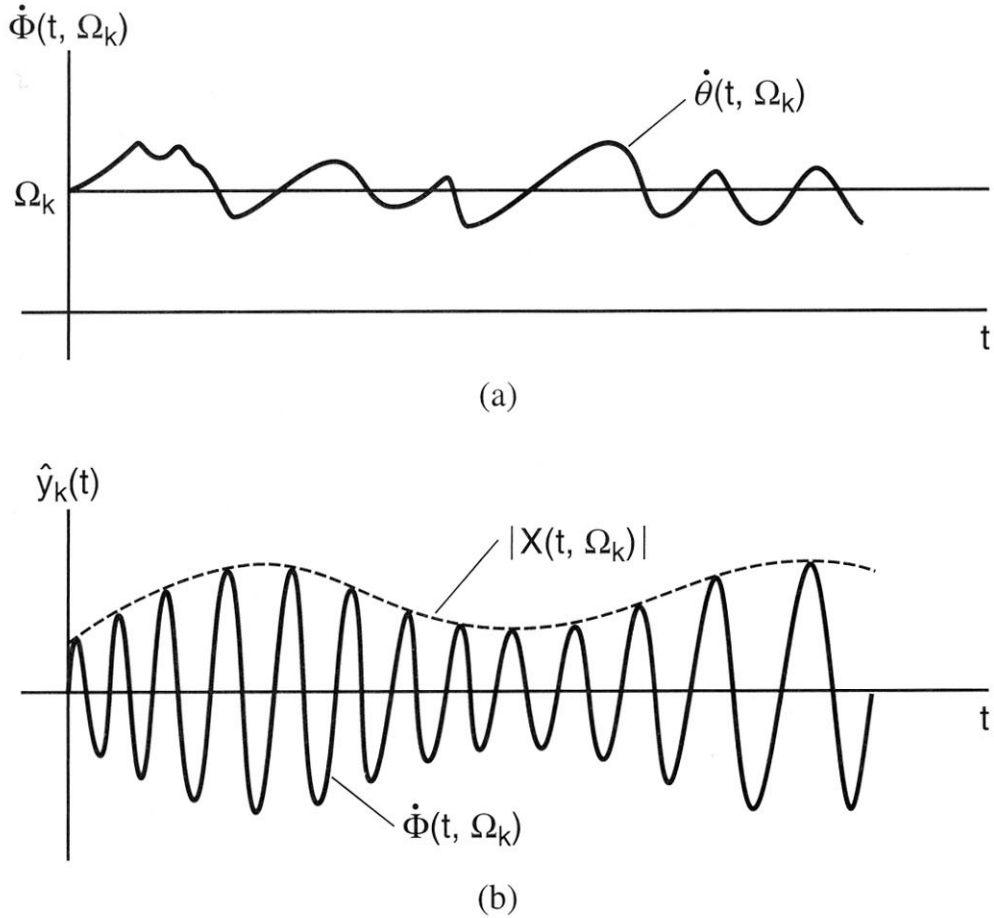


Figure 8.3 Interpretation of instantaneous frequency and amplitude in continuous time: (a) $\dot{\theta}(t, \Omega_k)$ is the deviation of the instantaneous frequency from the center frequency Ω_k of the kth filter (the frequency modulation) and $\dot{\Phi}(t, \Omega_k) = \Omega_k + \dot{\theta}(t, \Omega)$ is the instantaneous frequency; (b) the instantaneous amplitude $2|X(t, \Omega_k)|$ and instantaneous frequency $\dot{\Phi}(t, \Omega_k)$ characterize each filter bank sinewave output.

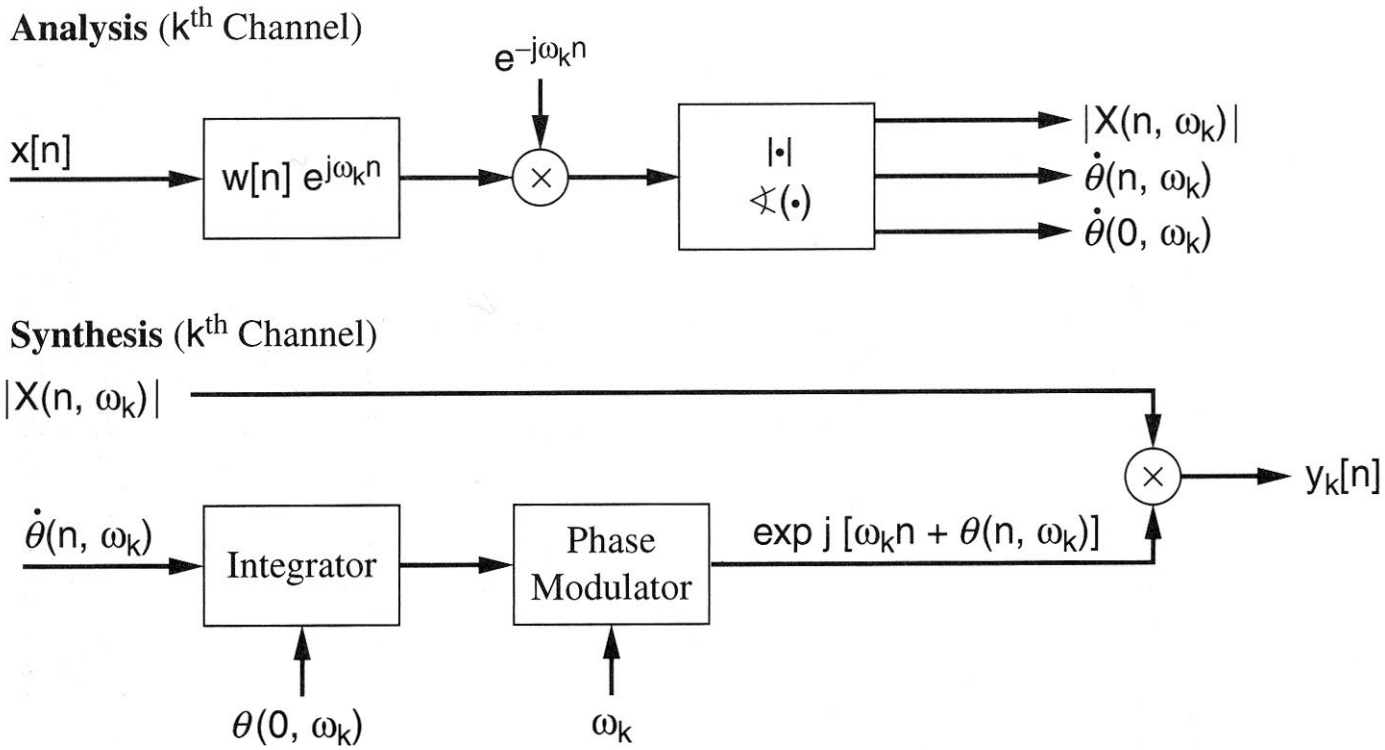


Figure 8.6 Analysis/synthesis structure in the phase vocoder.

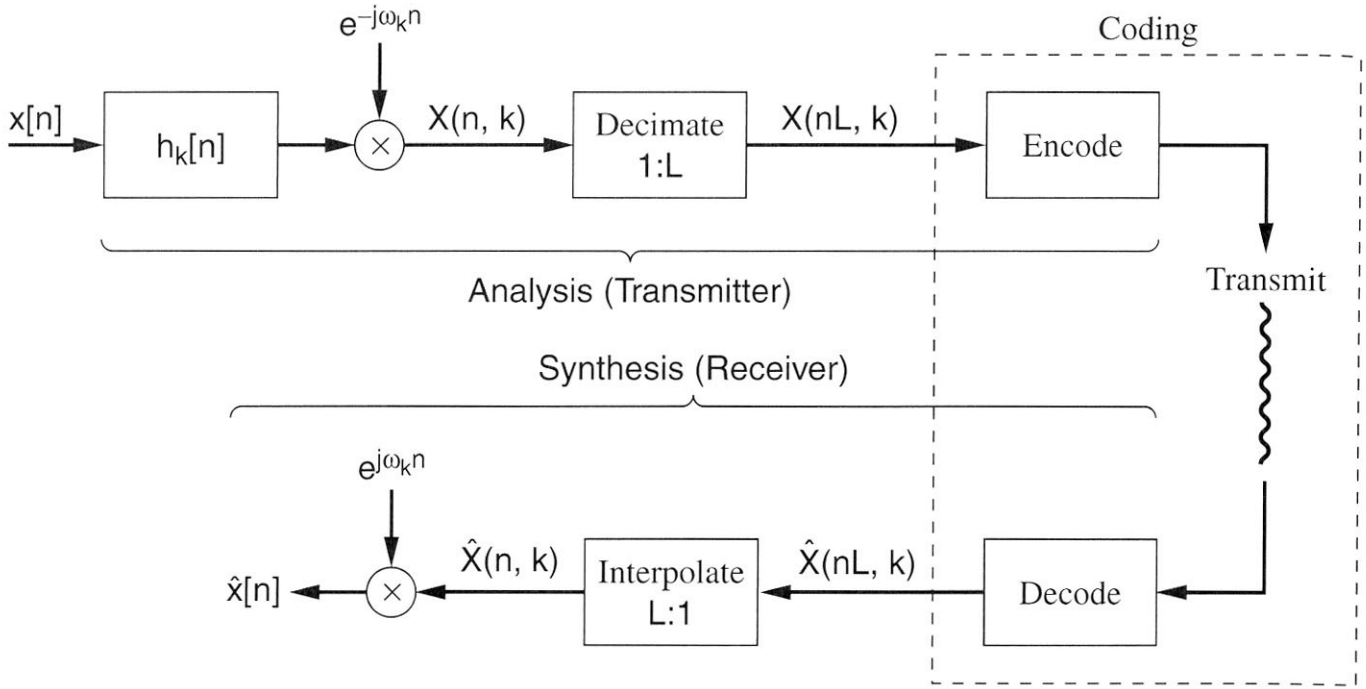


Figure 8.7 Filter-bank-based speech coder overview. The time decimation is limited by the bandwidth of each analysis filter, according to the Nyquist criterion, to avoid aliasing in frequency. In addition, the frequency decimation is limited by the duration of the analysis filter, i.e., the number of filters must be large enough to avoid aliasing in time. The number of samples/s over all filter bank channels (in time and frequency) is consequently larger than the input waveform sampling rate.

SOURCE: L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals* [42]. ©1978, Pearson Education, Inc. Used by permission.

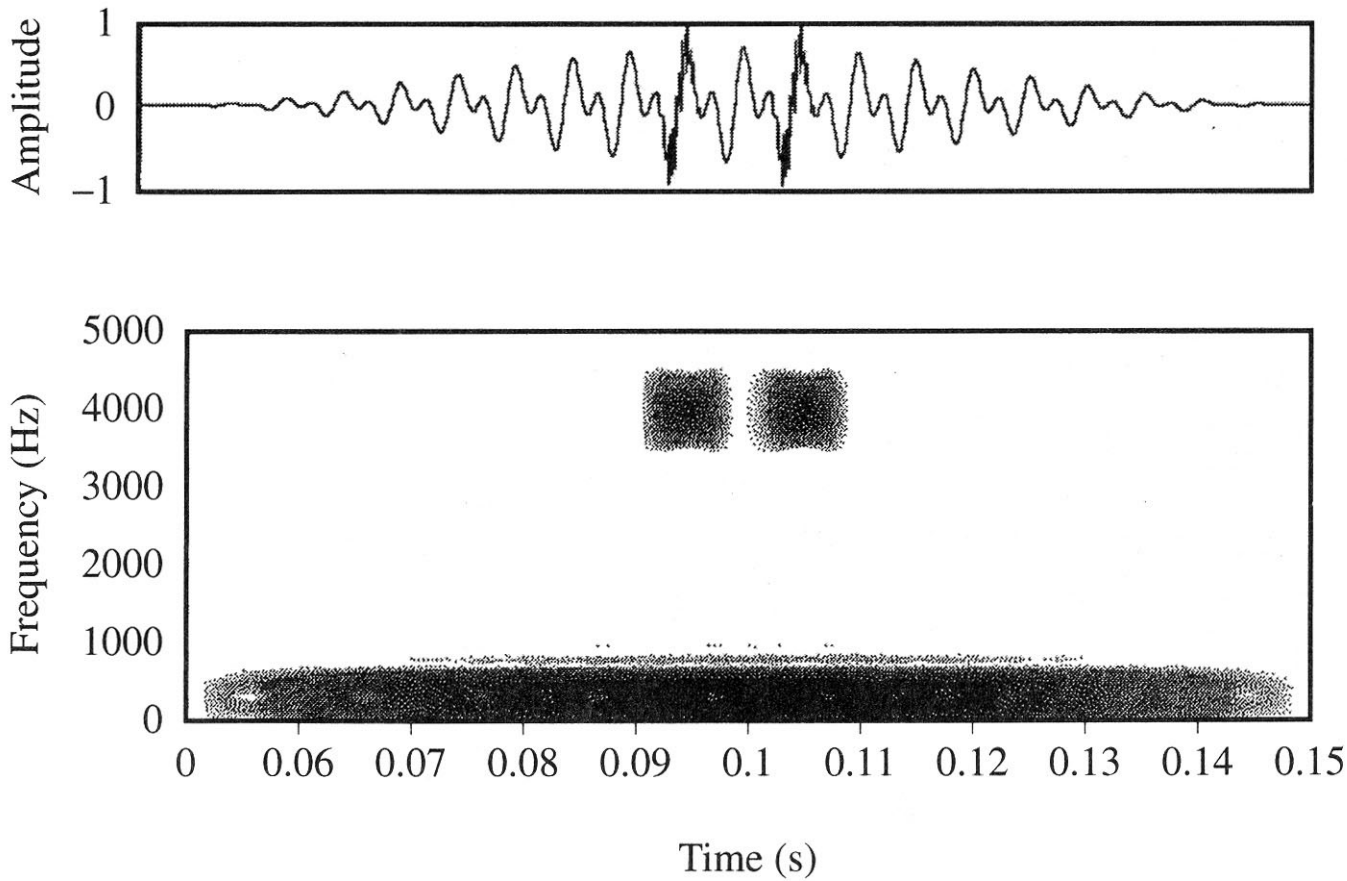


Figure 8.12 Low-frequency signal with superimposed high-frequency bursts: (a) waveform; (b) wideband spectrogram of (a).

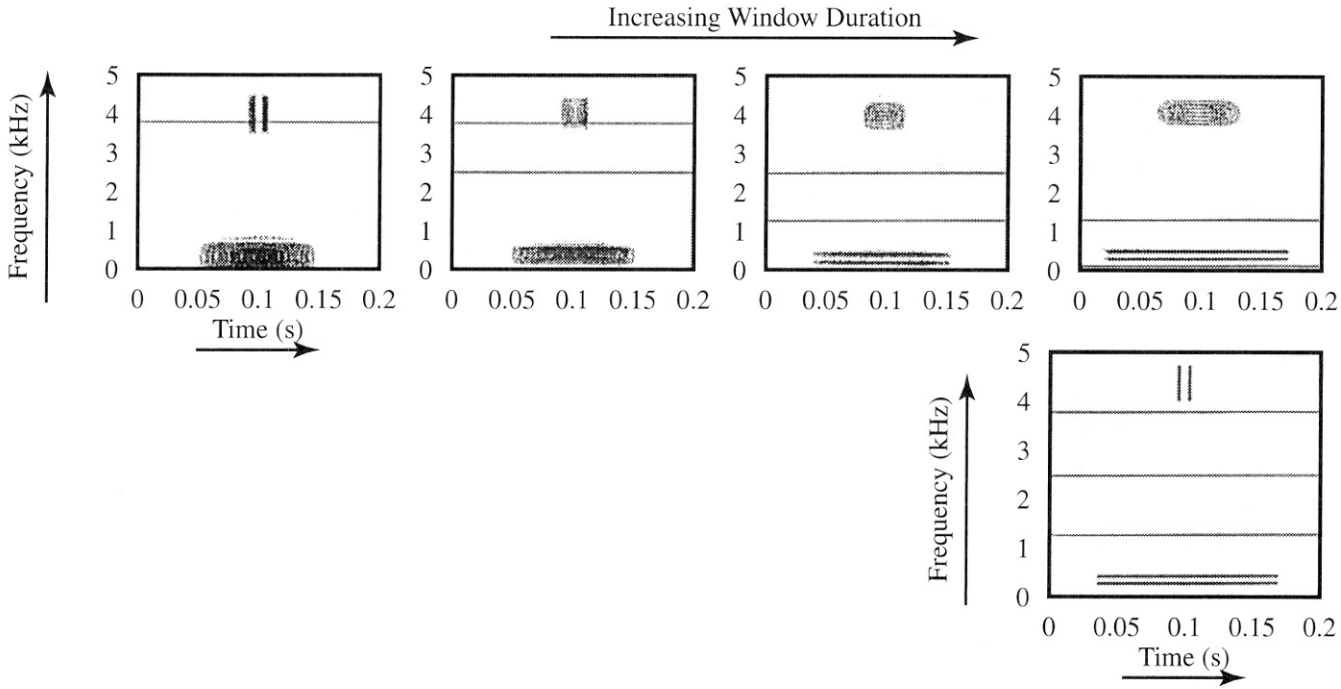


Figure 8.13 The wavelet transform as a collage of spectrograms for the multi-component signal of Figure 8.12a. A short window at high frequency gives good time resolution, while a long window at low frequency gives good frequency resolution. The lower right panel is obtained by piecing together regions of the upper four panels.

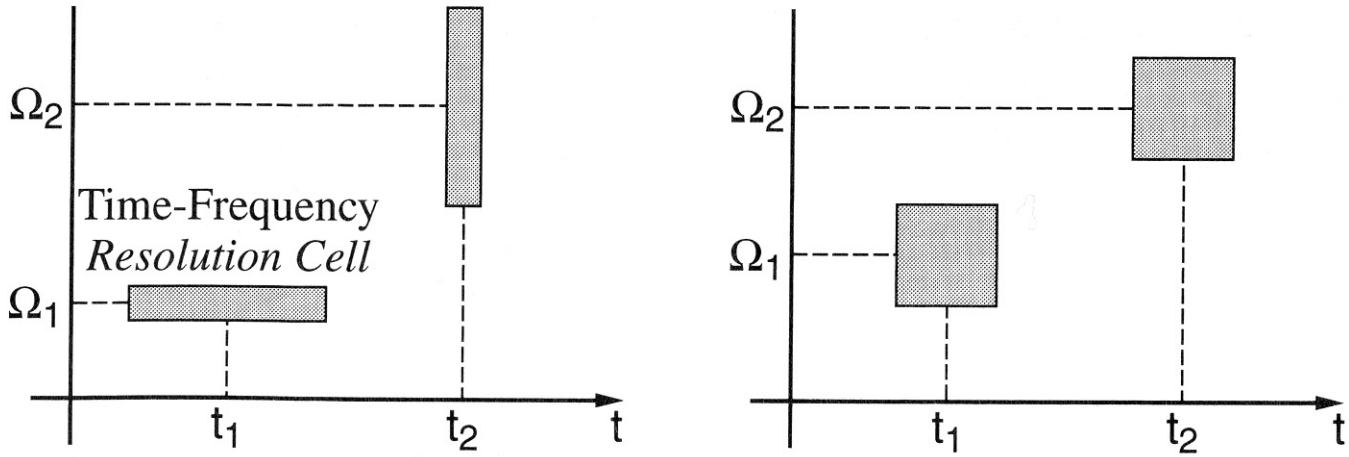


Figure 8.14 Adaptation of window size to frequency in the wavelet transform (left panel) in contrast to a fixed window in the STFT (right panel).

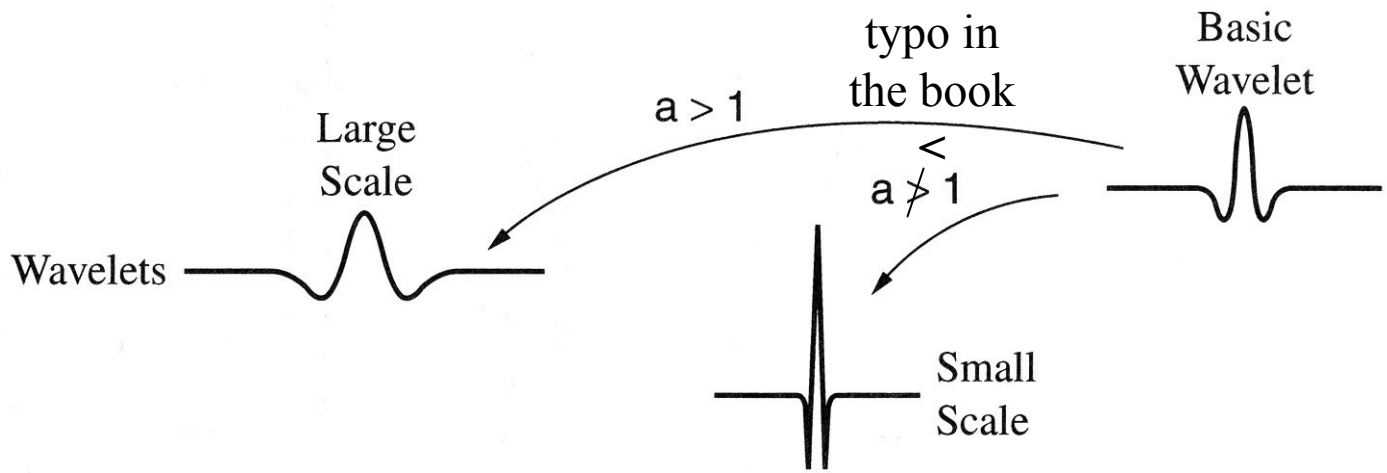


Figure 8.15 Schematic of a basic wavelet and its associated wavelets at different scales.

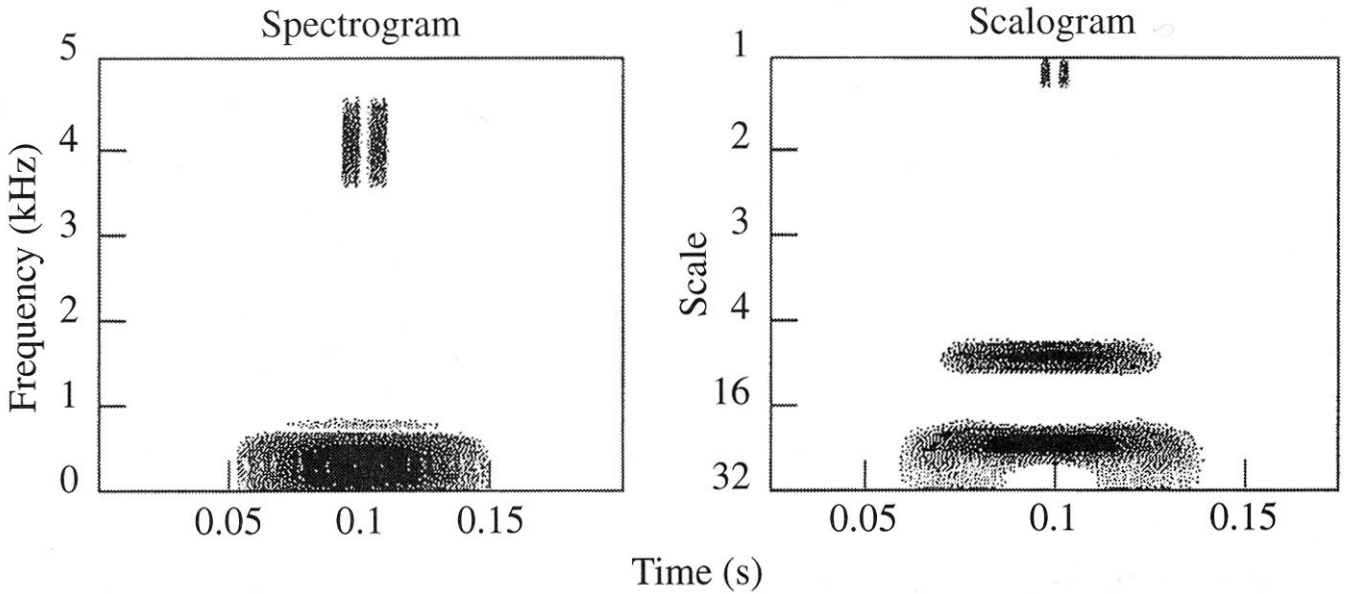


Figure 8.16 Comparison of the spectrogram $|X(\tau, \omega)|^2$ and scalogram $|X_w(\tau, a)|^2$ for the multi-component signal of Figure 8.12.

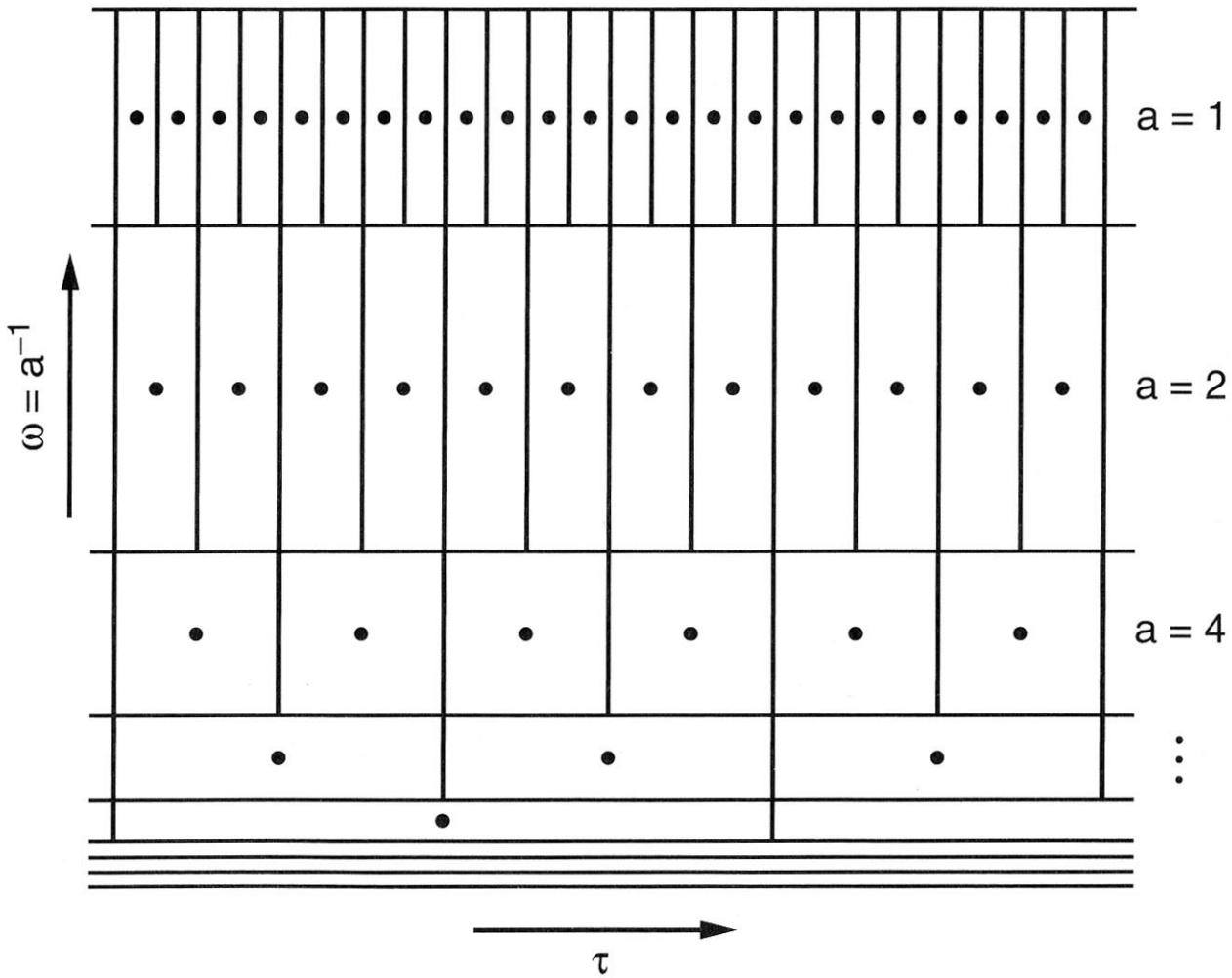


Figure 8.17 Sampling of scale and shift in a dyadic wavelet basis. The wavelets are partitioned in octave bands and the time shift is commensurate with bandwidth: $a_m = 1, 2, 4, \ldots 2^m \ldots$ and $\tau_n = n a_m$.

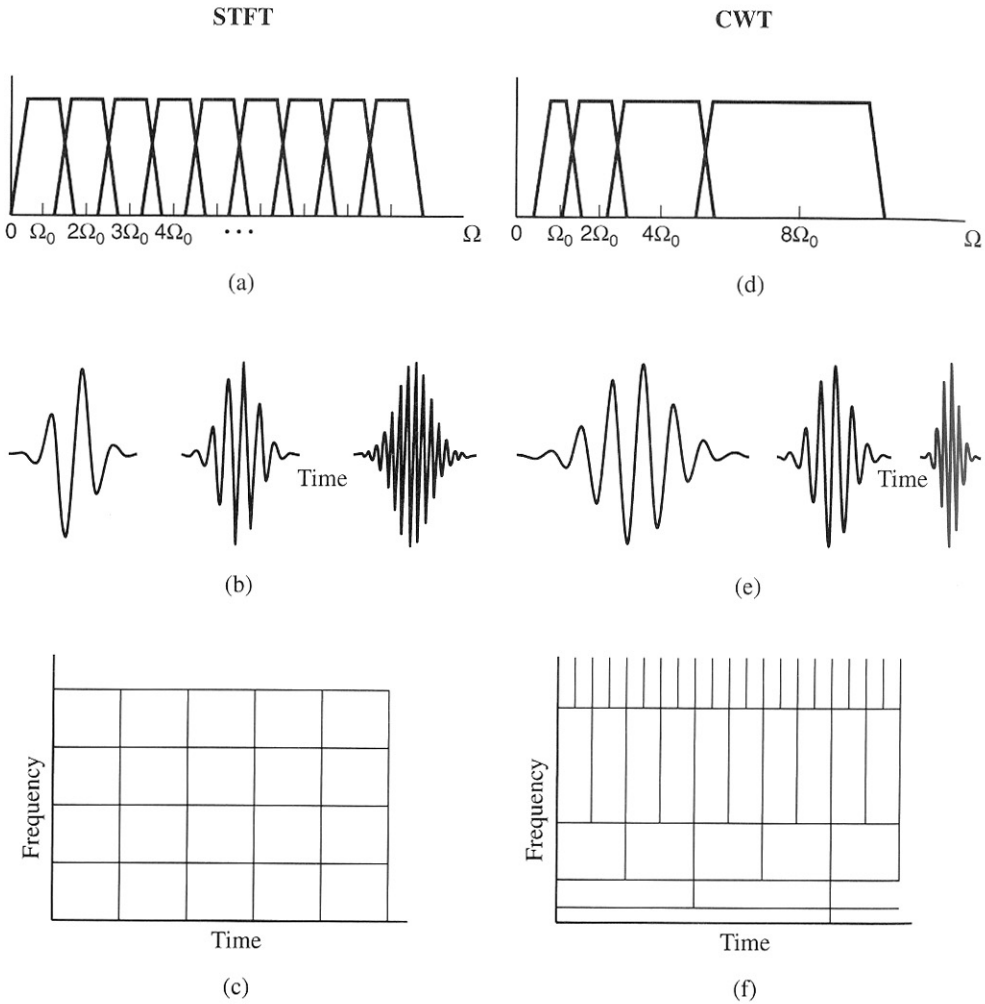


Figure 8.18 Comparison of the sampling requirements for the discrete STFT and the discrete dyadic wavelet transform from a filtering perspective. Panels (a) and (d) show the required filters in frequency, while (b) and (e) show their counterparts in time. The discrete STFT filters have constant bandwidth while the discrete dyadic wavelets have constant-Q bandwidth. Panels (c) and (f) give the respective time-frequency "tiles" that represent the essential concentration of the basis in the time-frequency plane.

SOURCE: O. Rioul and M. Vetterli, "Wavelets and Signal Processing" [43]. ©1991, IEEE. Used by permission.

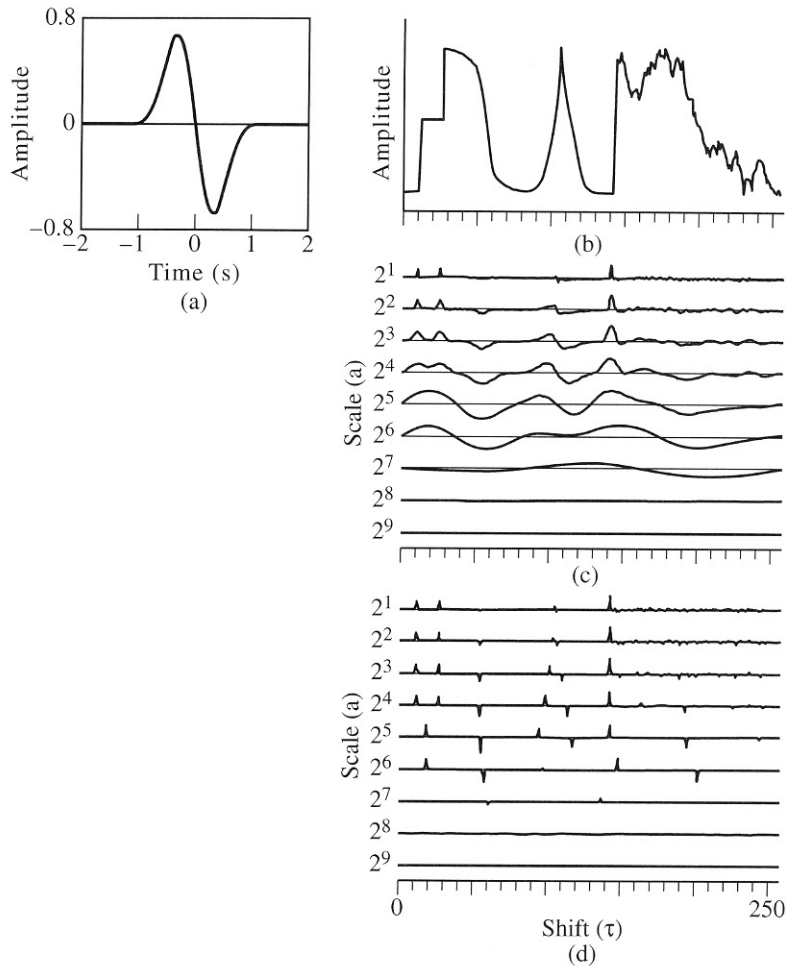
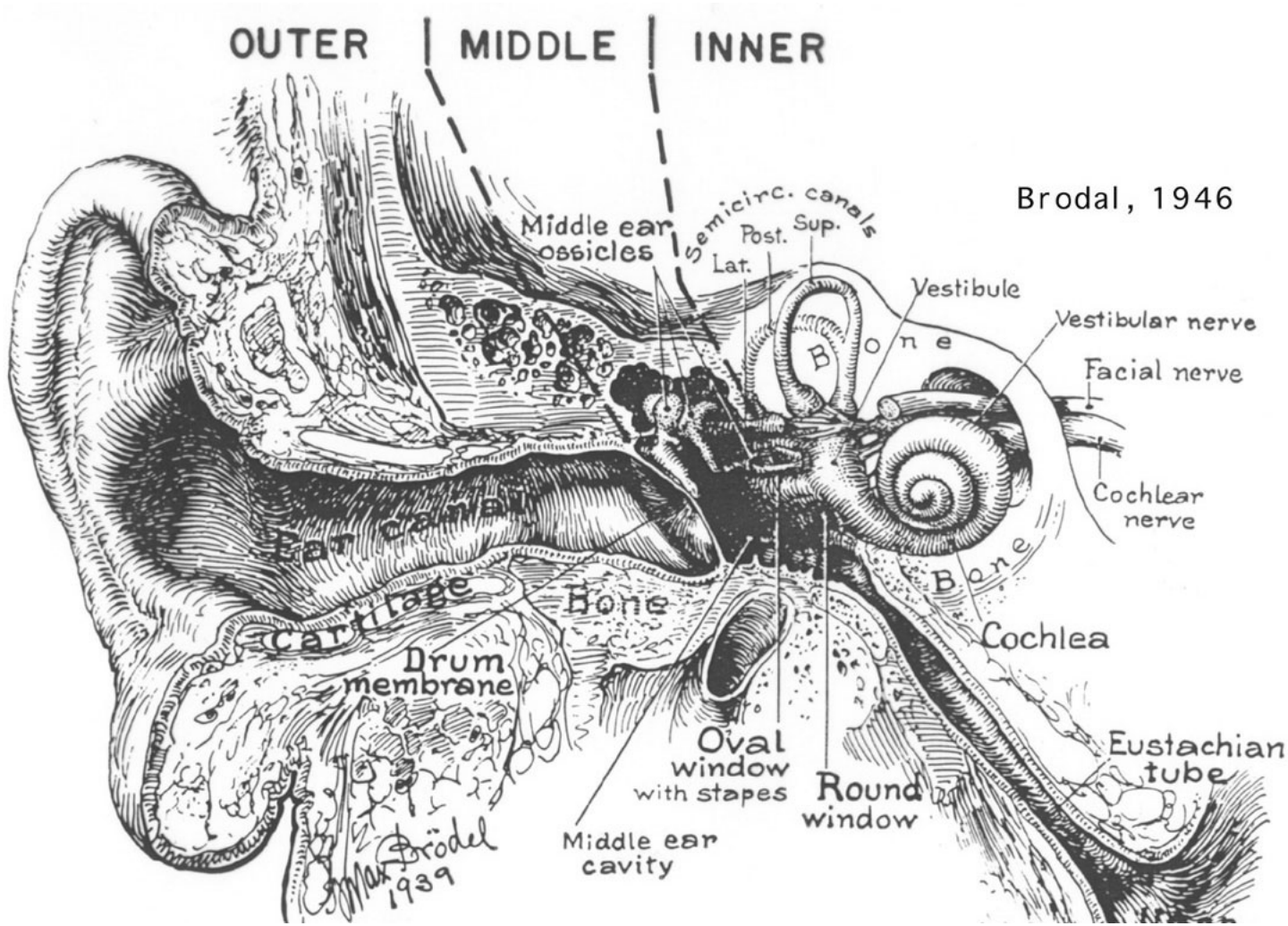
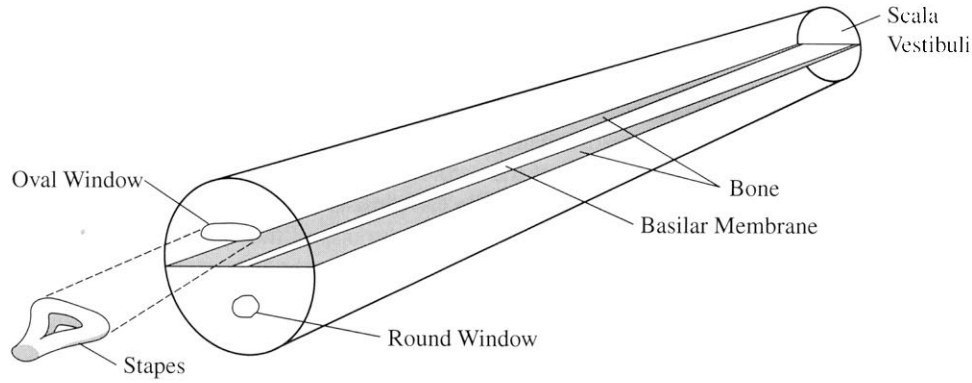


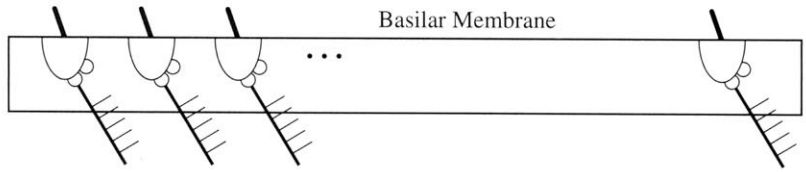
Figure 8.19 Signal representation by maxima of the discrete wavelet transform with respect to shift: (a) a wavelet chosen to approximate a differentiator; (b) a signal and its superimposed (essentially indistinguishable) reconstruction from wavelet maxima; (c) wavelet transform outputs $X_w(\tau, a)$ for sampled at a dyadic scale; (d) points of wavelet maxima of (c), i.e., $\max|X_w(\tau, a)|$ with respect to τ.

SOURCE: S. Mallat and W.L. Hwang, "Singularity Detection and Processing with Wavelets" [23]. ©1992, IEEE. Used by permission.

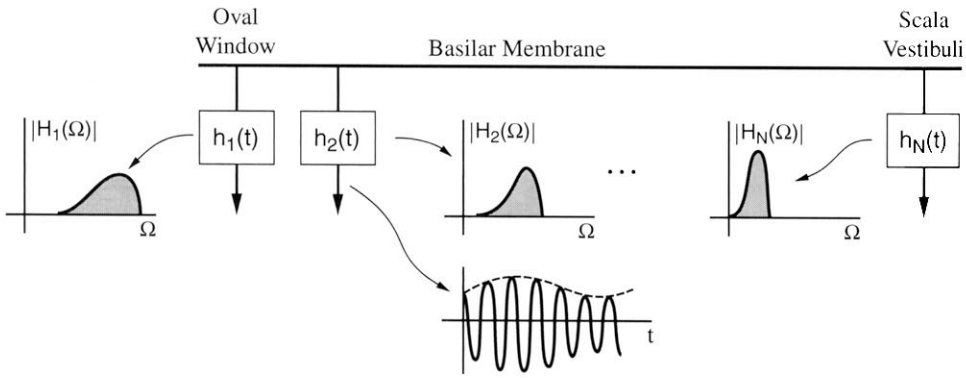

OUTER
MIDDLE
INNER
Brodal, 1946
Middle ear ossicles
Semicirc. canals
Post.
Sup.
Lat.
Vestibule
Vestibular nerve
Facial nerve
Cochlear nerve
Bone
Ear canal
Cartilage
Bone
Drum membrane
Cochlea
Eustachian tube
Oval window with stapes
Round window
Middle ear cavity
Max Brödel 1939



(a)



(b)



(c)

Figure 8.25 Schematic of front-end auditory processing and its model as a wavelet transform: (a) the uncoiled cochlea; (b) the transduction to neural firings of the deflection of hairs that protrude from the inner hair cells along the basilar membrane; (c) a signal processing abstraction of the cochlear filters along the basilar membrane. The filter tuning curves, i.e., frequency responses, are roughly constant-Q with bandwidth decreasing logarithmically from the oval window to the scala vestibuli.

SOURCE FOR PANEL (a): D.M. Green, *An Introduction to Hearing*, [13]. ©1976, D.M. Green. Used by permission.

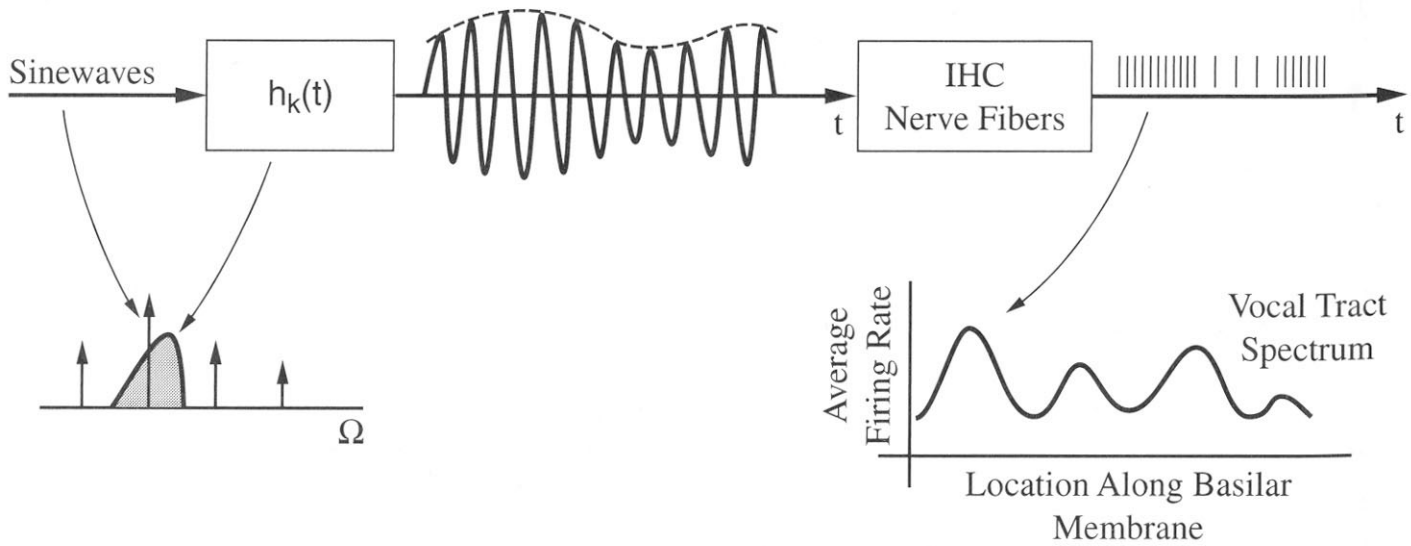


Figure 8.26 Auditory processing of a single slowly varying AM-FM sinewave as input to a cochlear filter, according to the place theory of hearing. The average firing rate is obtained by integrating the firing rate over many nerve fibers for a particular inner hair cell (IHC) and is roughly proportional to the input sinewave amplitude $A_p(t)$.

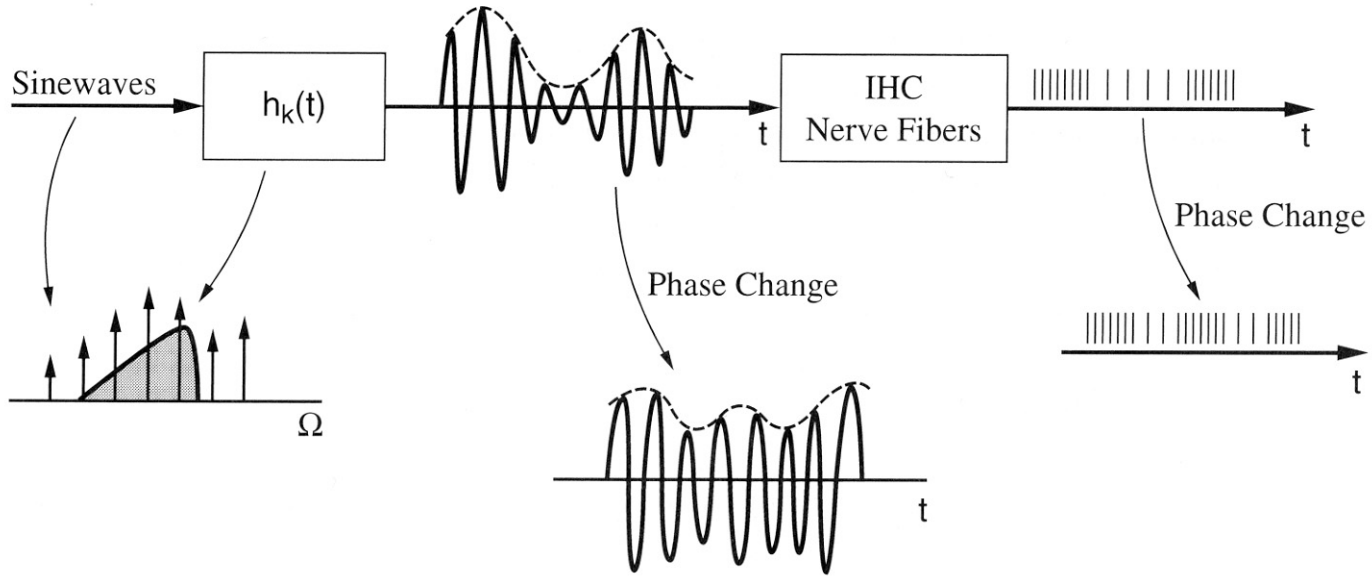
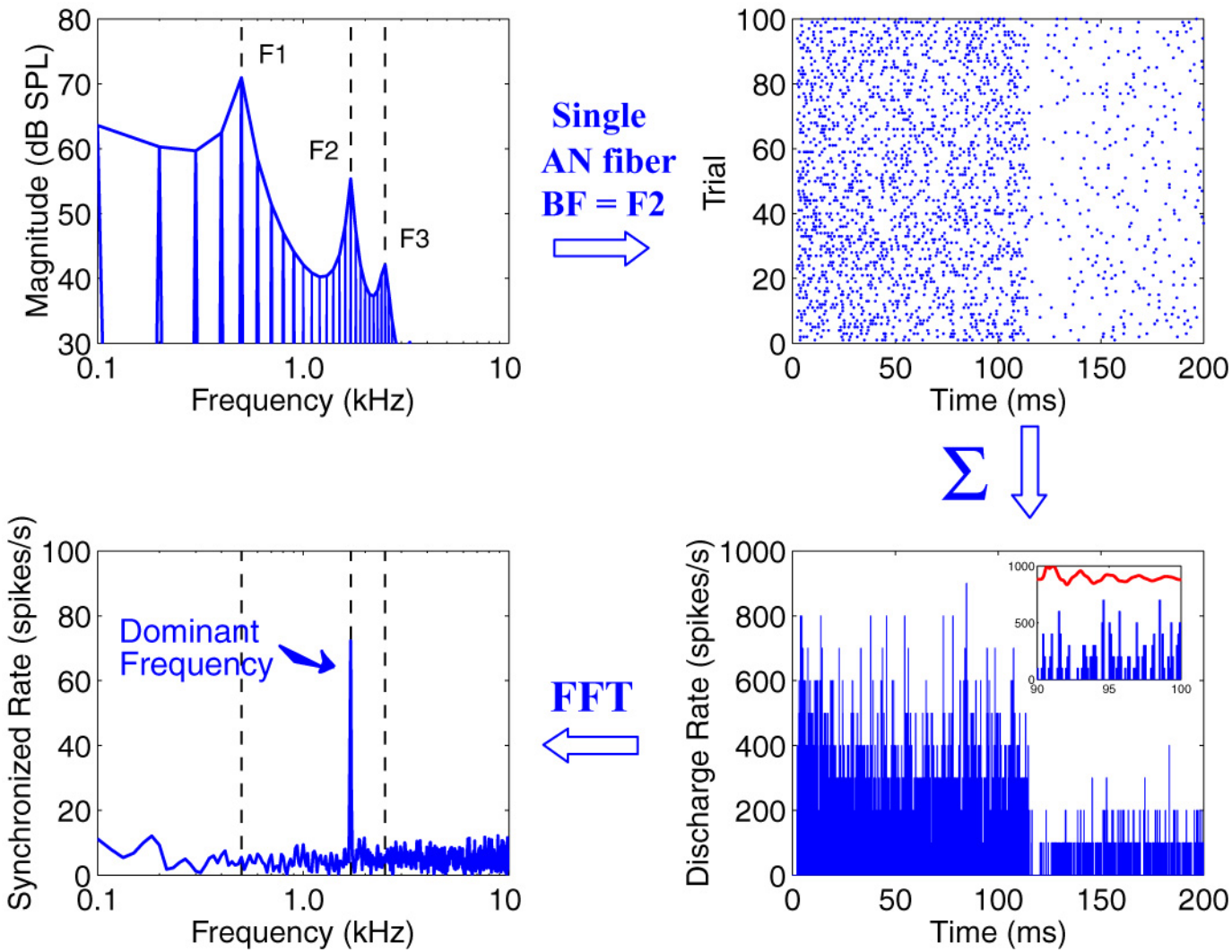
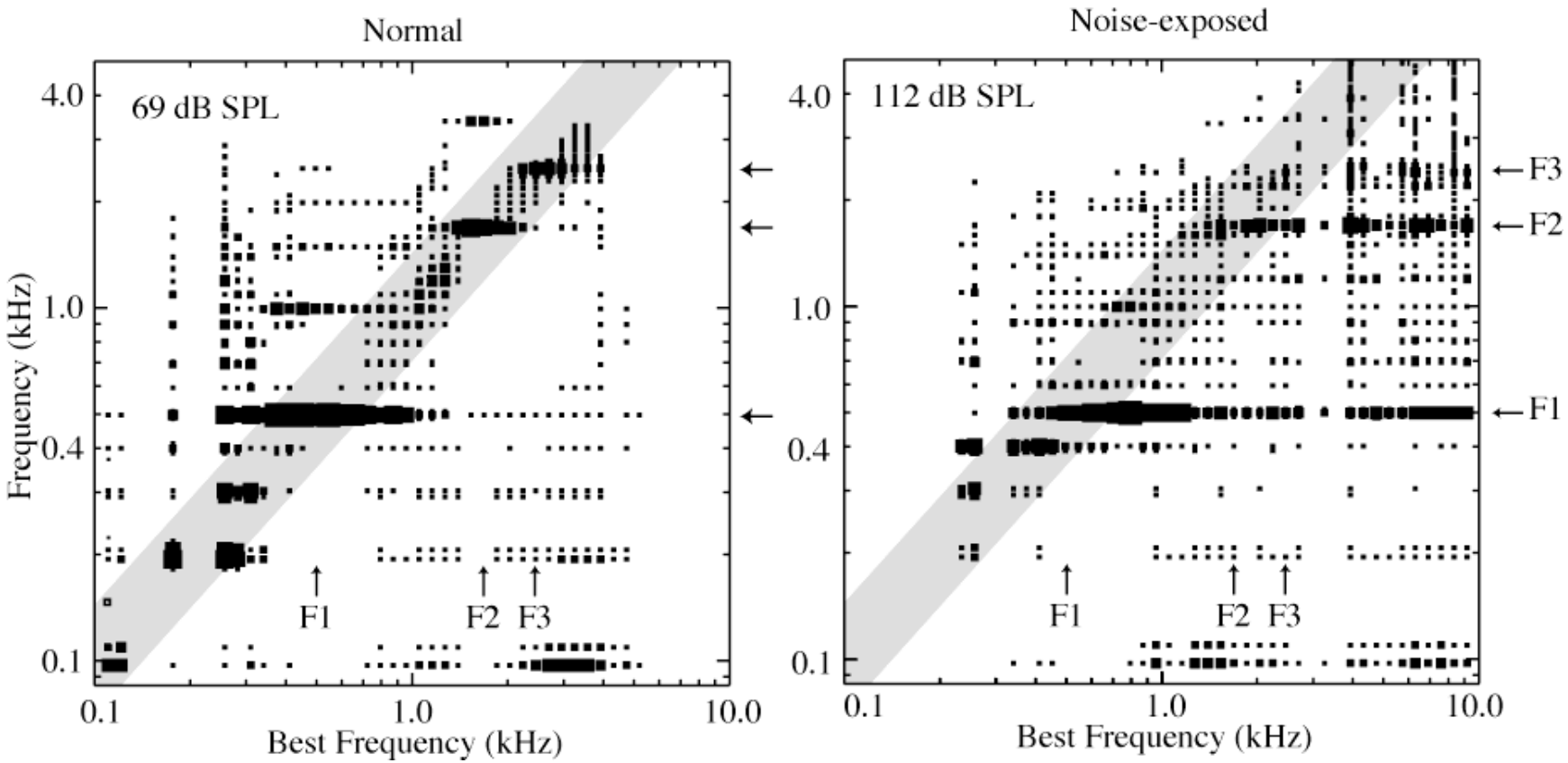


Figure 8.27 Auditory processing of a multiple of slowly varying AM-FM sinewaves as input to a cochlear filter, likely to occur with low pitch and at cochlear filters of high characteristic frequency. In this case, change in the phase relations of the input can alter the envelope shape and firing patterns of inner hair cell (IHC) nerve fibers, and thus, perhaps, perception of the sound.


Magnitude (dB SPL)
Frequency (kHz)
F1
F2
F3
Single
AN fiber
BF = F2
Trial
Time (ms)
Σ
Discharge Rate (spikes/s)
Time (ms)
FFT
Synchronized Rate (spikes/s)
Dominant
Frequency
Frequency (kHz)



(Schilling et al., 1998)

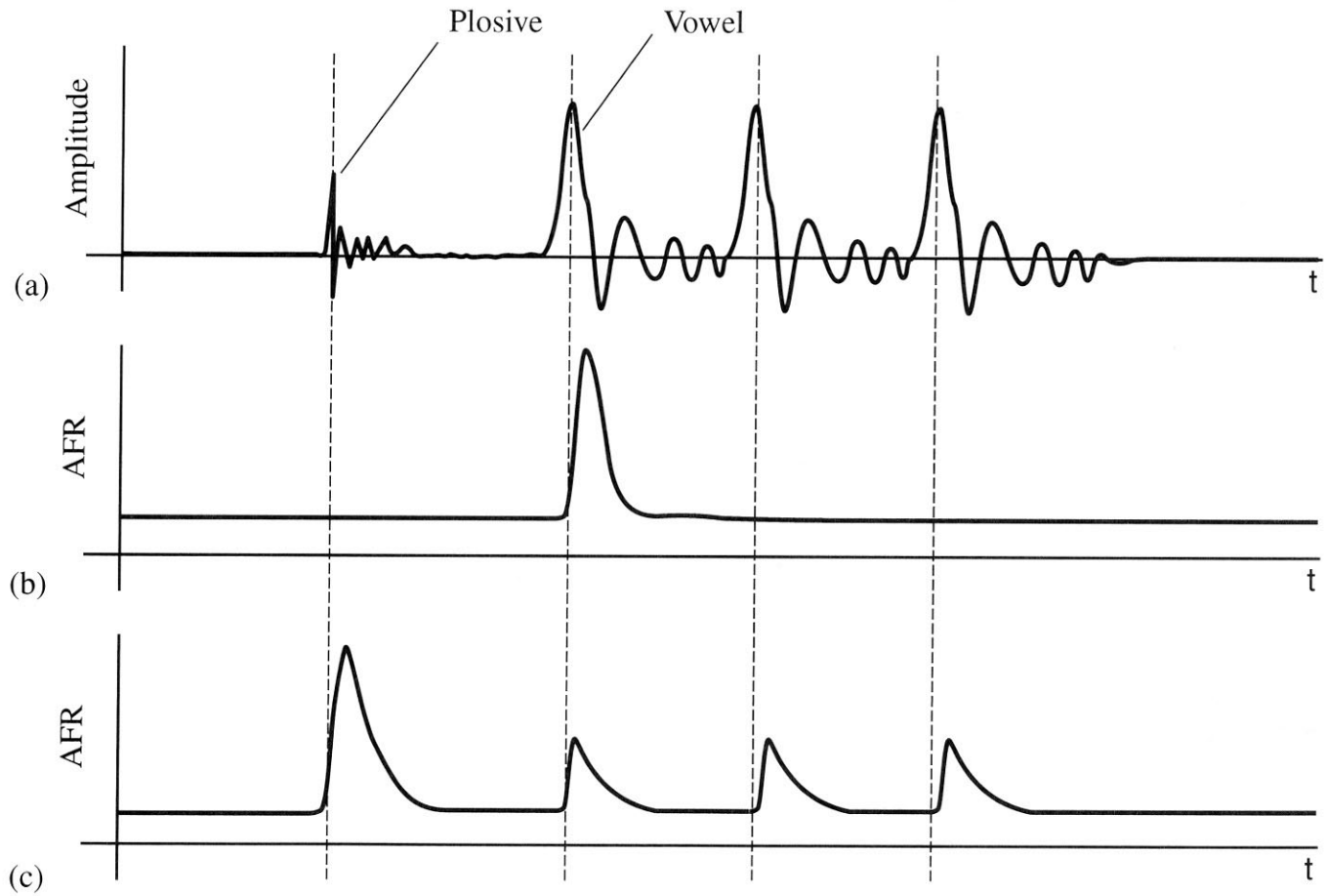


Figure 8.29 Schematized view of the phasic response of neural fibers in the auditory nerve to a plosive/vowel transition. Average firing rates (AFR) are illustrated for (b) a low-CF and (c) a high-CF channel of the auditory nerve for the speech waveform in panel (a). There is an average background discharge rate due to spontaneous emission of firings when no stimulus is present. In this schematic, the low-CF fiber responds to the vowel onset, while the high-CF fiber responds to the plosive and glottal pulse onsets. Observe that steady spectral components are suppressed in a phasic response.

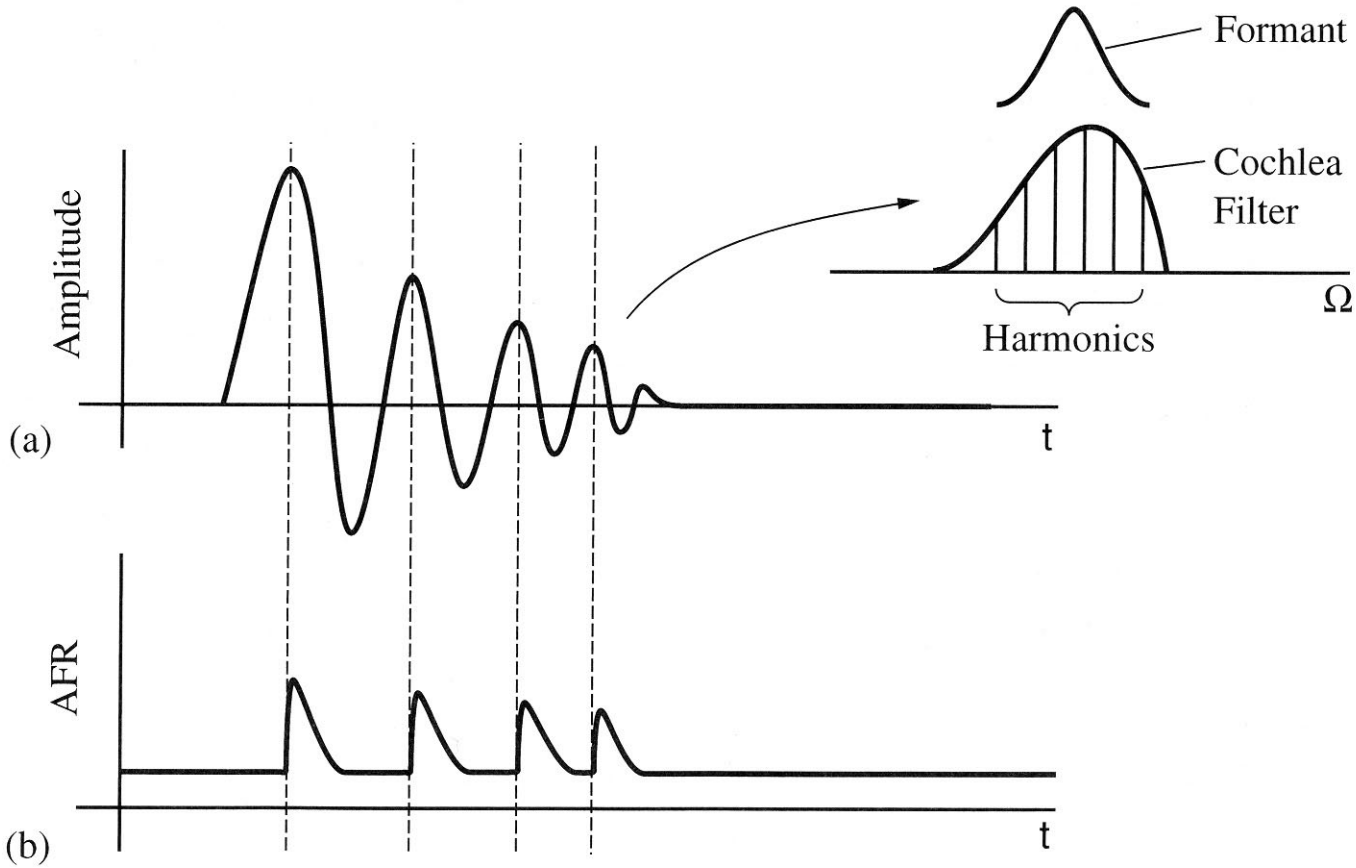


Figure 8.30 Schematic of average firing rate (AFR) [panel (b)] in estimation of changing formant frequency [panel (a)] by tonic auditory nerve component. Phase synchrony with respect to formant peaks allows fine time resolution.

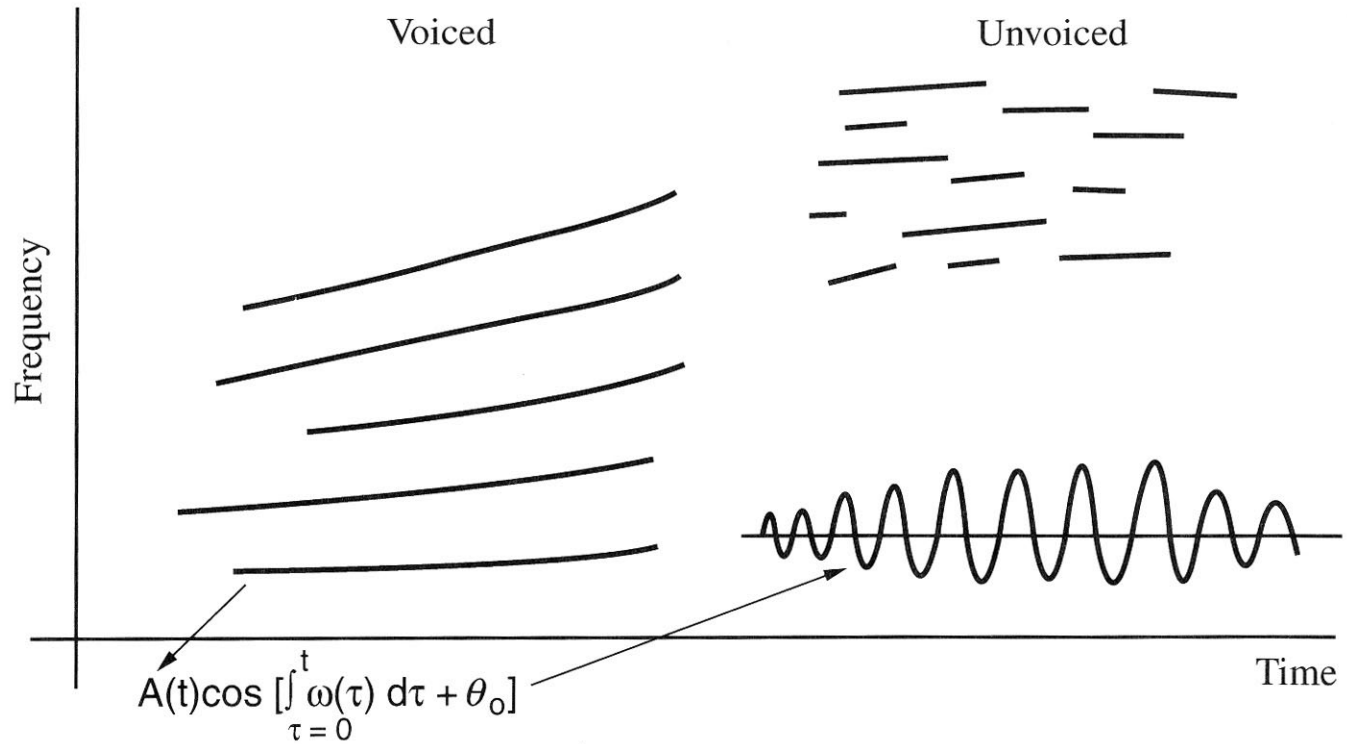


Figure 9.1 The sinewave model encompasses both voiced and unvoiced speech. Frequency tracks of voiced sinewaves are approximately harmonically related, while unvoiced tracks typically have no such relation and come and go randomly over short durations.

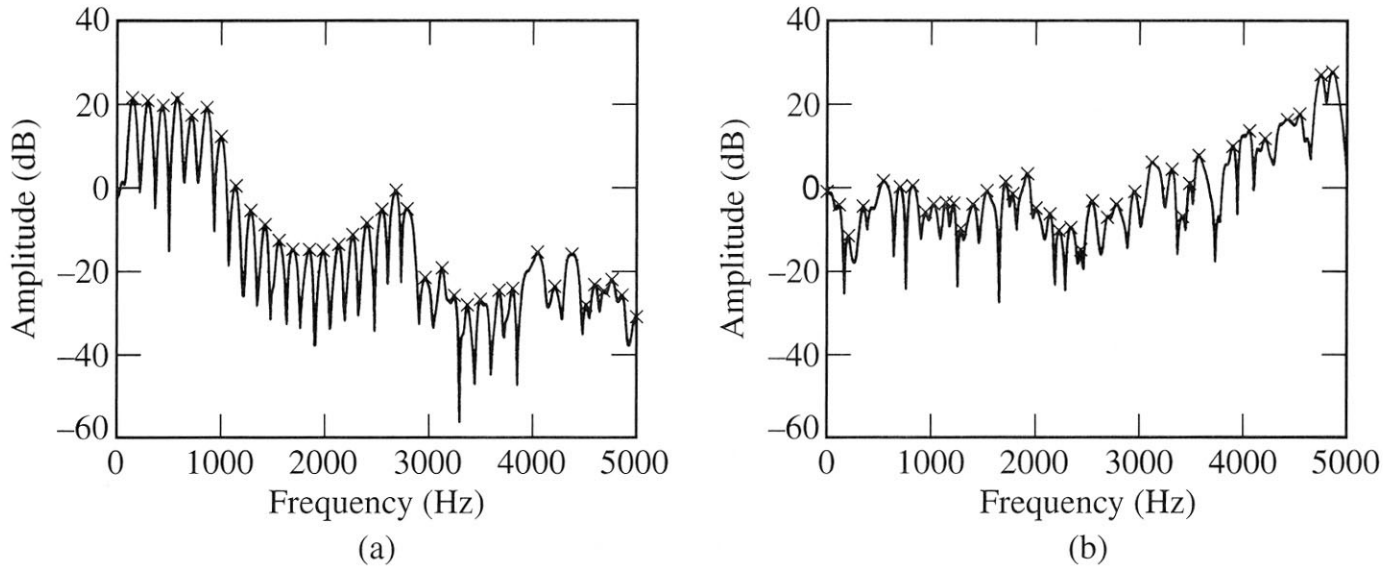


Figure 9.5 Typical STFT magnitude of voiced and unvoiced (fricative) speech: (a) voiced with an aspiration component; (b) unvoiced. Spectral peaks, whose locations are denoted by the crosses, determine which frequencies are selected to represent the speech waveform.

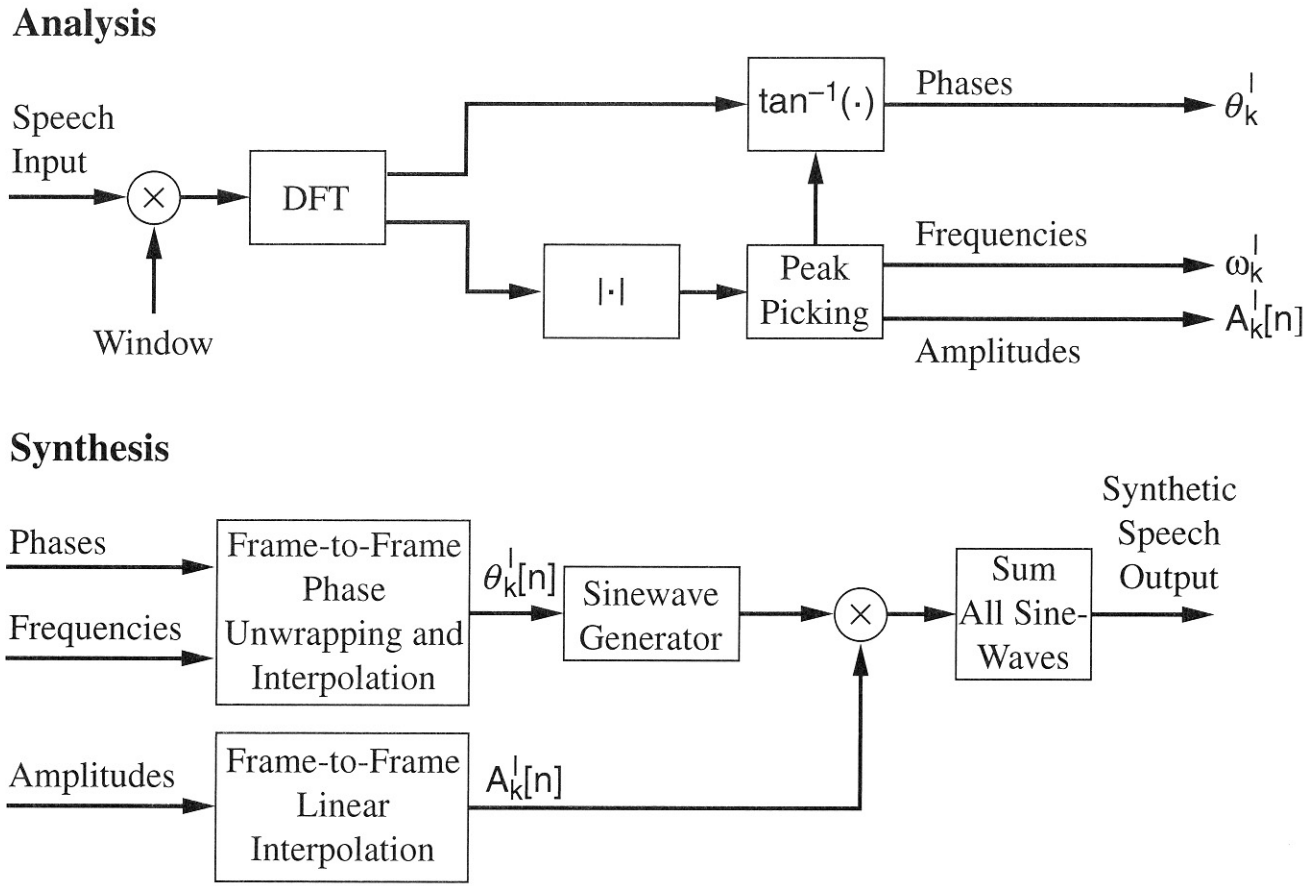


Figure 9.7 Block diagram of the baseline sinusoidal analysis/synthesis system.

SOURCE: R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation" [30]. ©1986, IEEE. Used by permission.

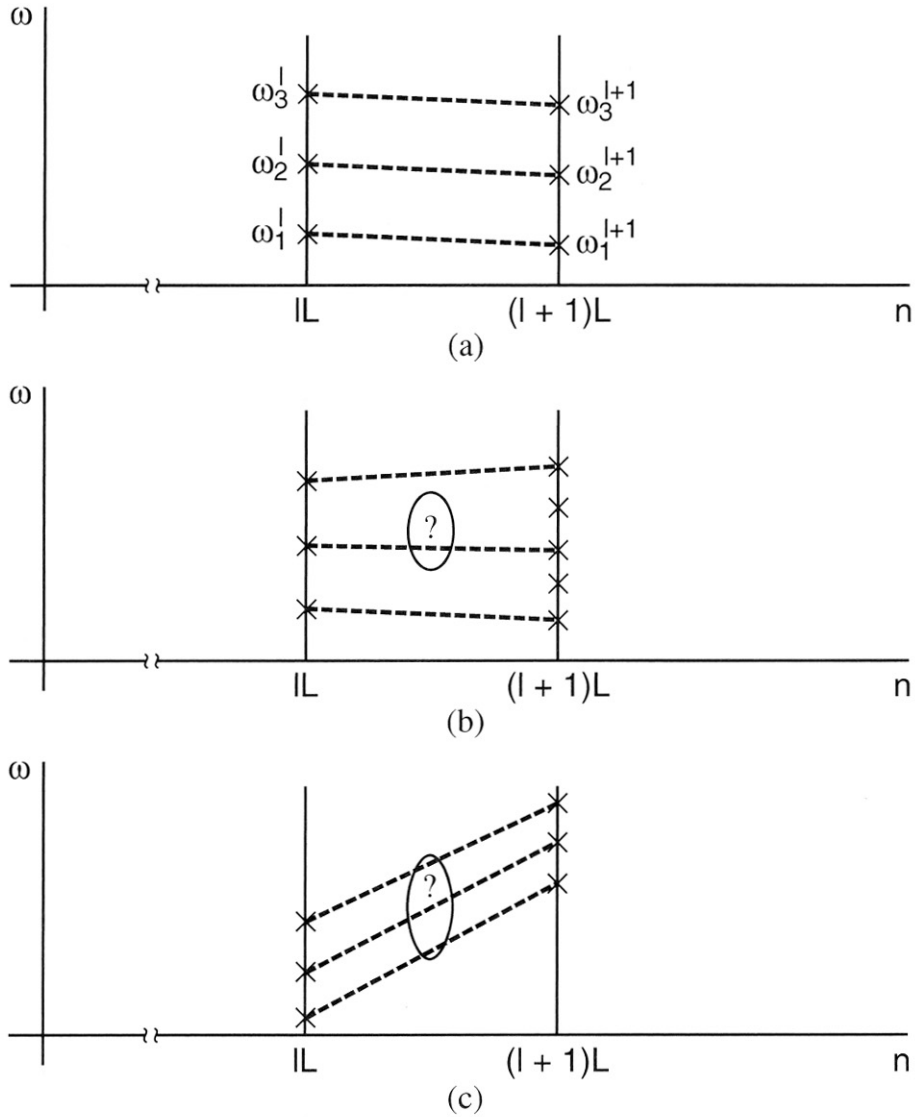


Figure 9.8 Problem of frequency matching: (a) slowly varying pitch; (b) rapidly varying pitch; (c) rapid voiced/unvoiced transition.

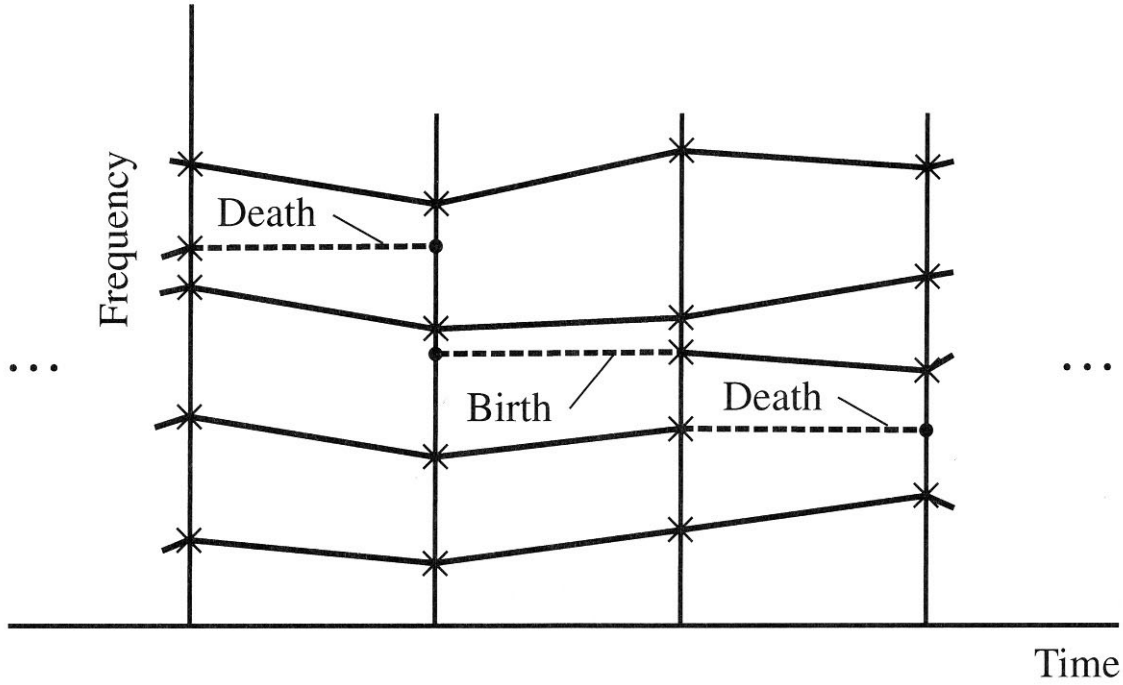


Figure 9.10 Different modes used in the birth/death frequency-matching process for determining frequency tracks. Note the death of two tracks during frames one and three and the birth of a track during the second frame.

SOURCE: R.J. McAulay and T.F. Quatieri, "Low Rate Speech Coding Based on the Sinusoidal Speech Model," chapter in *Advances in Speech Signal Processing* [34]. ©1992, Marcel Dekker, Inc. Courtesy of Marcel Dekker, Inc.

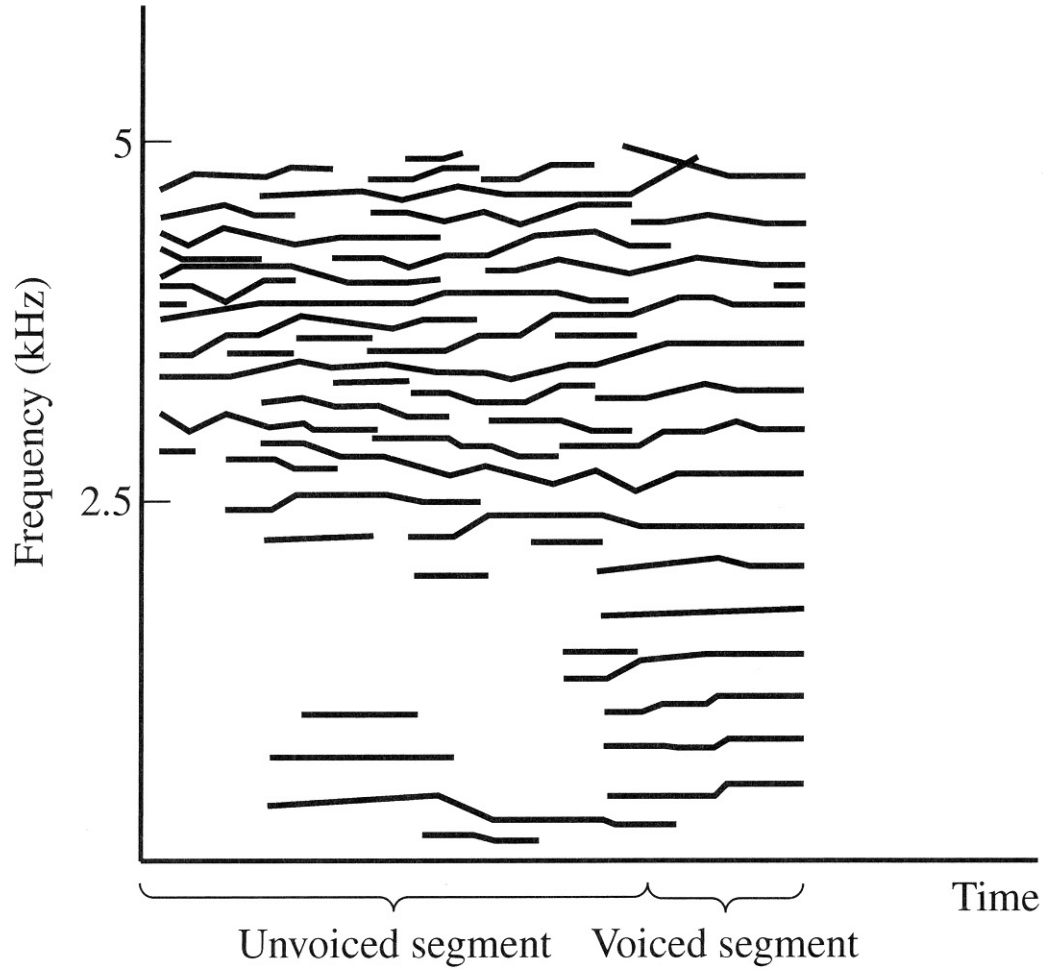


Figure 9.11 Typical frequency tracks for real speech.

SOURCE: R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation" [30]. ©1986, IEEE. Used by permission.