Chap 01-05,09

Fg1.1

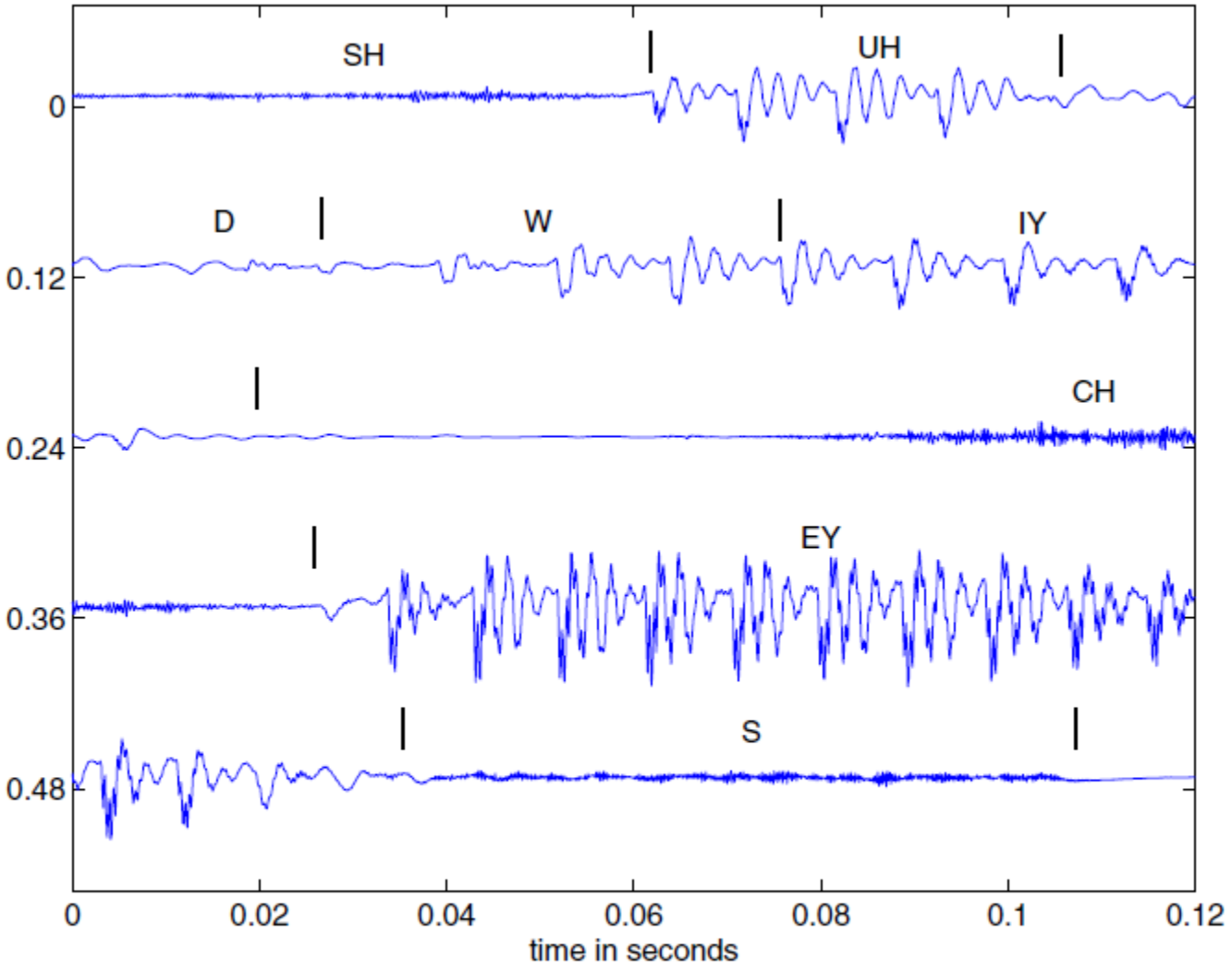

Fig. 1.1 A speech waveform with phonetic labels for the text message "Should we chase."

Fg1.2

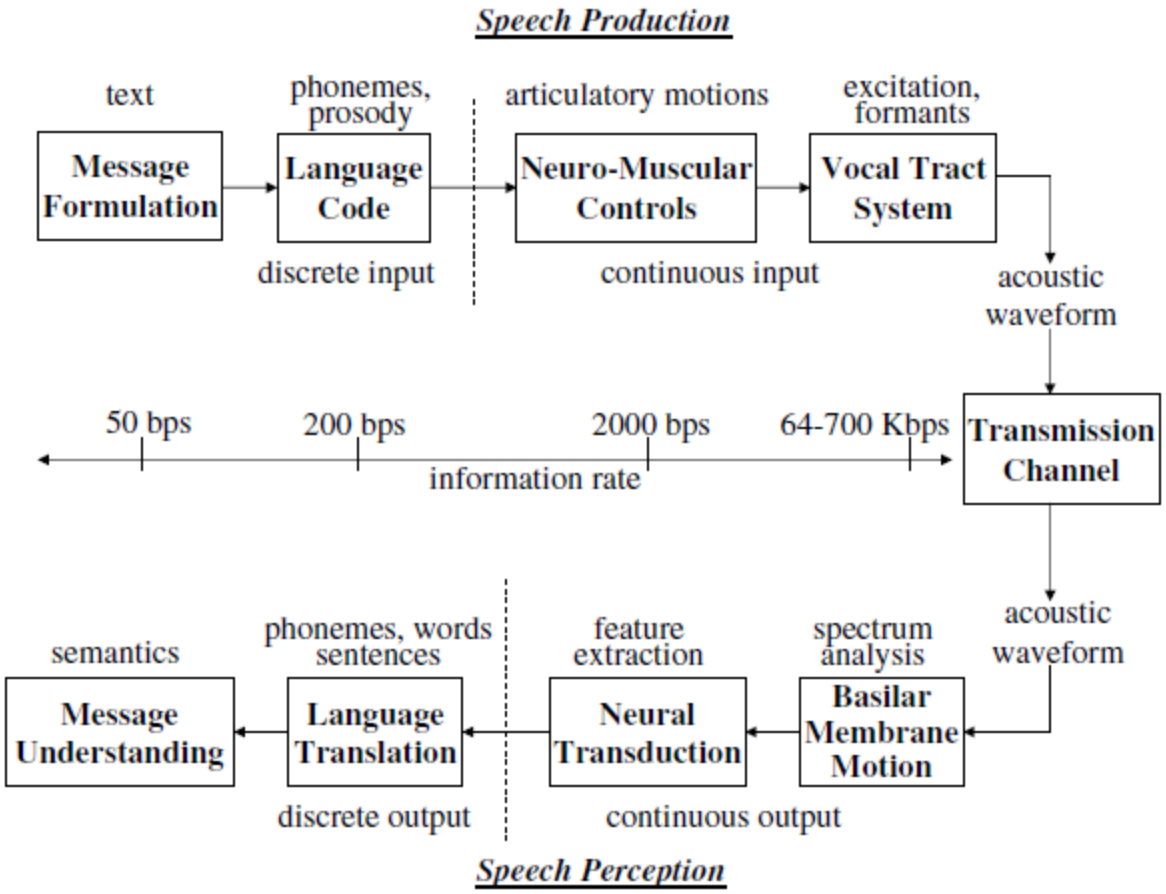

Fig. 1.2 The Speech Chain: from message, to speech signal, to understanding.

Fg1.3

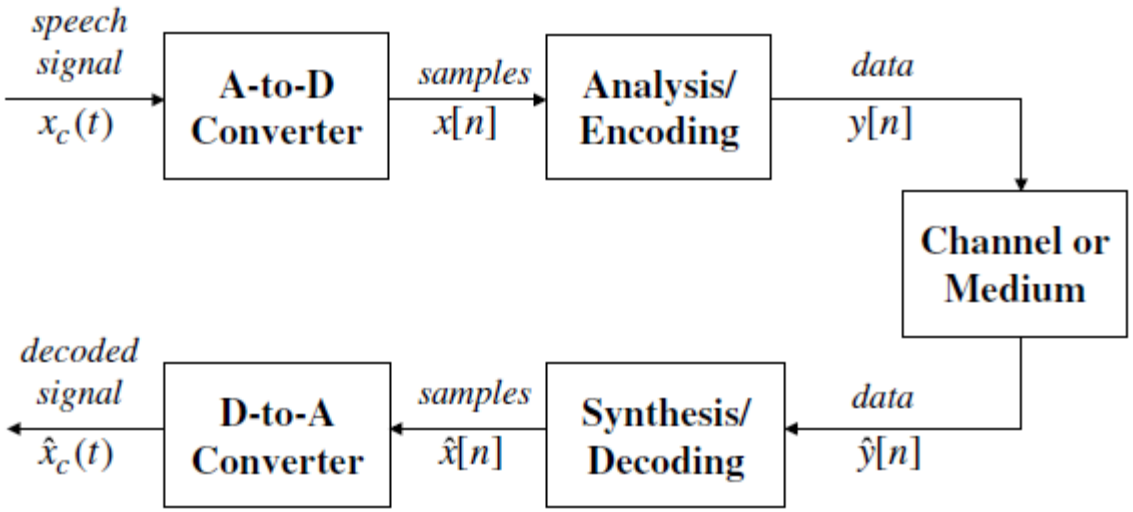

Fig. 1.3 Speech coding block diagram — encoder and decoder.

Fg1.4

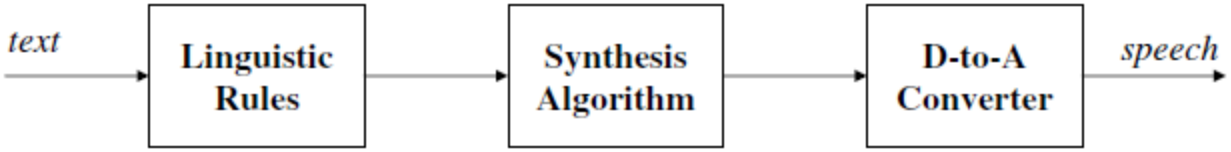

Fig. 1.4 Text-to-speech synthesis system block diagram.

Fg1.5

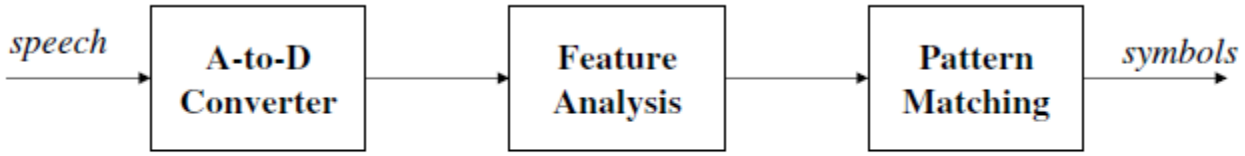

Fig. 1.5 Block diagram of general pattern matching system for speech signals.

Fg1.6

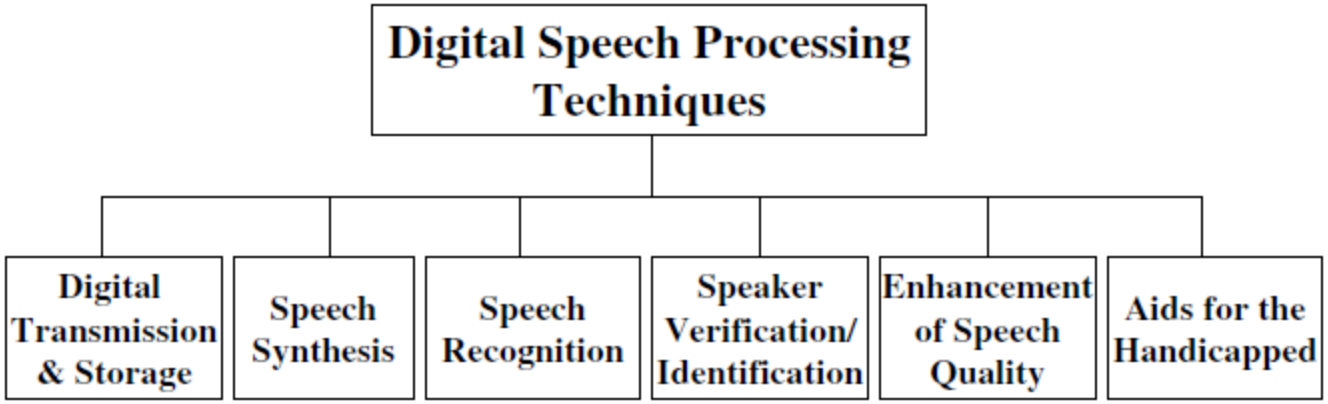

Fig. 1.6 Range of speech communication applications.

Fg1.7

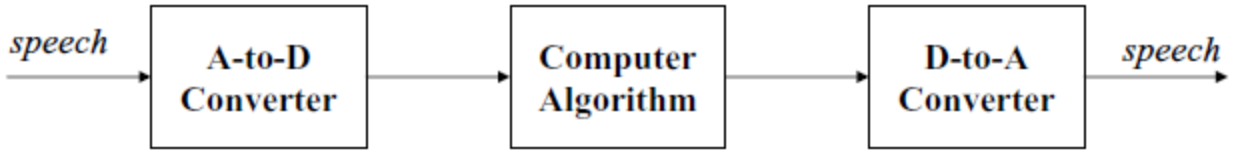

Fig. 1.7 General block diagram for application of digital signal processing to speech signals.

Tb2.1_1

Table 2.1 Condensed list of ARPAbet phonetic symbols for North American English.

Class	ARPAbet	Example	Transcription
Vowels and	IY	b*ee*t	[B IY T]
diphthongs	IH	b*i*t	[B IH T]
	EY	b*ai*t	[B EY T]
	EH	b*e*t	[B EH T]
	AE	b*a*t	[B AE T]
	AA	b*o*b	[B AA B]
	AO	b*or*n	[B AO R N]
	UH	b*oo*k	[B UH K]
	OW	b*oa*t	[B OW T]
	UW	b*oo*t	[B UW T]
	AH	b*u*t	[B AH T]
	ER	b*ir*d	[B ER D]
	AY	b*uy*	[B AY]
	AW	d*ow*n	[D AW N]
	OY	b*oy*	[B OY]

Tb2.1_2

Glides	Y	*y*ou	[Y UH]
	R	*r*ent	[R EH N T]
Liquids	W	*w*it	[W IH T]
	L	*l*et	[L EH T]
Nasals	M	*m*et	[M EH T]
	N	*n*et	[N EH T]
	NG	si*ng*	[S IH NG]
Stops	P	*p*at	[P AE T]
	B	*b*et	[B EH T]
	T	*t*en	[T EH N]
	D	*d*ebt	[D EH T]
	K	*k*it	[K IH T]
	G	*g*et	[G EH T]
Fricatives	HH	*h*at	[HH AE T]
	F	*f*at	[F AE T]
	V	*v*at	[V AE T]
	TH	*th*ing	[TH IH NG]
	DH	*th*at	[DH AE T]
	S	*s*at	[S AE T]
	Z	*z*oo	[Z UW]
	SH	*sh*ut	[SH AH T]
	ZH	a*z*ure	[AE ZH ER]
Affricates	CH	*ch*ase	[CH EY S]
	JH	*j*udge	[JH AH JH]

[a]This set of 39 phonemes is used in the CMU Pronouncing Dictionary available on-line at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Fg2.1

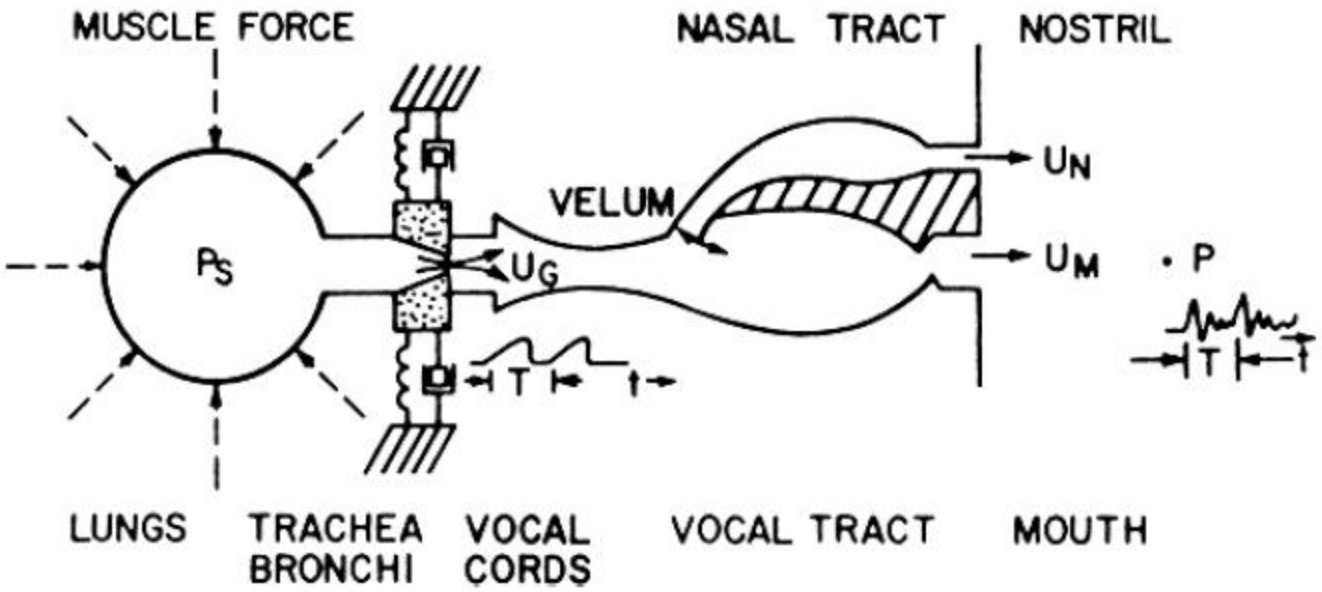

Fig. 2.1 Schematic model of the vocal tract system. (After Flanagan et al. [35].)

Fg2.2

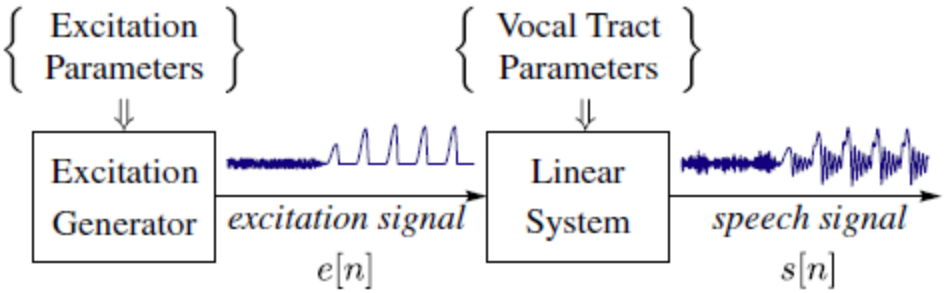

Fig. 2.2 Source/system model for a speech signal.

Fg3.1

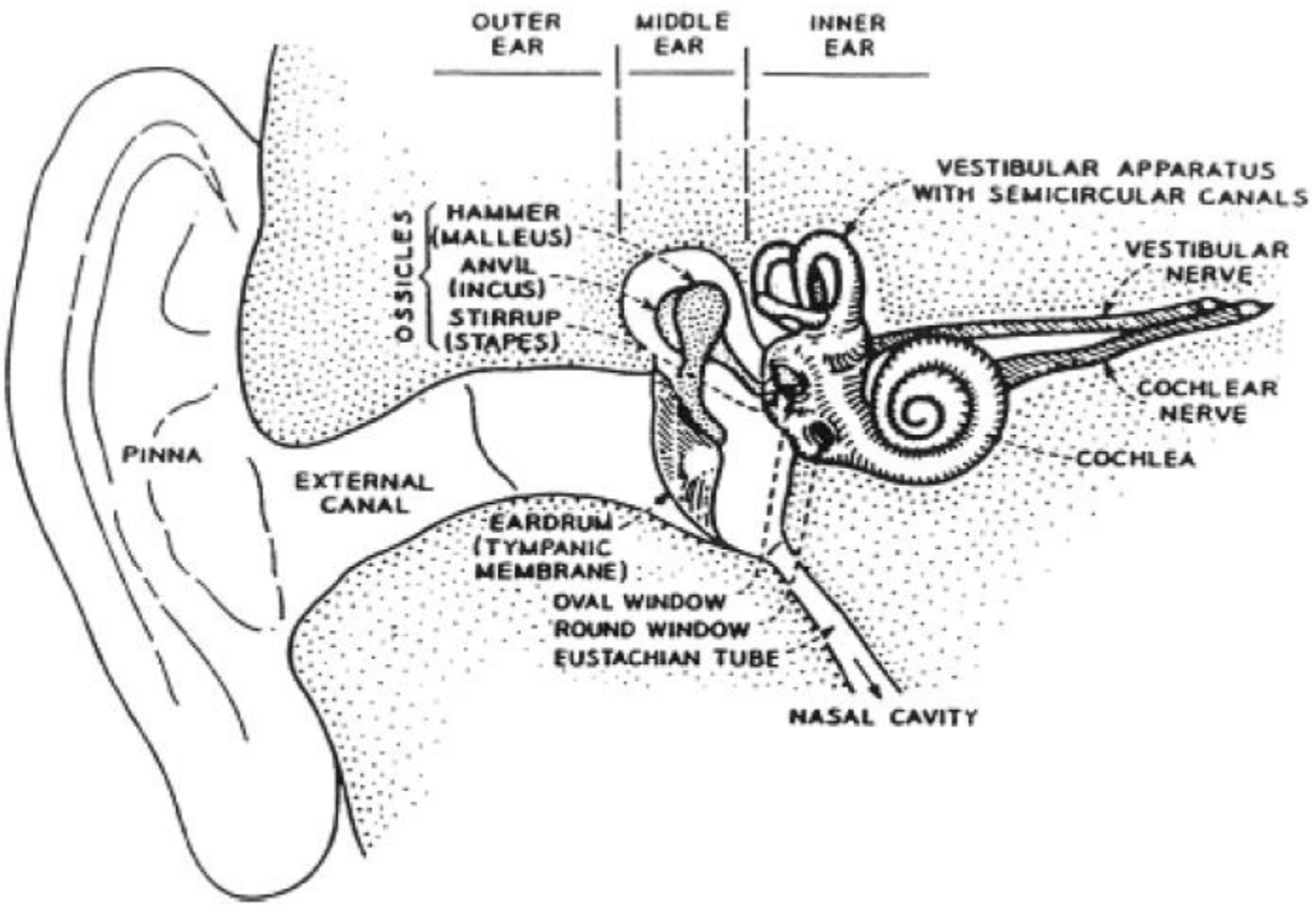

Fig. 3.1 Schematic view of the human ear (inner and middle structures enlarged). (After Flanagan [34].)

Fg3.2

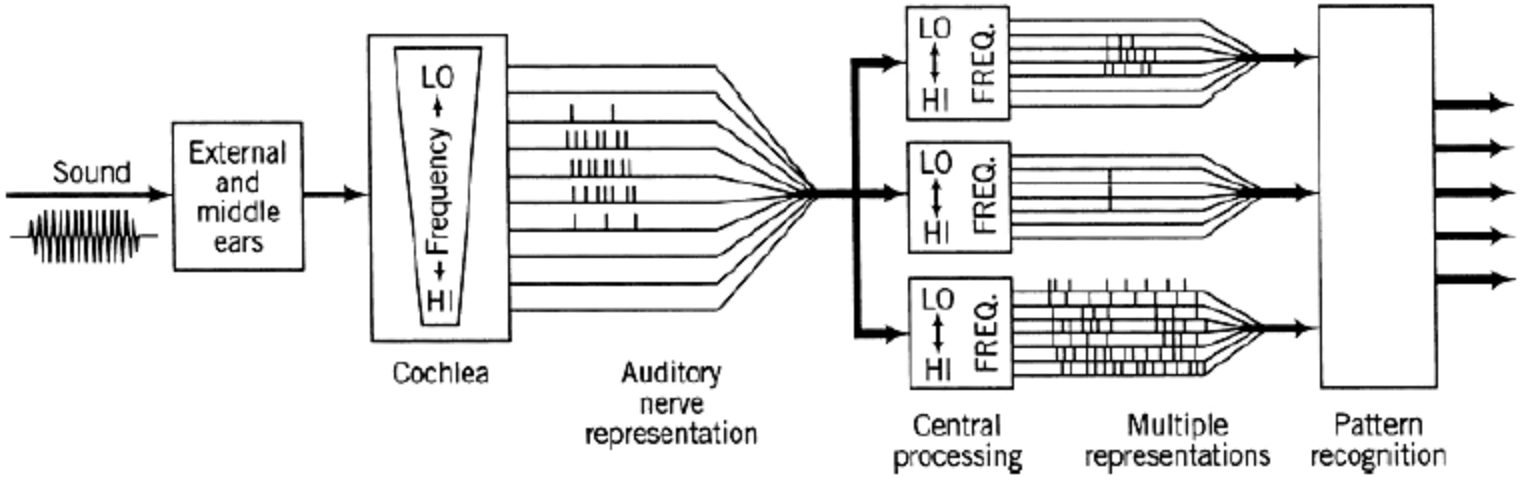

Fig. 3.2 Schematic model of the auditory mechanism. (After Sachs et al. [107].)

Fg3.3

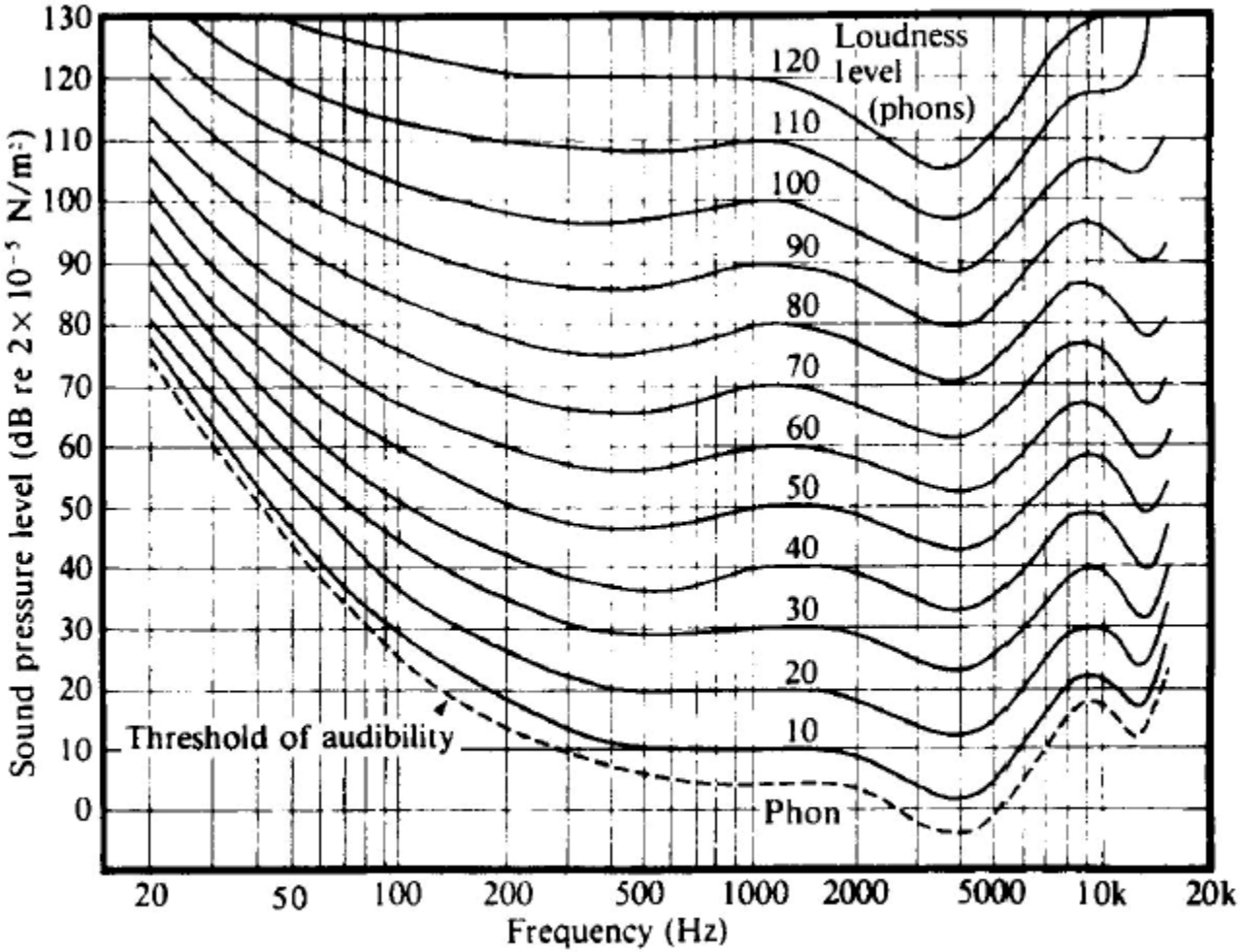

Fig. 3.3 Loudness level for human hearing. (After Fletcher and Munson [37].)

Fg3.4

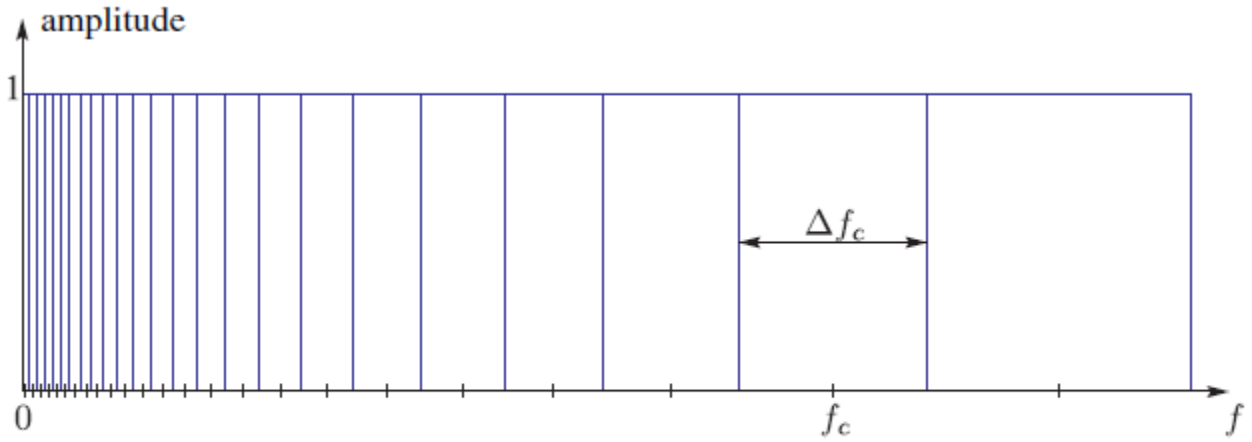

Fig. 3.4 Schematic representation of bandpass filters according to the critical band theory of hearing.

Fg3.5

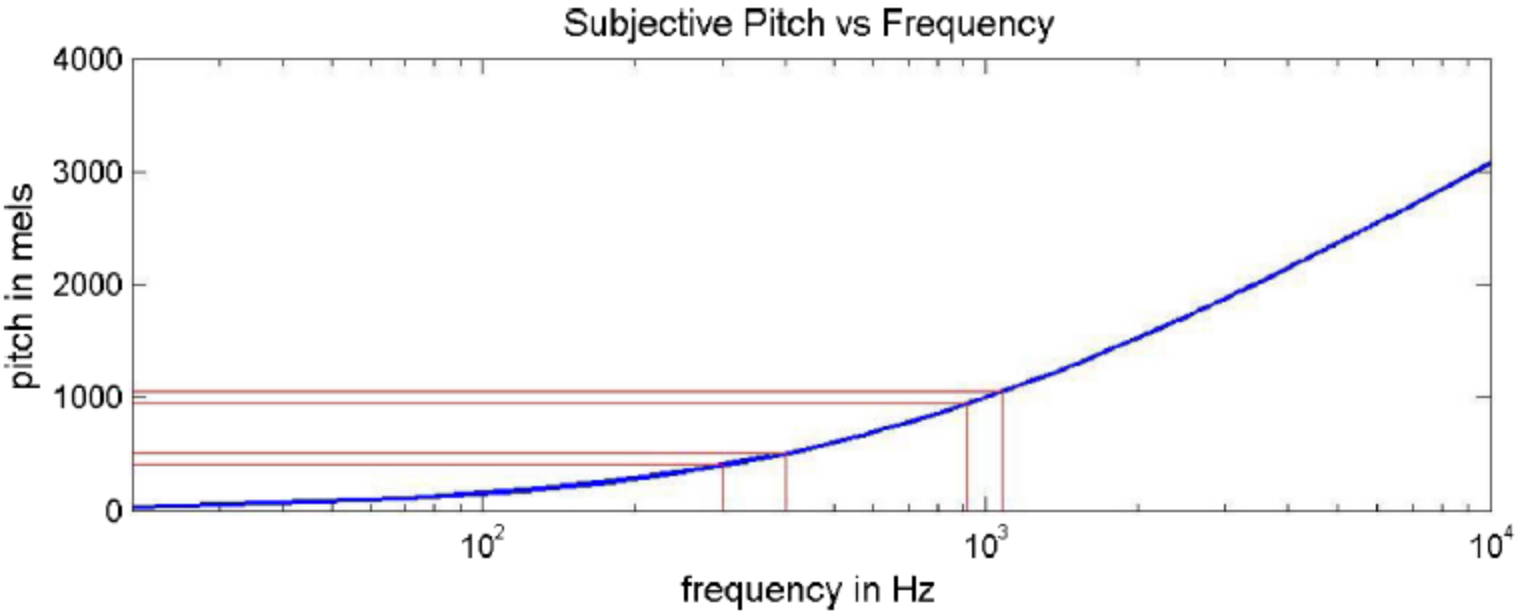

Fig. 3.5 Relation between subjective pitch and frequency of a pure tone.

Fg3.6

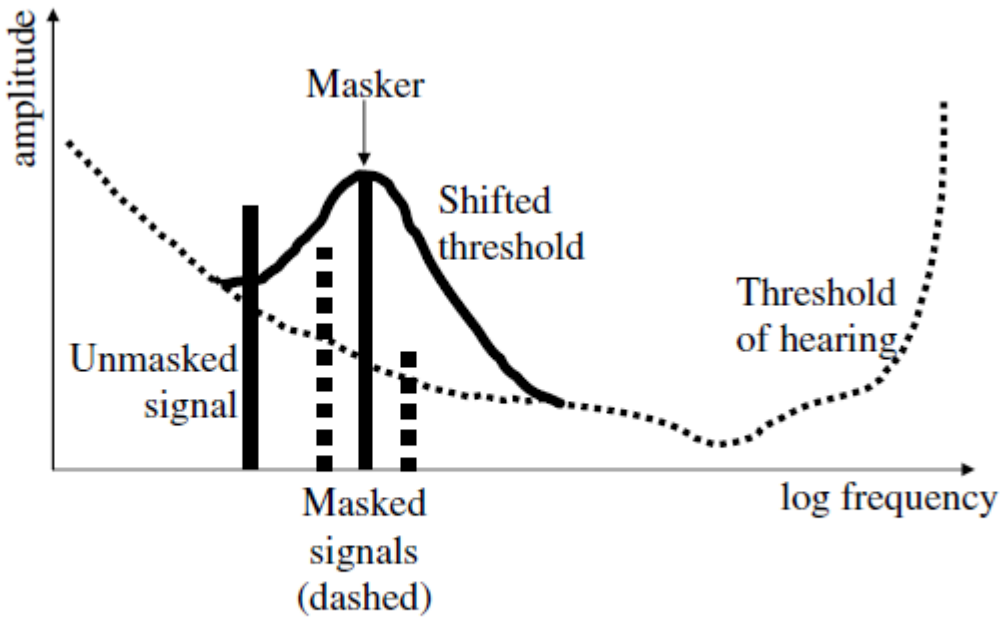

Fig. 3.6 Illustration of effects of masking.

Fg4.1

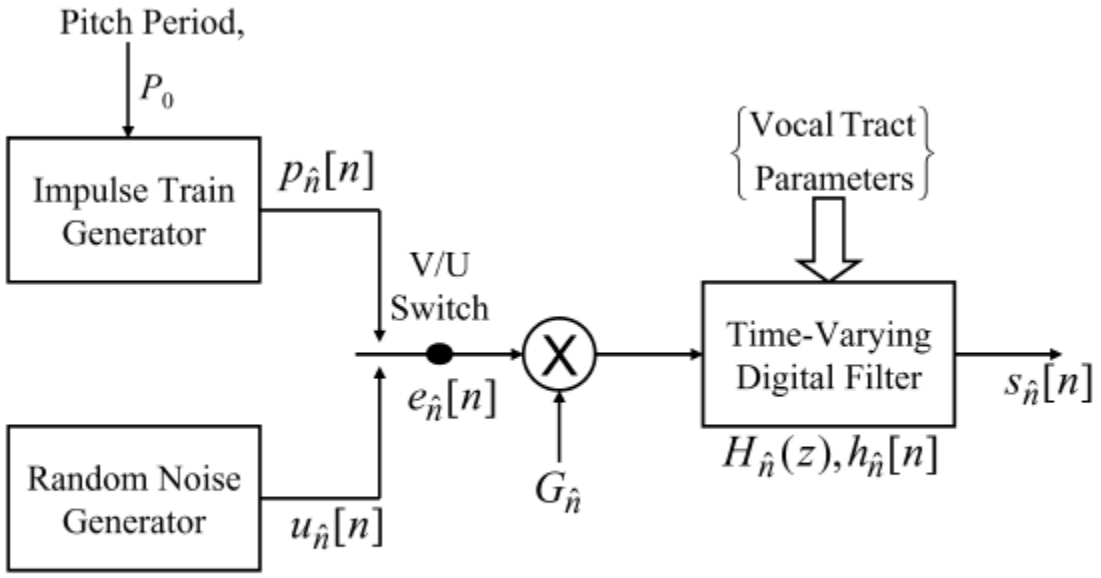

Fig. 4.1 Voiced/unvoiced/system model for a speech signal.

Fg4.2

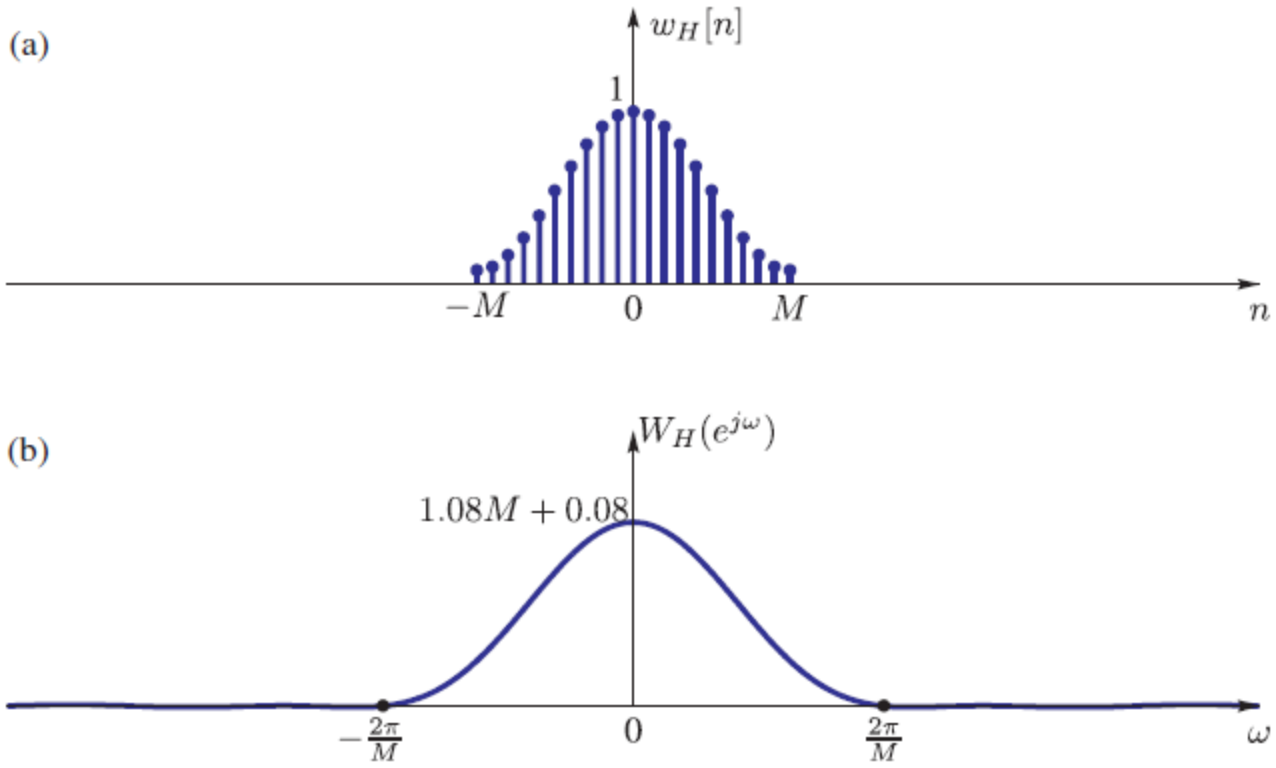

Fig. 4.2 Hamming window (a) and its discrete-time Fourier transform (b).

Fg4.3

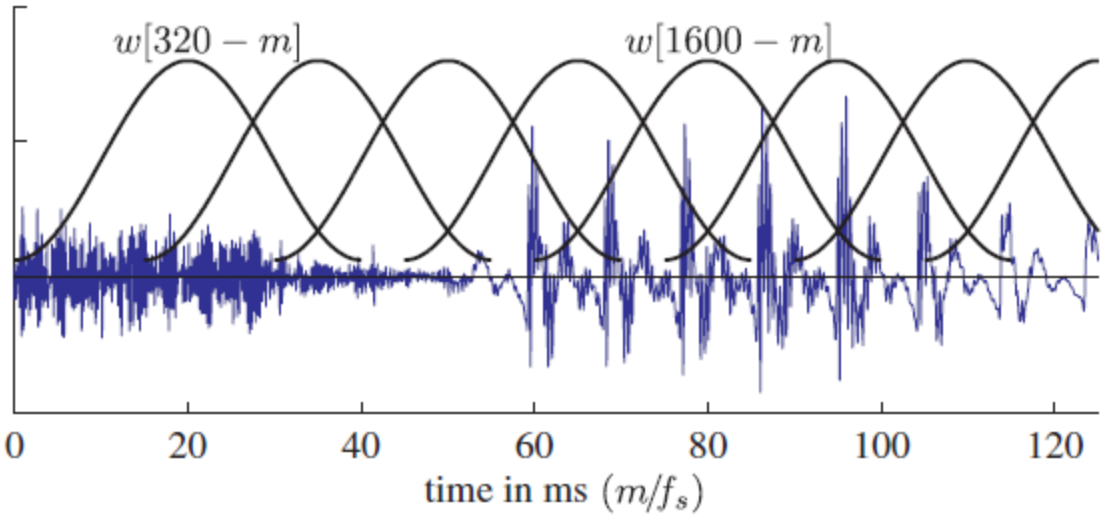

Fig. 4.3 Section of speech waveform with short-time analysis windows.

Fg4.4

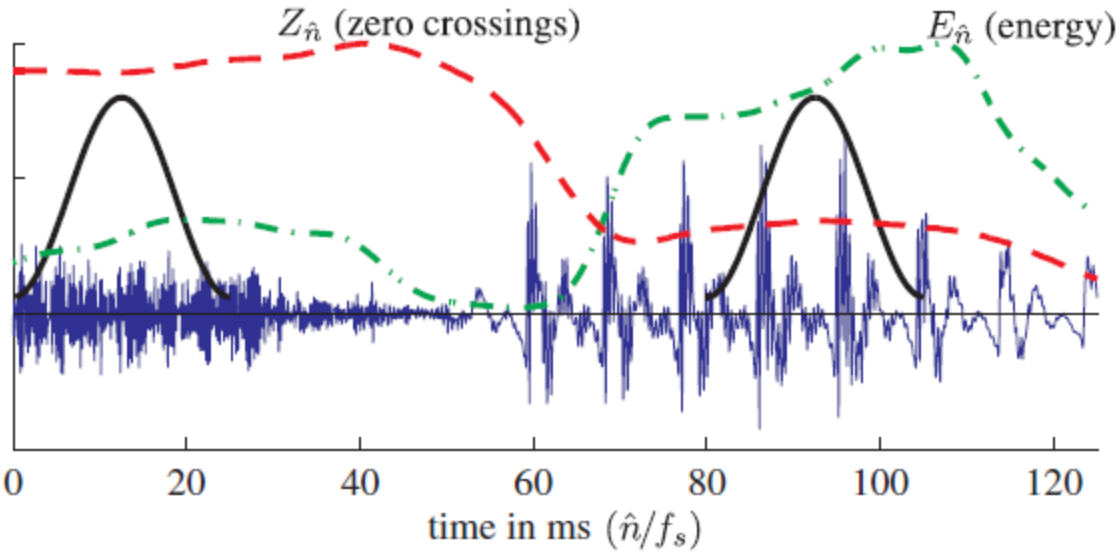

Fig. 4.4 Section of speech waveform with short-time energy and zero-crossing rate superimposed.

Fg4.5

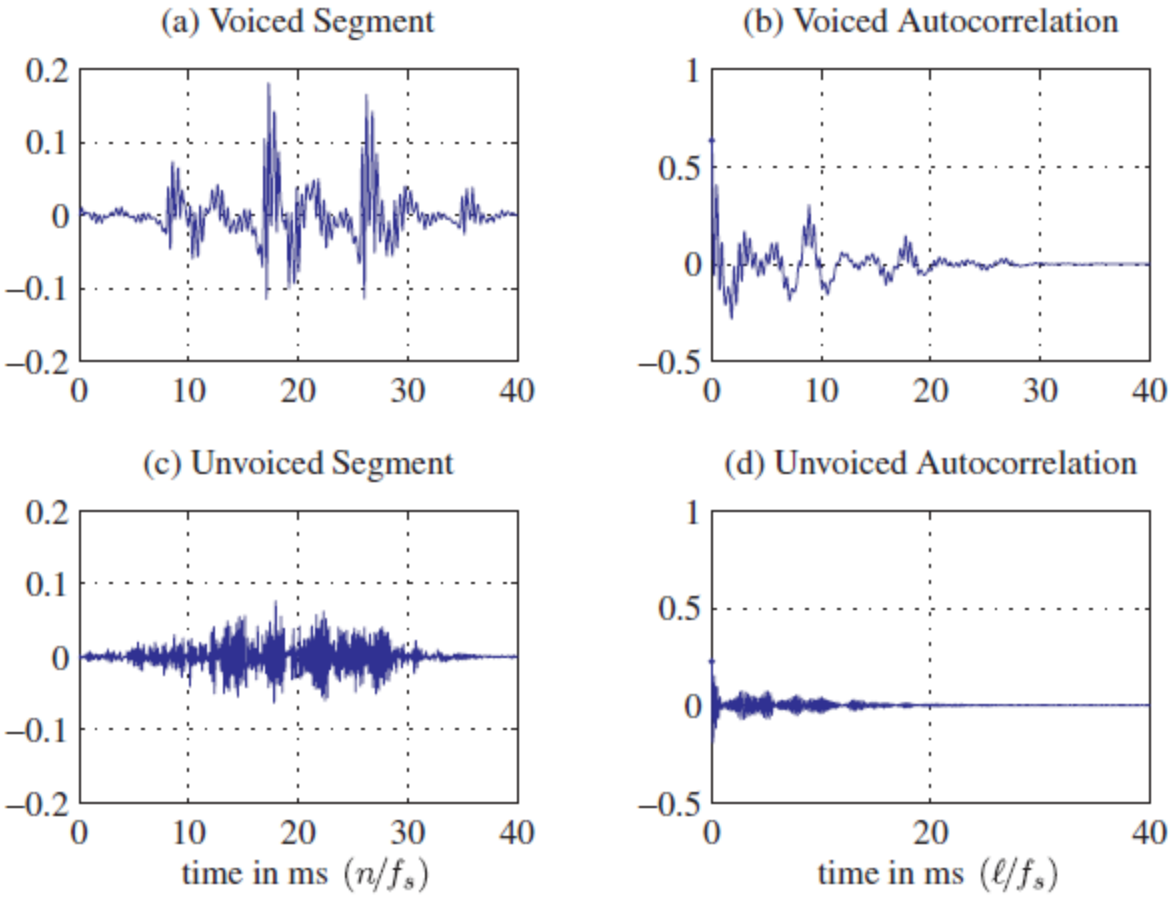

Fig. 4.5 Voiced and unvoiced segments of speech and their corresponding STACF.

Fg4.6

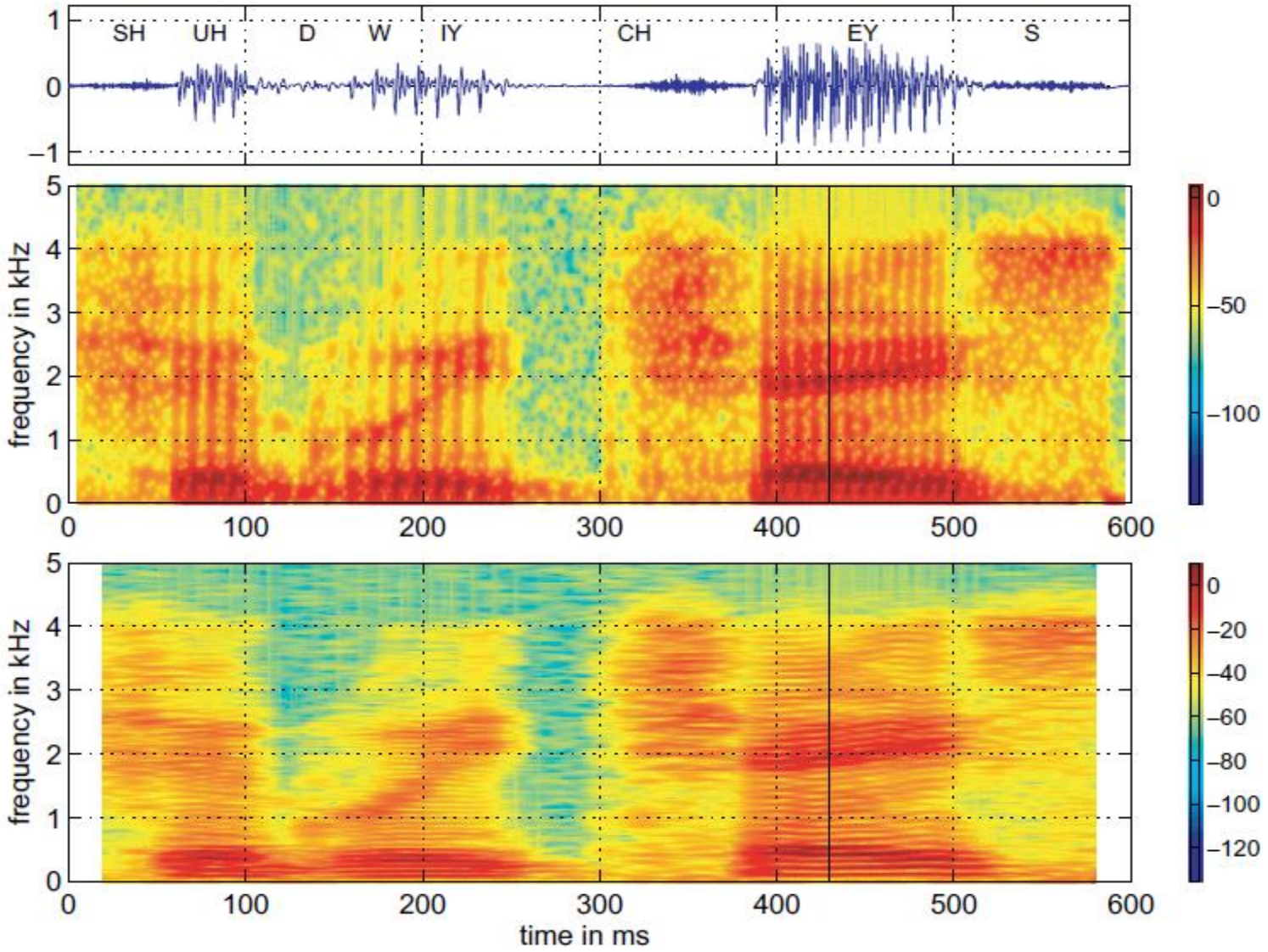

Fig. 4.6 Spectrogram for speech signal of Figure 1.1.

Fg4.7

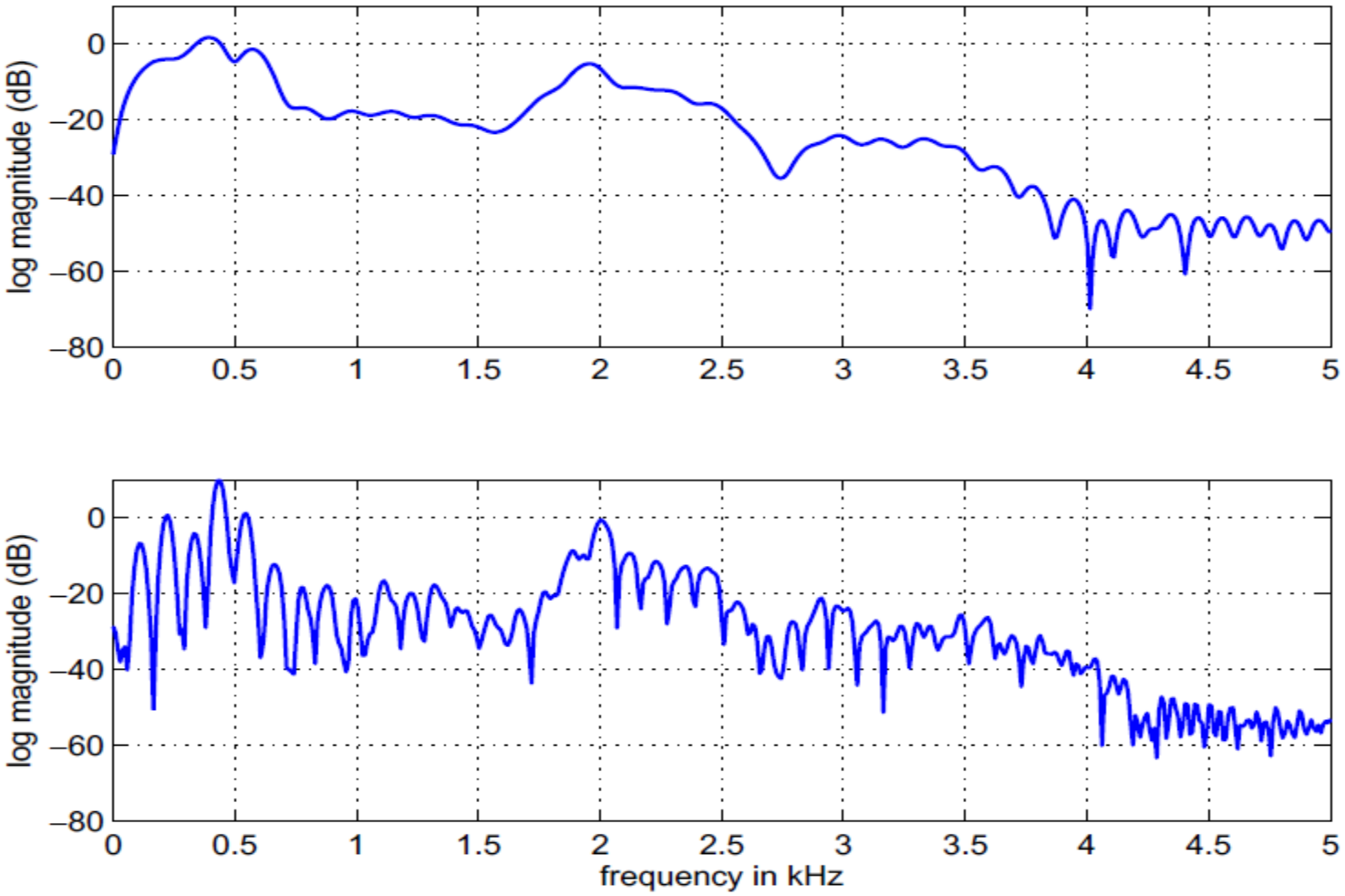

Fig. 4.7 Short-time spectrum at time 430 ms (dark vertical line in Figure 4.6) with Hamming window of length $M = 101$ in upper plot and $M = 401$ in lower plot.

Fg4.8

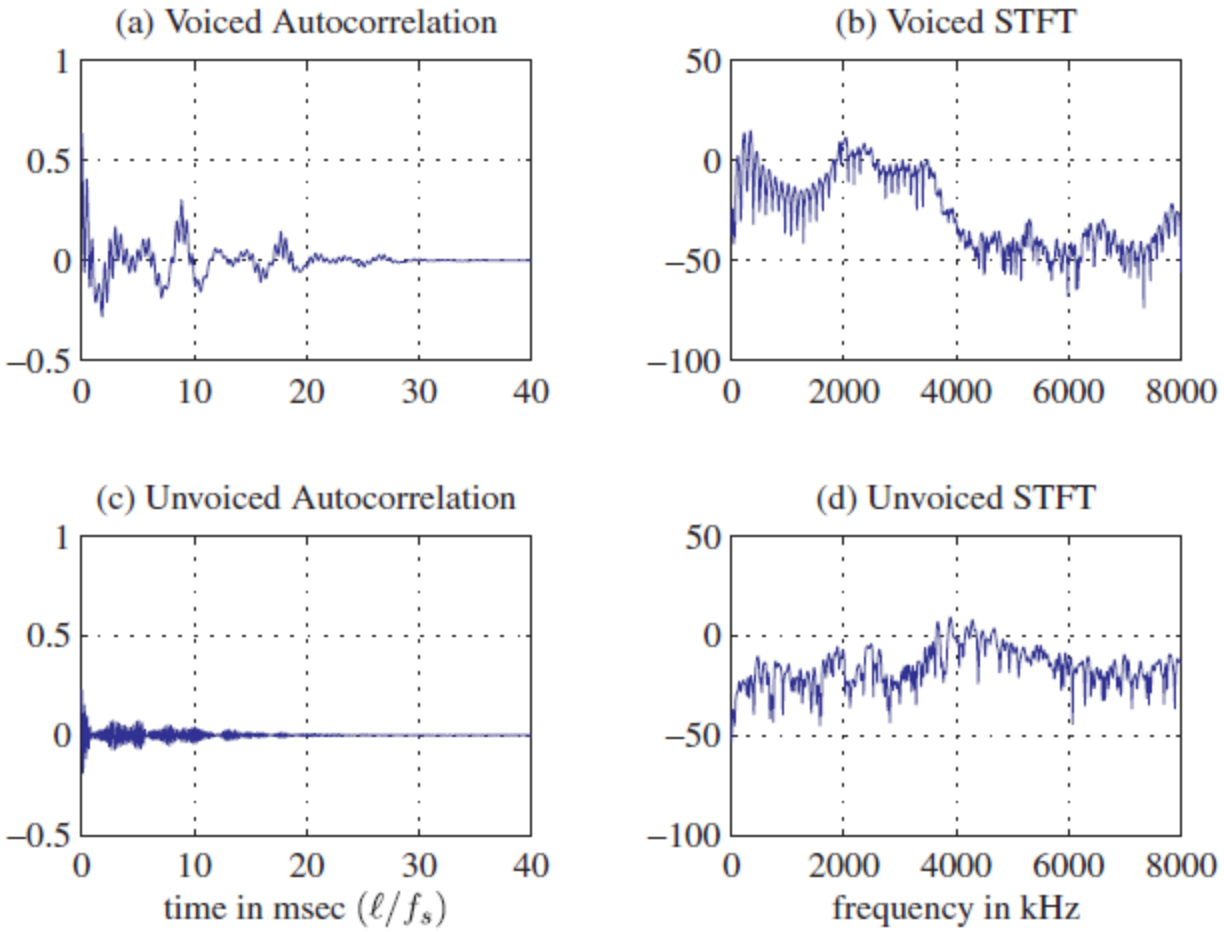

Fig. 4.8 STACF and corresponding STFT.

Fg4.9

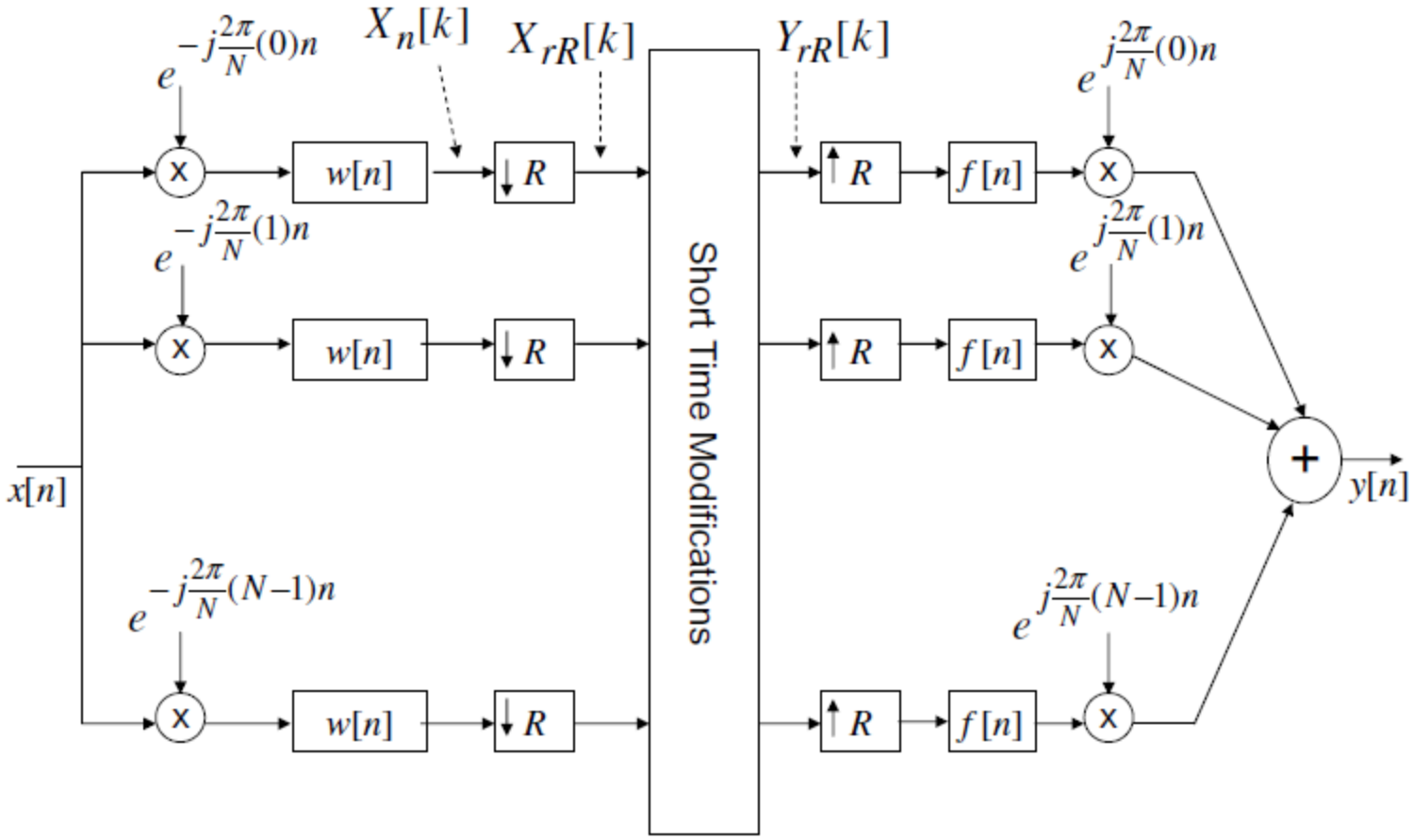

Fig. 4.9 Filterbank interpretation of short-time Fourier analysis and synthesis.

Fg5.1

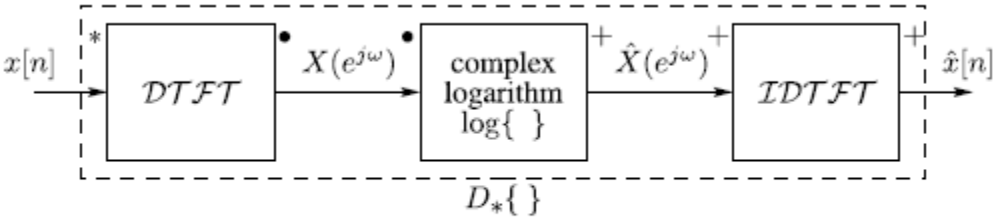

Fig. 5.1 Computing the complex cepstrum using the DTFT.

Fg5.2

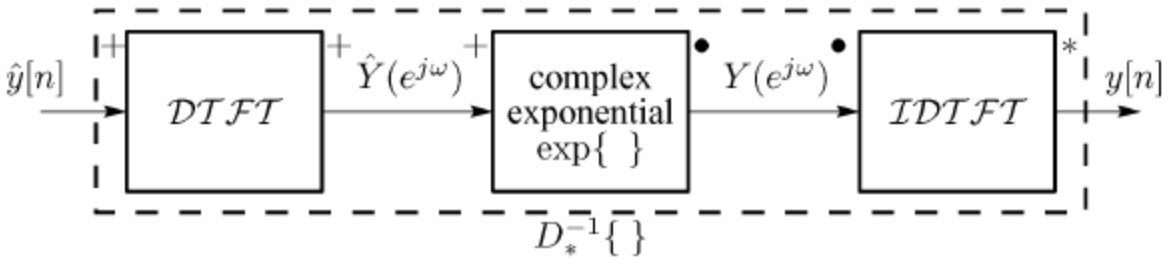

Fig. 5.2 The inverse of the characteristic system for convolution (inverse complex cepstrum).

Fg5.3

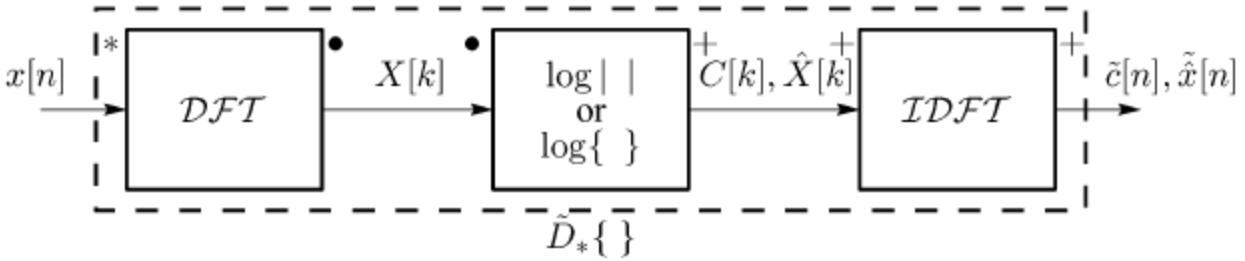

Fig. 5.3 Computing the cepstrum or complex cepstrum using the DFT.

Fg5.4

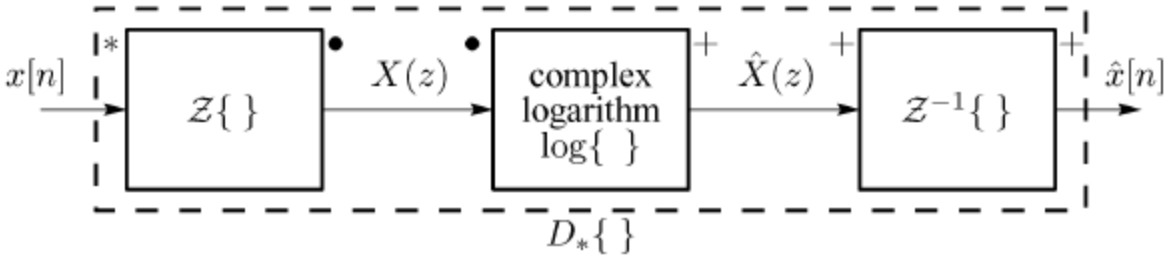

Fig. 5.4 z-transform representation of characteristic system for convolution.

Fg5.5

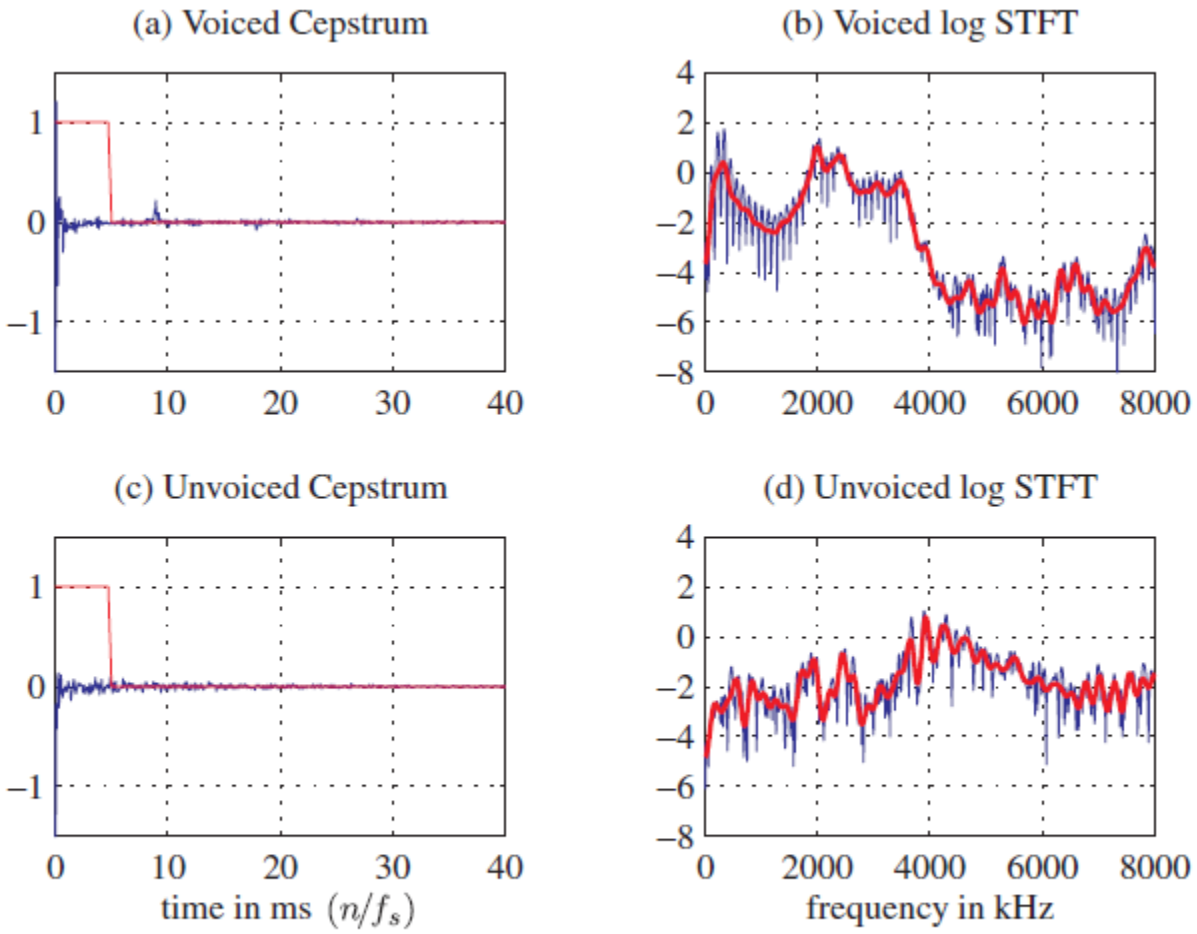

Fig. 5.5 Short-time cepstra and corresponding STFTs and homomorphically-smoothed spectra.

Fg5.6

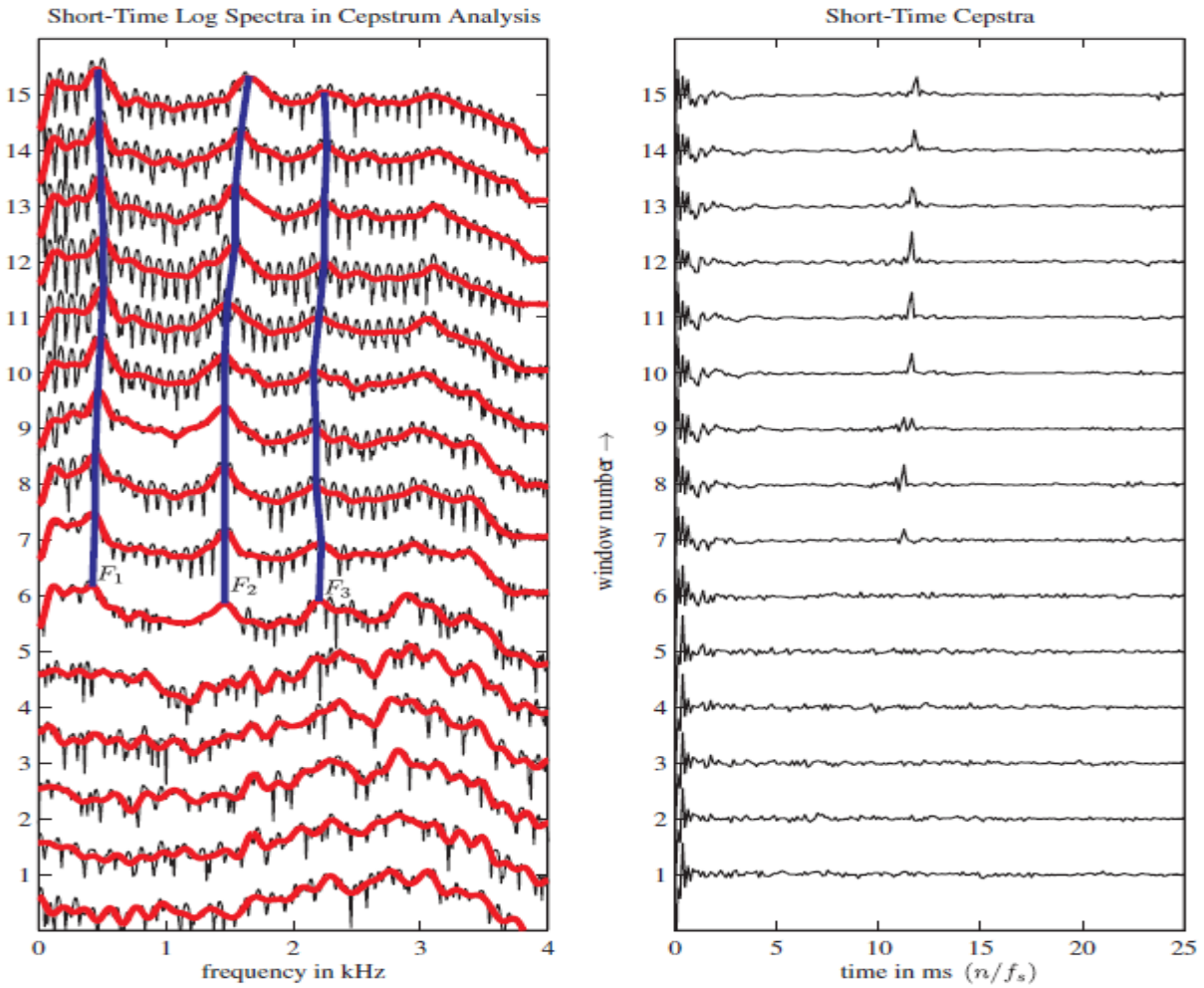

Fig. 5.6 Short-time cepstra and corresponding STFTs and homomorphically-smoothed spectra.

Fg5.7

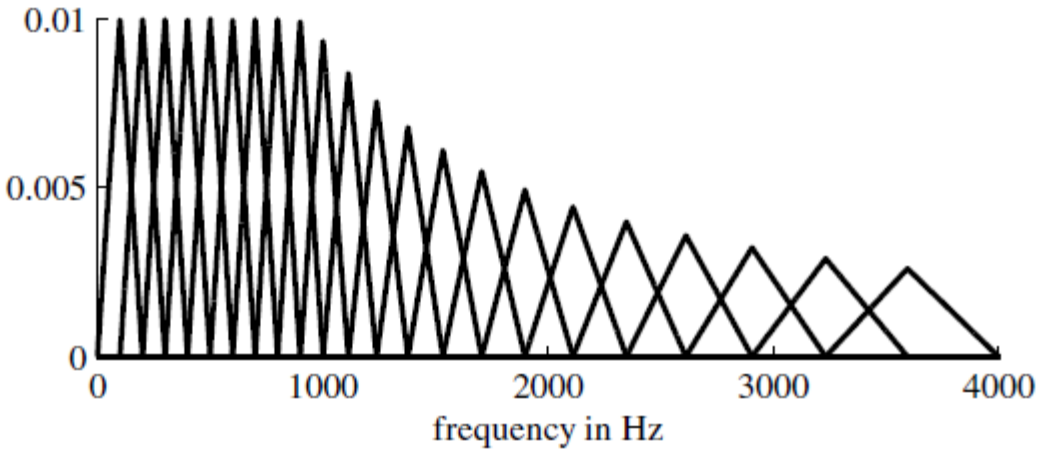

Fig. 5.7 Weighting functions for Mel-frequency filter bank.

Fg5.8

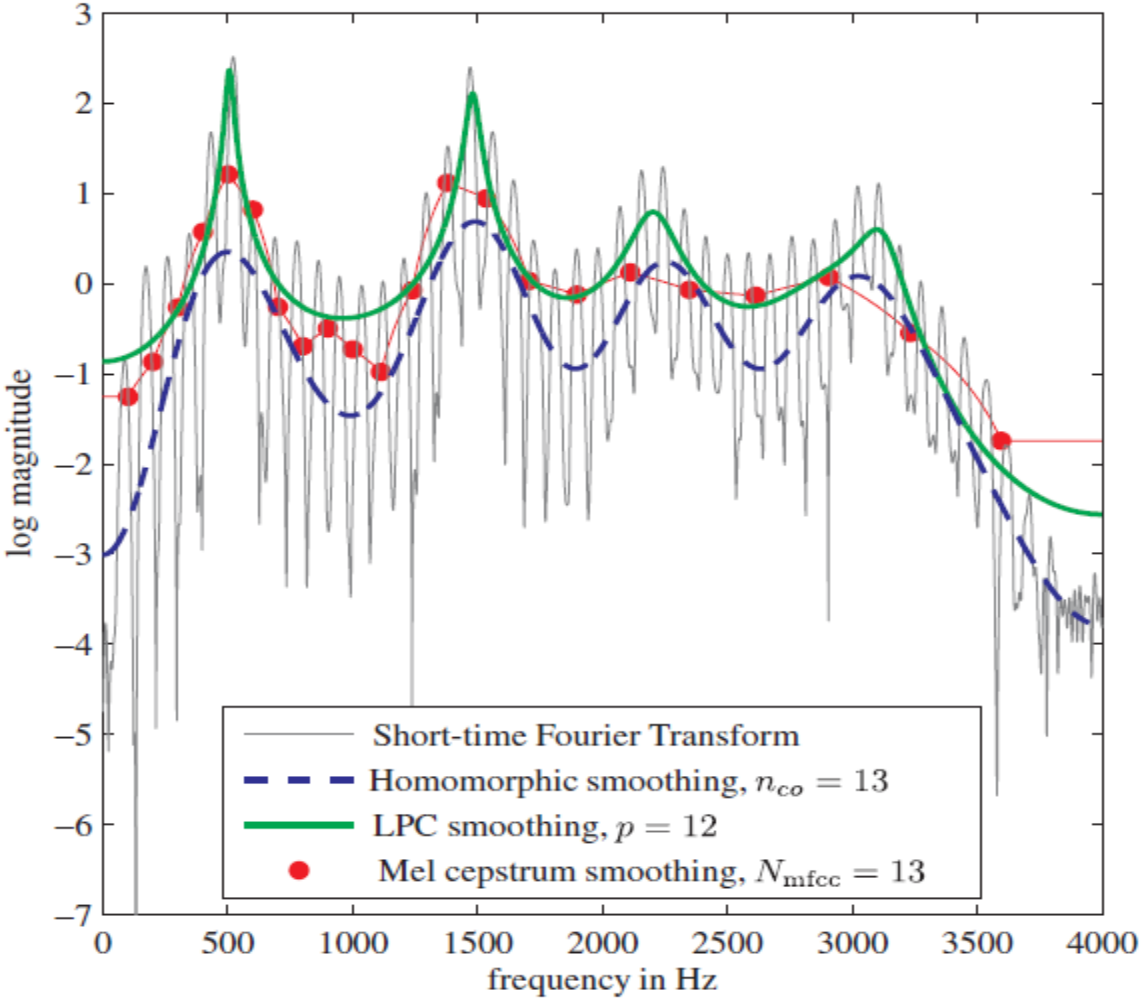

Fig. 5.8 Comparison of spectral smoothing methods.

Fg5.9

Fg9.1

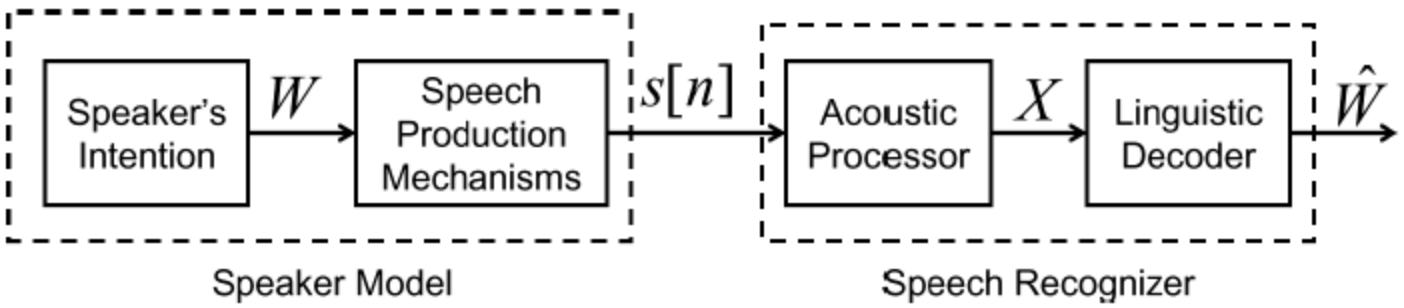

Fig. 9.1 Conceptual model of speech production and speech recognition processes.

Fg9.2

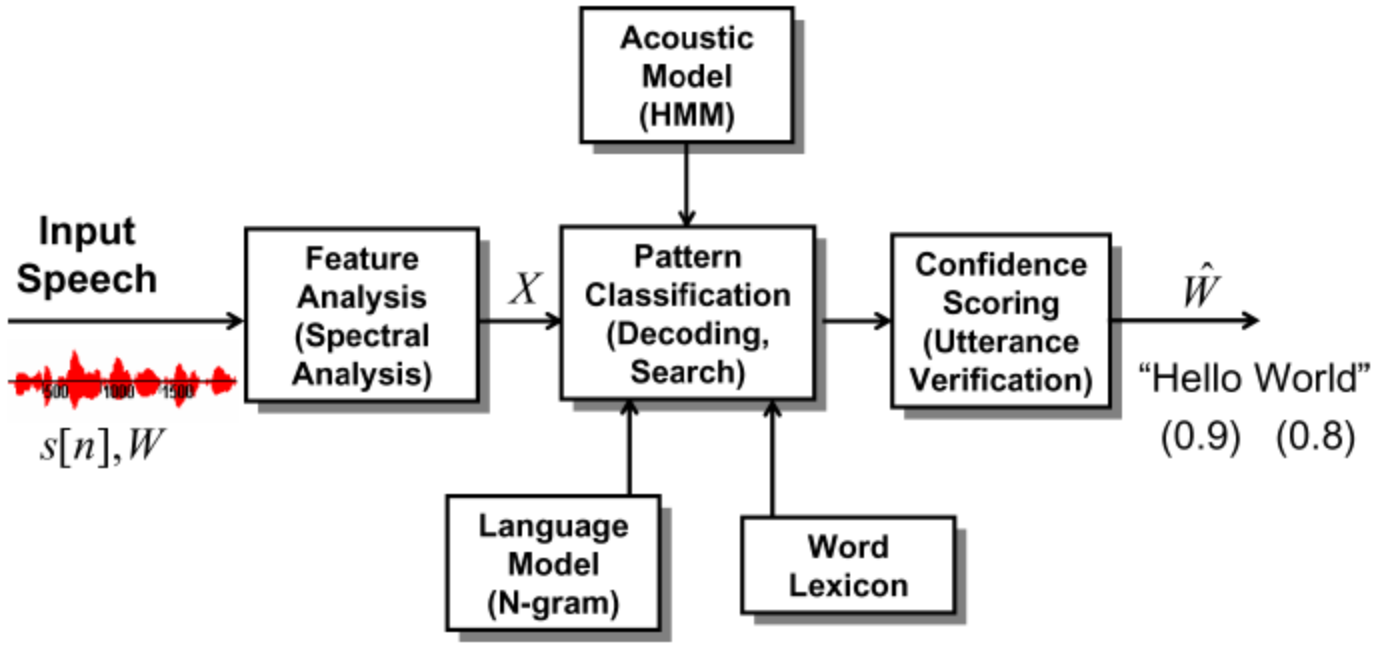

Fig. 9.2 Block diagram of an overall speech recognition system.

Fg9.3

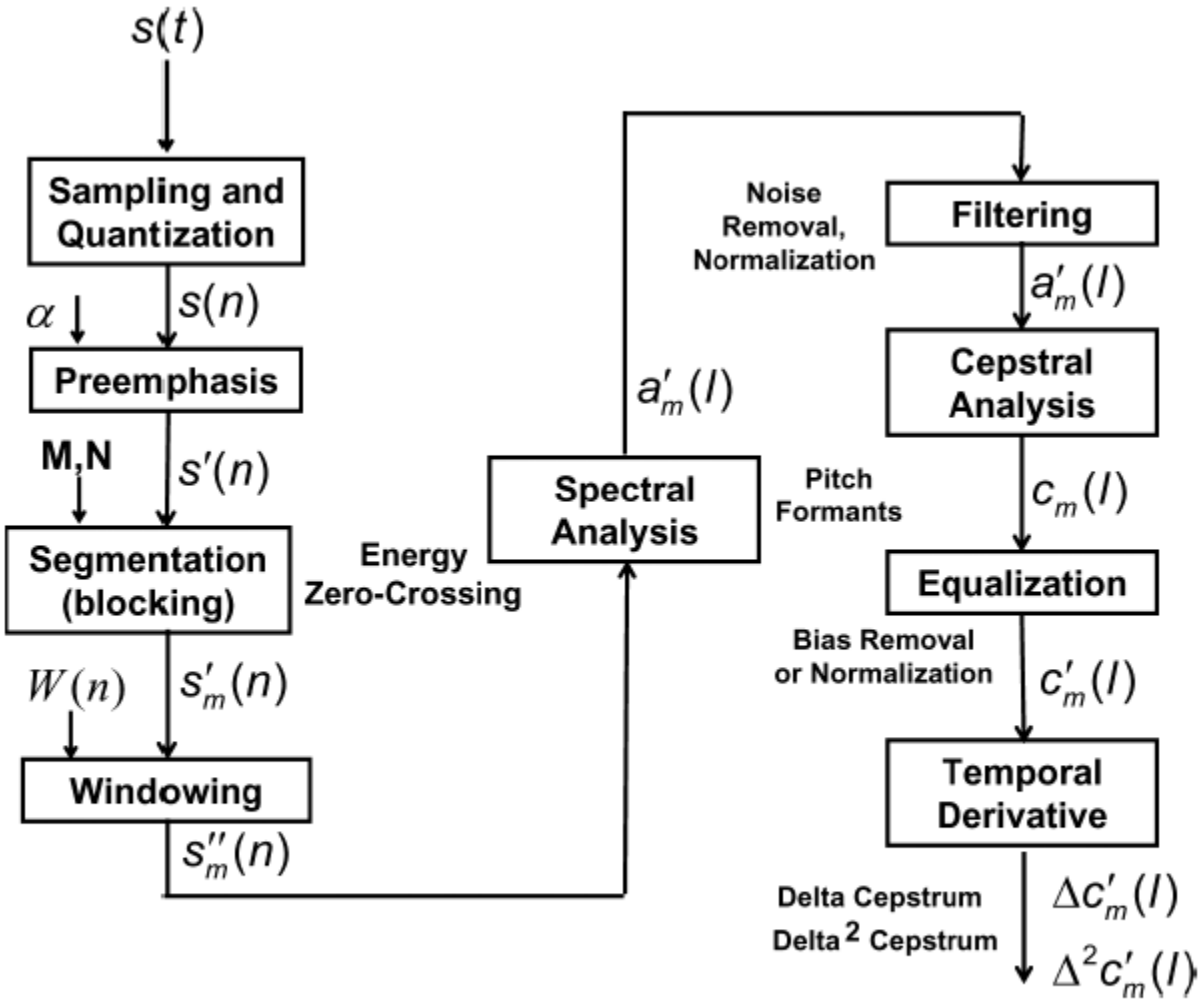

Fig. 9.3 Block diagram of feature extraction process for feature vector consisting of mfcc coefficients and their first and second derivatives.

Fg9.4

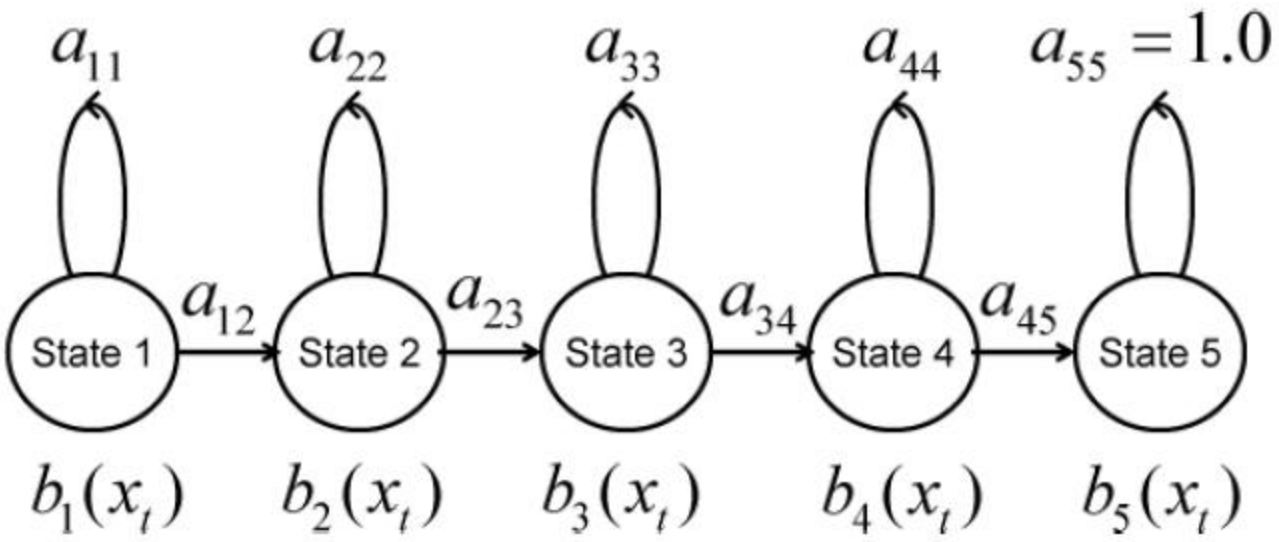

Fig. 9.4 Word-based, left-to-right, HMM with 5 states.

Fg9.5

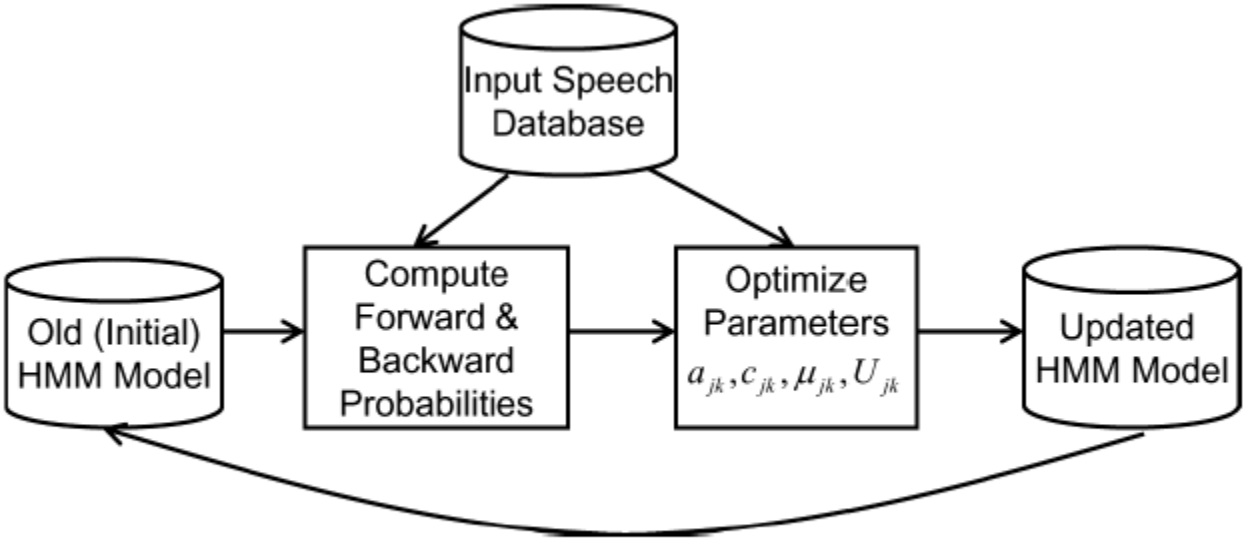

Fig. 9.5 The Baum–Welch training procedure based on a given training set of utterances.

Fg9.6

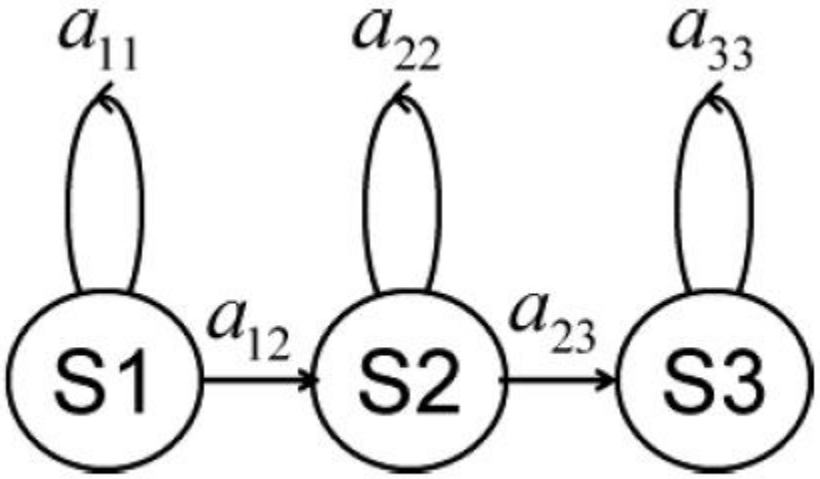

Fig. 9.6 Sub-word-based HMM with 3 states.

Fg9.7

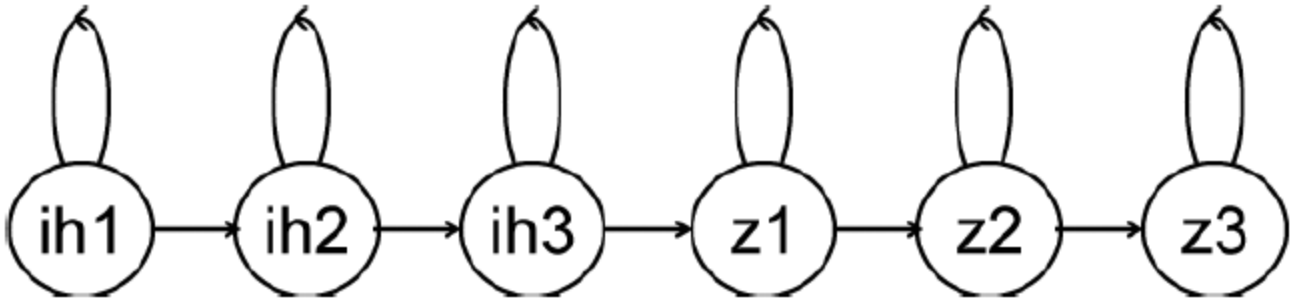

Fig. 9.7 Word-based HMM for the word /is/ created by concatenating 3-state subword models for the sub-word units /ih/ and /z/.

Fg9.8

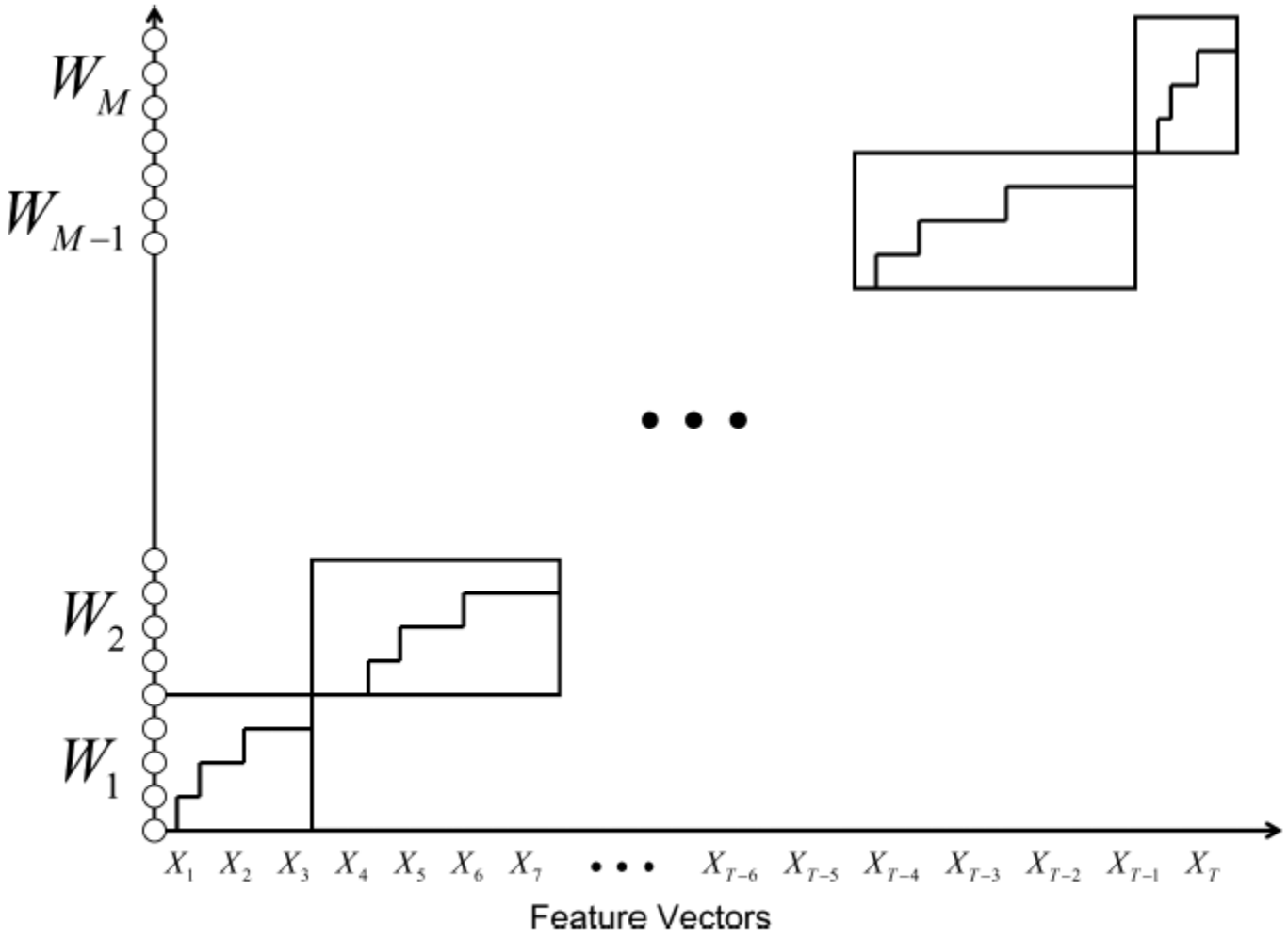

Fig. 9.8 Alignment of concatenated HMM word models with acoustic feature vectors based on either a Baum–Welch or Viterbi alignment procedure.

Fg9.9

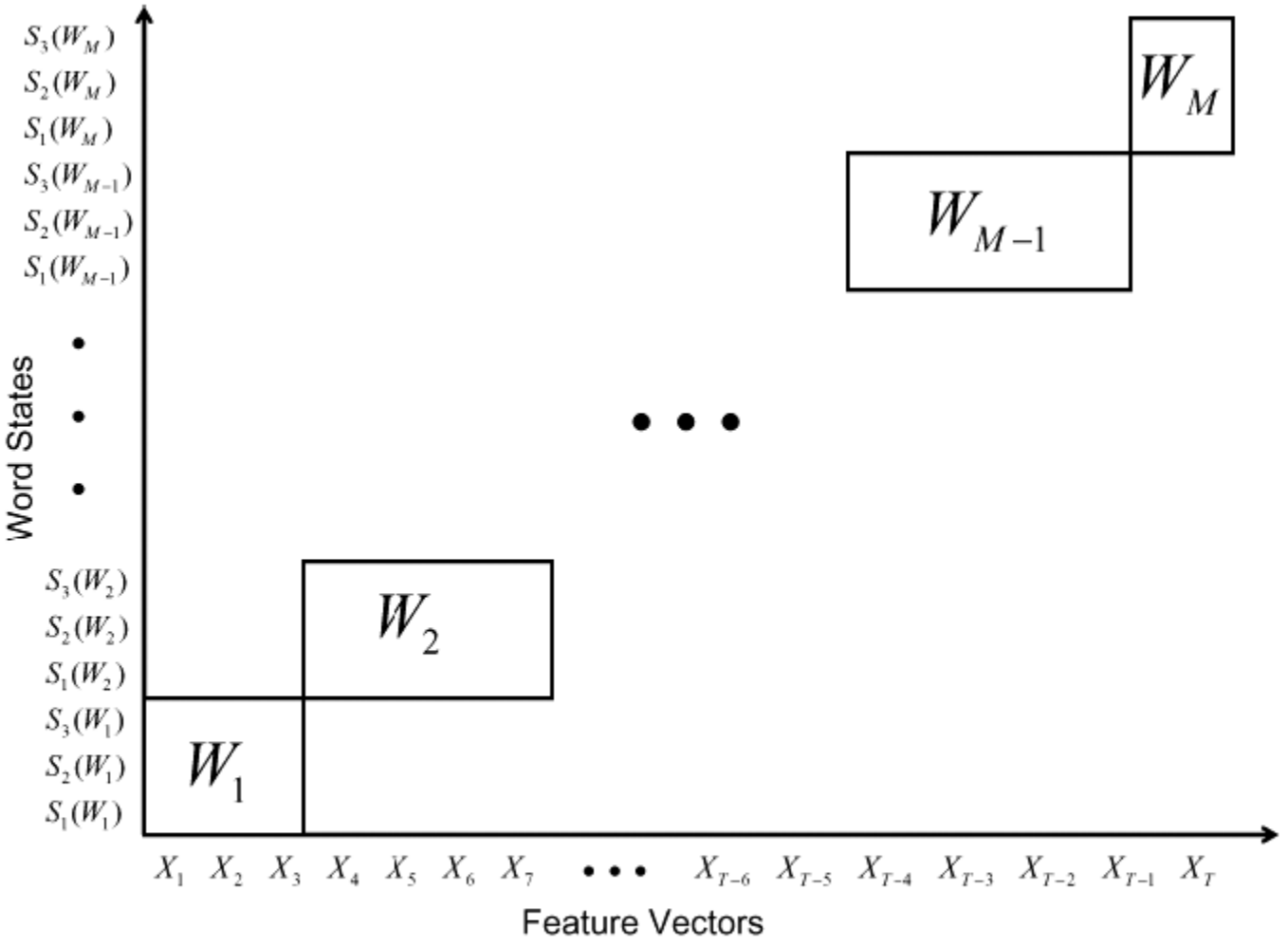

Fig. 9.9 Illustration of time alignment process between unknown utterance feature vectors and set of M concatenated word models.

Fg9.10

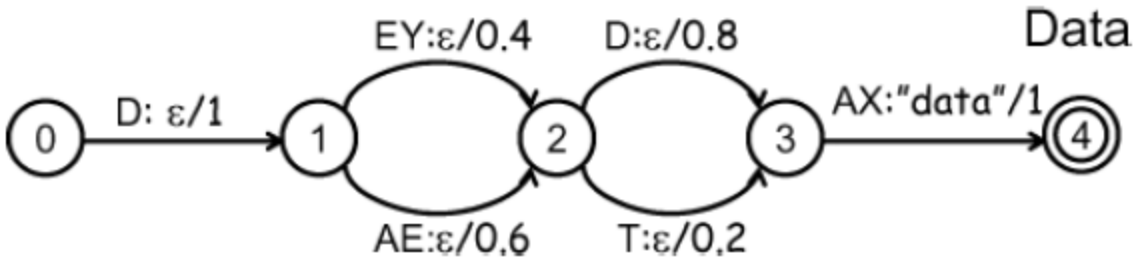

Fig. 9.10 Word pronunciation transducer for four pronunciations of the word /data/. (After Mohri [81].)

Fg9.11

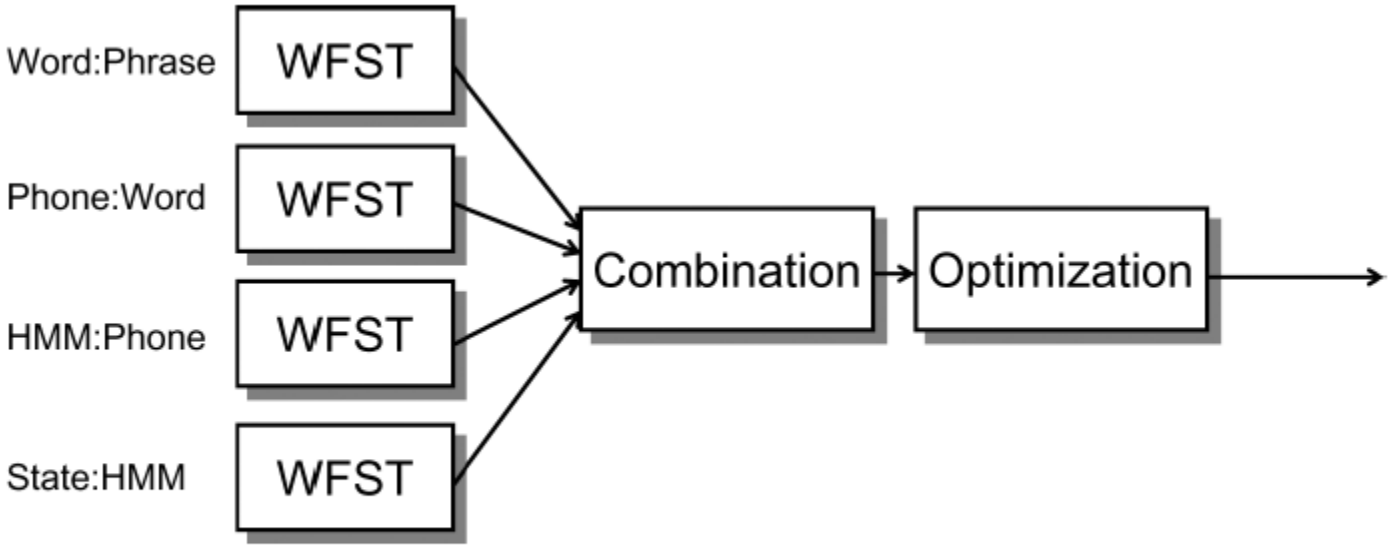

Fig. 9.11 Use of WFSTs to compile a set of FSNs into a single optimized network to minimize redundancy in the network. (After Mohri [81].)

Tb9.1

Table 9.1 Word error rates for a range of speech recognition systems.

Corpus	Type of speech	Vocabulary size	Word error rate (%)
Connected digit strings (TI Database)	Spontaneous	11 (0–9, oh)	0.3
Connected digit strings (AT&T Mall Recordings)	Spontaneous	11 (0–9, oh)	2.0
Connected digit strings (AT&T HMIHY©)	Conversational	11 (0–9, oh)	5.0
Resource management (RM)	Read speech	1000	2.0
Airline travel information system (ATIS)	Spontaneous	2500	2.5
North American business (NAB & WSJ)	Read text	64,000	6.6
Broadcast News	Narrated News	210,000	≈15
Switchboard	Telephone conversation	45,000	≈27
Call-home	Telephone conversation	28,000	≈35

HMM in ASR

Fg2.1

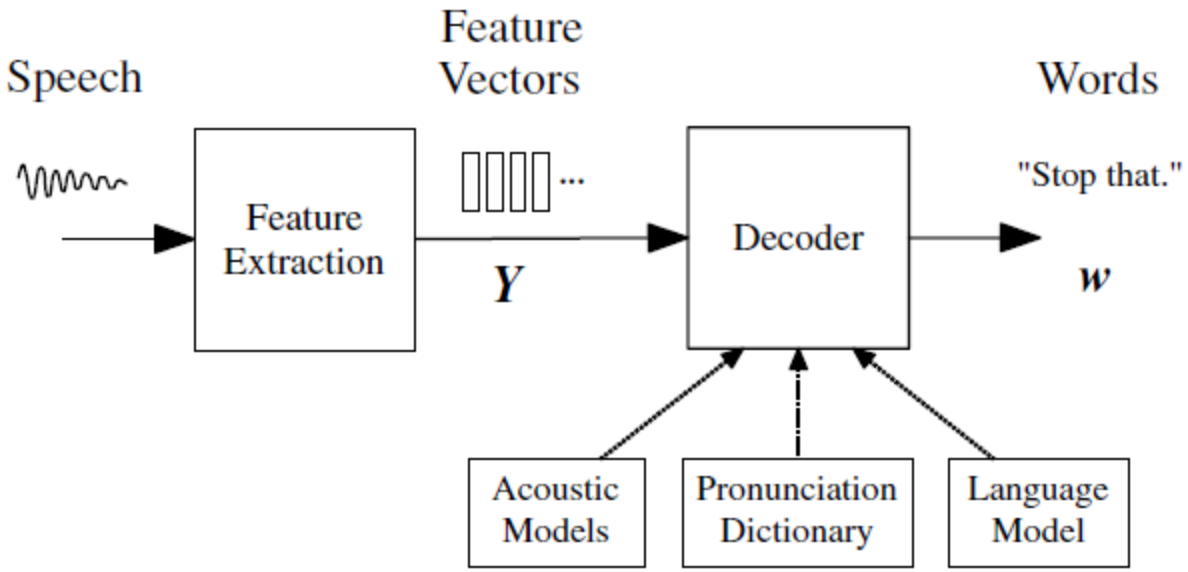

Fig. 2.1 Architecture of a HMM-based Recogniser.

Fg2.2

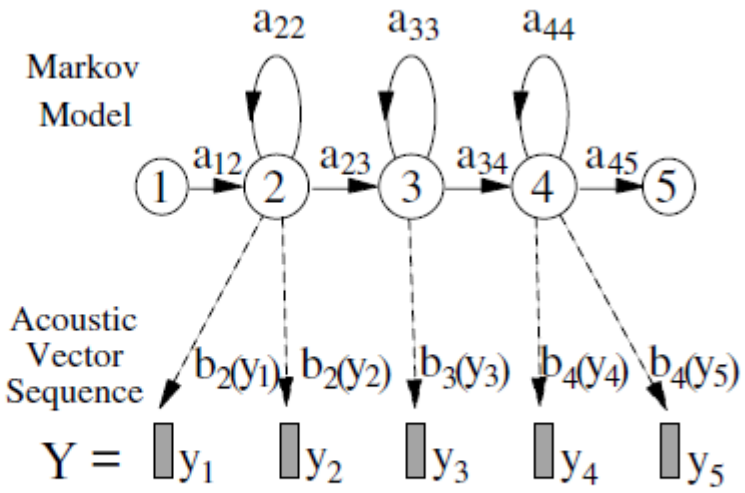

Fig. 2.2 HMM-based phone model.

Fg2.3

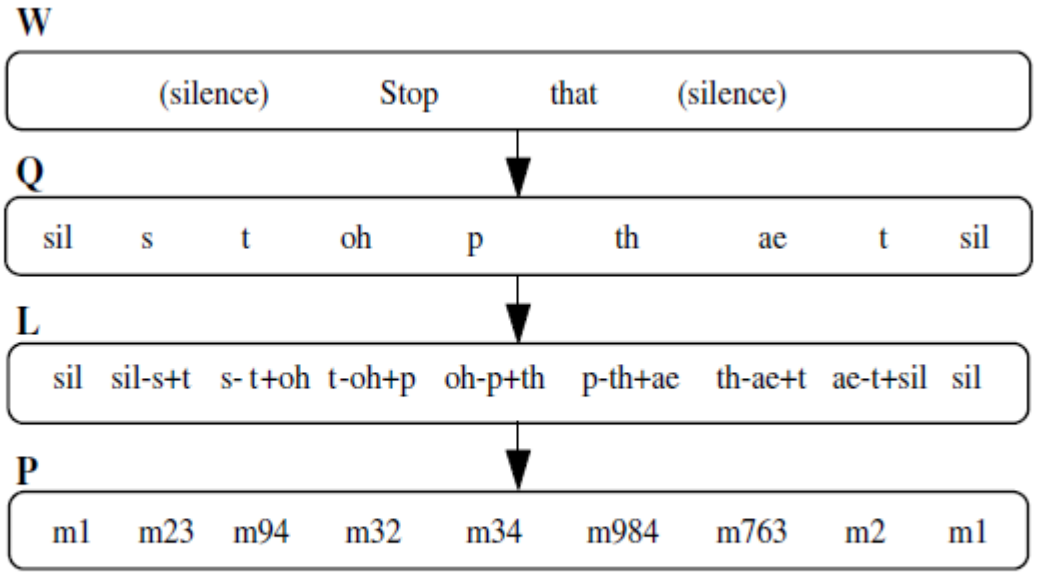

Fig. 2.3 Context dependent phone modelling.

Fg2.4

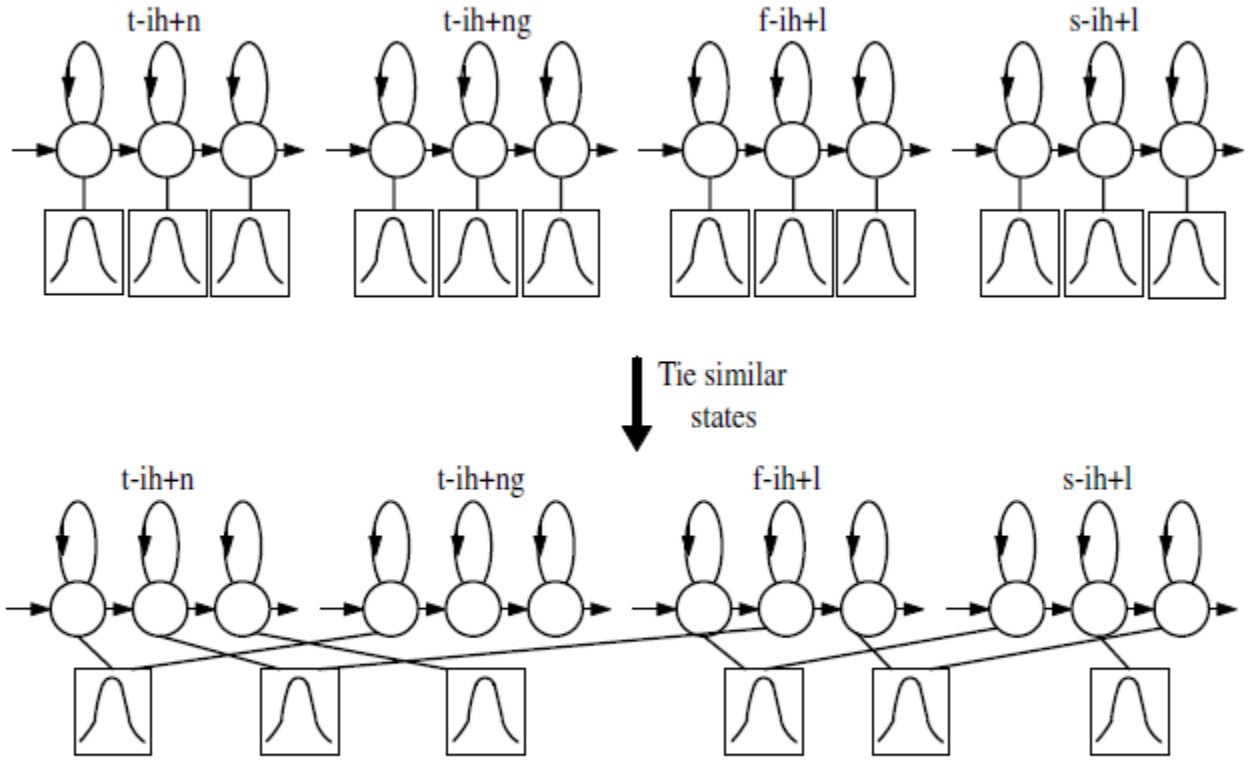

Fig. 2.4 Formation of tied-state phone models.

Fg2.5

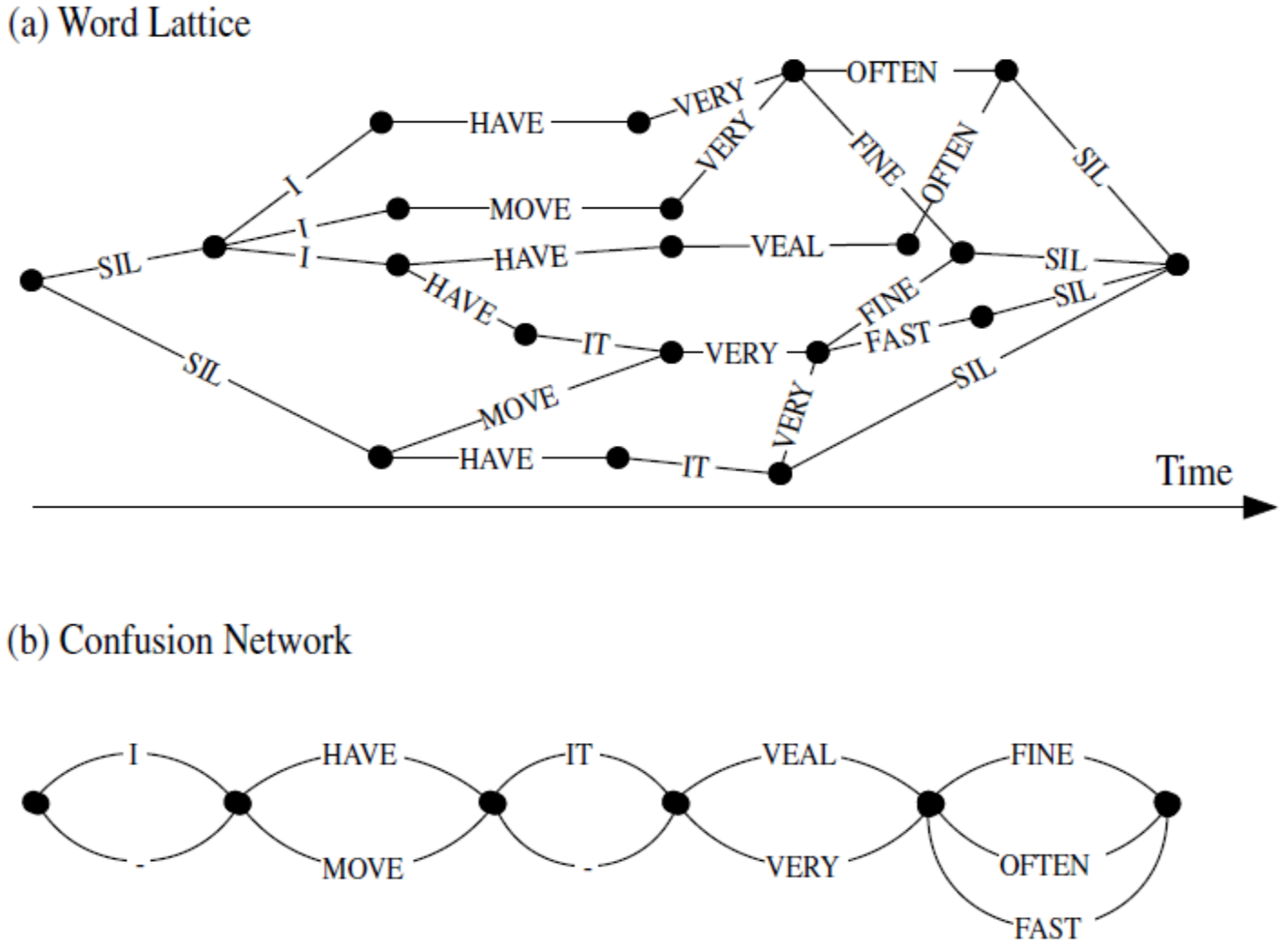

Fig. 2.6 Example lattice and confusion network.