

Speech Analysis Synthesis and Perception

Third Edition

James L. Flanagan
Jont B. Allen
Mark A. Hasegawa-Johnson

2008

Contents

1	Voice Communication	1
1.1	Speech as a Communication Channel	2
1.2	Entropy of the Speech Source	5
1.3	Conditional Entropy of Received Speech	5
1.4	Capacity of the Acoustic Channel	9
1.5	Organization of this Book	10
2	The Mechanism of Speech Production	13
2.1	Physiology of the Vocal Apparatus	13
2.2	The Sounds of Speech	17
2.2.1	Vowels	19
2.2.2	Consonants	19
2.3	Quantitative Description of Speech	25
3	Acoustical Properties of the Vocal System	27
3.1	The Vocal Tract as an Acoustic System	27
3.2	Equivalent Circuit for the Lossy Cylindrical Pipe	29
3.2.1	The Acoustic “L”	31
3.2.2	The Acoustic “R”	31
3.2.3	The Acoustic “C”	33
3.2.4	The Acoustic “G”	33
3.2.5	Summary of the Analogous Acoustic Elements	36
3.3	The Radiation Load at the Mouth and Nostrils	37
3.4	Spreading of Sound about the Head	39
3.5	The Source for Voiced Sounds	40
3.5.1	Glottal Excitation	40
3.5.2	Sub-Glottal Impedance	42
3.5.3	Glottal Impedance	43
3.5.4	Source-Tract Coupling Between Glottis and Vocal Tract	49
3.5.5	High-Impedance Model of the Glottal Source	51
3.6	Turbulent Noise Sources	51
3.7	The Source for Transient Excitation	53
3.8	Some Characteristics of Vocal Tract Transmission	56
3.8.1	Effect of Radiation Load upon Mode Pattern	58
3.8.2	Effect of Glottal Impedance upon Mode Pattern	60
3.8.3	Effect of Cavity Wall Vibration	61
3.8.4	Two-Tube Approximation of the Vocal Tract	64
3.8.5	Excitation by Source Forward in Tract	65
3.8.6	Effects of the Nasal Tract	70

3.8.7	Four-Tube, Three-Parameter Approximation of Vowel Production	73
3.8.8	Multitube Approximations and Electrical Analogs of the Vocal Tract	75
3.9	Fundamentals of Speech and Hearing in Analysis-Synthesis Telephony	77
4	Speech Synthesis	79
4.1	Mechanical Speaking Machines; Historical Efforts	79
4.2	Electrical Methods for Speech Synthesis	83
4.2.1	Spectrum Reconstruction Techniques	83
4.2.2	“Terminal Analog” Synthesizers	86
4.2.3	Transmission-Line Analogs of the Vocal System	97
4.2.4	Excitation of Electrical Synthesizers	100
4.2.5	Vocal Radiation Factors	121
4.2.6	Speech Synthesis by Computer Simulation	121
5	The Ear and Hearing	135
5.1	Mechanism of the Ear	135
5.1.1	The Outer Ear	136
5.1.2	The Middle Ear	136
5.1.3	The Inner Ear	139
5.1.4	Mechanical-to-Neural Transduction	140
5.1.5	Neural Pathways in the Auditory System	145
5.2	Computational Models for Ear Function	153
5.2.1	Basilar Membrane Model	154
5.2.2	Middle Ear Transmission	156
5.2.3	Combined Response of Middle Ear and Basilar Membrane	158
5.2.4	An Electrical Circuit for Simulating Basilar Membrane Displacement	160
5.2.5	Computer Simulation of Membrane Motion	162
5.2.6	Transmission Line Analogs of the Cochlea	166
5.3	Illustrative Relations between Subjective and Physiological Behavior	168
5.3.1	Pitch Perception	169
5.3.2	Binaural Lateralization	171
5.3.3	Threshold Sensitivity	175
5.3.4	Auditory Processing of Complex Signals	178
6	Perception of Speech and Speech-Like Sounds	181
6.1	Differential vs, Absolute Discrimination	182
6.2	Differential Discriminations Along Signal Dimensions Related to Speech	183
6.2.1	Limens for Vowel Formant Frequencies	183
6.2.2	Limens for Formant Amplitude	183
6.2.3	Limens for Formant Bandwidth	184
6.2.4	Limens for Fundamental Frequency	184
6.2.5	Limens for Excitation Intensity	184
6.2.6	Limens for Glottal Zeros	185
6.2.7	Discriminability of Maxima and Minima in a Noise Spectrum	185
6.2.8	Other Close-Comparison Measures Related to Speech	186
6.2.9	Differential Discriminations in the Articulatory Domain	189
6.3	Absolute Discrimination of Speech and Speech-Like Sounds	191
6.3.1	Absolute Identification of Phonemes	191
6.3.2	Absolute Identification of Syllables	193
6.3.3	Effects of Learning and Linguistic Association in Absolute Identification of Speech-Like Signals	199

6.3.4	Influence of Linguistic Association Upon Differential Discriminability	202
6.4	Effects of Context and Vocabulary Upon Speech Perception	204
6.5	The Perceptual Units of Speech	206
6.5.1	Models of Speech Perception	208
6.6	Subjective Evaluation of Transmission Systems	210
6.6.1	Articulation Tests	210
6.6.2	Quality Tests	211
6.7	Calculating Intelligibility Scores from System Response and Noise Level: The Articulation Index	214
6.8	Supplementary Sensory Channels for Speech Perception	215
6.8.1	Visible Speech Translator	216
6.8.2	Tactile Vocoder	216
6.8.3	Low Frequency Vocoder	218
7	Techniques for Speech Analysis	219
7.1	Spectral Analysis of Speech	220
7.1.1	Short-Time Frequency Analysis	220
7.1.2	Measurement of Short-Time Spectra	222
7.1.3	Choice of the Weighting Function, $h(t)$	225
7.1.4	The Sound Spectrograph	226
7.1.5	Short-Time Correlation Functions and Power Spectra	231
7.1.6	Average Power Spectra	235
7.1.7	Measurement of Average Power Spectra for Speech	236
7.2	Formant Analysis of Speech	237
7.2.1	Formant-Frequency Extraction	239
7.2.2	Measurement of Formant Bandwidth	251
7.3	Analysis of Voice Pitch	252
7.4	Articulatory Analysis of the Vocal Mechanism	255
7.5	Automatic Recognition of Speech	259
7.6	Automatic Recognition and Verification of Speakers	264
8	Statistical Speech Recognition	269
8.1	Classification of Short-Time Spectra	270
8.1.1	The Maximum A Posteriori Probability Classification Rule	270
8.1.2	Gaussian Models of the Speech Spectrum	270
8.1.3	Mixture Gaussian Models	270
8.1.4	Sources of Error in Pattern Classification	270
8.1.5	Linear and Discriminant Features	270
8.1.6	Multilayer and Kernel-Based Neural Networks	270
8.1.7	Talker Adaptation	270
8.2	Recognition of Words	270
8.2.1	Linear Time Warping	270
8.2.2	Dynamic Time Warping	270
8.2.3	Hidden Markov Models: Testing	270
8.2.4	Hidden Markov Models: Training	270
8.2.5	Pronunciation Modeling	270
8.2.6	Context-Dependent Recognition Units	270
8.2.7	Landmarks, Events, and Islands of Certainty	270
8.3	Recognition of Utterances	270
8.3.1	Static Search Graph: Finite State Methods	270
8.3.2	Regular Grammars for Dialog Systems	270

8.3.3	N-Grams and Backoff	270
8.3.4	Dynamic Search Graph: Stack-Based Methods	270
8.3.5	Dynamic Search Graph: Bayesian Networks	270
8.3.6	Multi-Pass Recognition	270
8.3.7	System Combination	270
9	Systems for Digital Transmission of Speech	271
9.1	Assessment of Speech Perceptual Quality	272
9.1.1	Subjective Measures	272
9.1.2	Objective Measures: Broadband	272
9.1.3	Objective Measures: Critical Band	272
9.1.4	Automatic Prediction of Subjective Measures	272
9.1.5	Computationally Efficient Measures	272
9.2	Quantization	272
9.3	Transform and Sub-Band Coding	272
9.3.1	Analytic Rooter	273
9.3.2	Transform Coding: Error Analysis	277
9.3.3	Expansion of the Speech Waveform	277
9.3.4	Expansion of the Short-Time Amplitude Spectrum	280
9.3.5	Expansion of the Short-Time Autocorrelation Function	281
9.4	Correlation Vocoders	286
9.4.1	Channel Vocoders	289
9.4.2	Design Variations in Channel Vocoders	290
9.4.3	Vocoder Performance	291
9.4.4	Linear Transformation of Channel Signals	296
9.4.5	Discrete Cosine Transform	297
9.4.6	Sub-Band Coder	297
9.4.7	Sinusoidal Transform Coding	297
9.4.8	“Peak-Picker”	297
9.5	Predictive Quantization	298
9.5.1	Delta Modulation	298
9.5.2	Predictive Coding of Speech	301
9.6	Parametric Models of the Spectral Envelope	310
9.6.1	Homomorphic Vocoders	310
9.6.2	Maximum Likelihood Vocoders	312
9.6.3	Linear Prediction Vocoders	315
9.6.4	Articulatory Vocoders	317
9.6.5	Pattern-Matching Vocoders	318
9.7	Formant Vocoders	319
9.8	Parametric Models of the Spectral Fine Structure	322
9.8.1	Voice-Excited Vocoders	322
9.8.2	Self-Excited LPC	324
9.8.3	Code-Excited LPC	324
9.8.4	Multipulse and Algebraic Codes	324
9.8.5	The LPC-10 Vocoder	324
9.8.6	Multiband and Harmonic Vector Excitations	324
9.8.7	Voice-Excited Formant Vocoders	324
9.8.8	Frequency-Dividing Vocoders	325
9.9	Rate-Distortion Tradeoffs for Speech Coding	327
9.9.1	Multiplexing and Digitalization	327
9.9.2	Multiplexing and Digitalization of Formant Vocoders	328

9.9.3	Time-Assignment Transmission of Speech	331
9.9.4	Multiplexing Channel Vocoders	333
9.9.5	Error Protection Coding	336
9.9.6	The Rate-Distortion Curve	336
9.9.7	Variable-Rate and Embedded Coding	336
9.9.8	Joint Source-Channel Coding	336
References	337

List of Figures

1.1	Conversation over lunch: Renoir's <i>Luncheon of the Boating Party</i> , 1881. (Phillips Collection, Washington D.C.)	3
1.2	Schematic diagram of a general communication system. X =source message, Y =received message, S =transmitted signal, R =received signal, N =noise. (After Shannon and Weaver, 1949)	4
1.3	Typical confusion matrix (6300Hz bandwidth, -6dB SNR). Entry (i, j) in the matrix lists the number of times that a talker said consonant x_i , and a listener heard consonant y_j . Each consonant was uttered as the first phoneme in a CV syllable; the vowel was always /a/. (After Miller and Nicely, 1955)	8
1.4	(a) Mutual information between spoken and perceived consonant labels, as a function of SNR, over an acoustic channel with 6300Hz bandwidth (200-6500Hz). (b) Mutual information between spoken and perceived consonant labels, at 12dB SNR, over lowpass and highpass acoustic channels with the specified cutoff frequencies. The lowpass channel contains information between 200Hz and the cutoff; bit rate is shown with a solid line. The highpass channel contains information between the cutoff and 6500Hz; bit rate is shown with a dashed line. (After Miller and Nicely, 1955)	8
2.1	Schematic diagram of the human vocal mechanism	14
2.2	Cut-away view of the human larynx. (After Farnsworth.) VC-vocal cords; AC-arytenoid cartilages; TC-thyroid cartilage	15
2.3	Technique for high-speed motion picture photography of the vocal cords. (After Farnsworth)	16
2.4	Successive phases in one cycle of vocal cord vibration. The total elapsed time is approximately 8 msec	16
2.5	Schematic vocal tract profiles for the production of English vowels. (Adapted from Potter, Kopp and Green)	20
2.6	Vocal tract profiles for the fricative consonants of English. The short pairs of lines drawn on the throat represent vocal cord operation. (Adapted from Potter, Kopp and Green)	22
2.7	Articulatory profiles for the English stop consonants. (After Potter, Kopp and Green)	23
2.8	Vocal profiles for the nasal consonants. (After Potter, Kopp and Green)	23
2.9	Vocal tract configurations for the beginning positions of the glides and semivowels. (After Potter, Kopp and Green)	24
3.1	Schematic diagram of functional components of the vocal tract	28
3.2	Incremental length of lossy cylindrical pipe. (a) acoustic representation; (b) electrical equivalent for a one-dimensional wave	29
3.3	Equivalent four-pole networks for a length l of uniform transmission line. (a) T-section; (b) π -section	30
3.4	Relations illustrating viscous loss at the wall of a smooth tube	32

3.5	Relations illustrating heat conduction at the wall of a tube	34
3.6	Normalized acoustic radiation resistance and reactance for (a) circular piston in all infinite baffle; (b) circular piston in a spherical baffle whose radius is approximately three times that of the piston; (c) pulsating sphere. The radius of the radiator, whether circular or spherical, is a	38
3.7	Spatial distributions of sound pressure for a small piston in a sphere of 9cm radius. Pressure is expressed in db relative to that produced by a simple spherical source of equal strength	40
3.8	Life-size mannequin for measuring the relation between the mouth volume velocity and the sound pressure at an external point. The transducer is mounted in the mannequin's head.	41
3.9	Distribution of sound pressure about the head, relative to the distribution for a simple source; (a) horizontal distribution for the mannequin; (b) vertical distribution for the mannequin	41
3.10	Schematic diagram of the human subglottal system	42
3.11	An equivalent circuit for the subglottal system	42
3.12	Simple orifice approximation to the human glottis	43
3.13	Model of the human glottis. (After Berg)	45
3.14	Simplified circuit for the glottal source	45
3.15	Ratios of glottal inertance (L_g) to viscous and kinetic resistance (R_v, R_k) as a function of glottal area (A)	47
3.16	Glottal area and computed volume velocity waves for single vocal periods. F_0 is the fundamental frequency; P_s is the subglottal pressure. The subject is an adult male phonating TIPA/æ/. (After Flanagan, 1958)	48
3.17	Calculated amplitude spectrum for the glottal area wave AII shown in Fig. 3.16. (After Flanagan, 1961)	48
3.18	Small-signal equivalent circuit for the glottal source. (After Flanagan, 1958)	49
3.19	Simplified representation of the impedance looking into the vocal tract at the glottis	50
3.20	Equivalent circuit for noise excitation of the vocal tract	53
3.21	Approximate vocal relations for stop consonant production	54
3.22	Relation between glottal and mouth volume currents for the unconstricted tract. The glottal impedance is assumed infinite and the radiation impedance is zero	56
3.23	Magnitude and phase of the glottis-to-mouth transmission for the vocal tract approximation shown in Fig. 3.22	58
3.24	Equivalent circuit for the unconstricted vocal tract taking into account the radiation load. The glottal impedance is assumed infinite	58
3.25	Equivalent circuit for the unconstricted vocal tract assuming the glottal impedance to be finite and the radiation impedance to be zero	60
3.26	Representation of wall impedance in the equivalent T-section for a length l of uniform pipe	62
3.27	Two-tube approximation to the vocal tract. The glottal impedance is assumed infinite and the radiation impedance zero	64
3.28	Two-tube approximations to the vowels /i,æ,a,	66
3.29	First formant ($F1$) versus second formant ($F2$) for several vowels. Solid points are averages from Peterson and Barney's (1952) data for real speech uttered by adult males. Circles are for the two-tube approximation to the vowels shown in Fig. 3.28	66
3.30	Two-tube approximation to the vocal tract with excitation applied forward of the constriction	67
3.31	Two-tube approximation to the fricative TIPA/s/. The undamped pole-zero locations are obtained from the reactance plots	68
3.32	Measured spectra for the fricative TIPA/s/ in real speech. (After Hughes and Halle)	68

3.33	Two-tube approximation to the vocal tract with the source of excitation applied at the tube junction	69
3.34	Measured spectra for the fricative TIPA/f/ in real speech. (After Hughes and Halle)	70
3.35	An equivalent circuit for the combined vocal and nasal tracts. The pharynx, mouth and nasal cavities are assumed to be uniform tubes.	70
3.36	A simple approximation to the vocal configuration for the nasal consonant TIPA/m/	71
3.37	Reactance functions and undamped mode pattern for the articulatory approximation to TIPA/m/ shown in Fig. 3.36	72
3.38	Measured spectrum for the nasal consonant TIPA/m/ in real speech. (After Fant, 1960)	72
3.39	Nomogram for the first three undamped modes (F_1, F_2, F_3) of a fourtube approximation to the vocal tract (Data adapted from Fant, 1960). The parameter is the mouth area, A_4 . Curves 1, 2, 3 and 4 represent mouth areas of 4, 2, 0.65 and 0.16 cm ² , respectively. Constant quantities are $A_l = A_3 = 8$ cm ² , $l_4 = 1$ cm and $A_2 = 0.65$ cm ² . Abscissa lengths are in cm	73
4.1	WHEATSTONE'S construction of VON KEMPELEN'S speaking machine	80
4.2	Mechanical vocal tract of RIESZ	81
4.3	Key control of RIESZ's mechanical talker	82
4.4	(a) Mechanical model of the vocal tract for simulating fricative consonants. (b) Measured sound spectrum for a continuant sound similar to /	82
4.5	Schematic diagram of the Voder synthesizer (After DUDLEY, RIESZ and WATKINS)	84
4.6	(a) Functional diagram of a spectrogram play-back device. (After COOPER.) (b) Spectrograms of real speech and an abstracted, hand-painted version of the same. Both displays can be synthesized on the pattern play-back machine. (After BORST)	85
4.7	Feedback circuit for producing a transmission having uniformly spaced complex conjugate poles	88
4.8	Front excitation of a straight pipe by a pressure source	90
4.9	Simplified configuration illustrating coupling between oral and nasal cavities	92
4.10	(a) Cascade connection of isolated RLC resonators for simulation of vocal transmission for vowel sounds. Each pole-pair or vocal resonance is simulated by a series circuit. (b) Cascaded pole and zero circuit for simulating low frequency behavior of a side branch resonator. The zero pair is approximated by the transmission of a simple series circuit	93
4.11	Circuit operations for simulating the time-domain response of Eq. (4.30)	95
4.12	Circuit for simulating the vowel function impulse response [see Eq. (4.33)]	96
4.13	T-circuit equivalents for a length l of uniform cylindrical pipe. (a) Exact circuit, (b) first-term approximations to the impedance elements	98
4.14	Ladder network approximations to the vocal tract. The impedance elements of the network are those shown in Fig. 4.13b	98
4.15	Continuously controllable transmission line analog of the vocal system. (After ROSEN; HECKER) Speech Synthesis	99
4.16	Single periods of measured glottal area and calculated volume velocity functions for two men (A and B) phonating the vowel /æ/ under four different conditions of pitch and intensity. F_0 is the fundamental frequency and P_s the sub glottal pressure. The velocity wave is computed according to the technique described in Section 3.5.3. (After FLANAGAN, 1958)	101
4.17	Triangular approximation to the glottal wave. The asymmetryfactor is k	101
4.18	Complex frequency loci of the zeros of a triangular pulse. The s -plane is normalized in terms of $\omega\tau_0$ and $\sigma\tau_0$. The asymmetry constant k is the parameter. (After DUNN, FLANAGAN and GESTRIN)	104

4.19	Imaginary parts of the complex zeros of a triangular pulse as a function of asymmetry. The imaginary frequency is normalized in terms of $\omega\tau_0$ and the range of asymmetry is $0 \leq k \leq \infty$. (After DUNN, FLANAGAN and GESTRIN)	105
4.20	Amplitude spectra for two triangular pulses, $k = 1$ and $k = 11/12$. (After DUNN, FLANAGAN and GESTRIN)	106
4.21	Four symmetrical approximations to the glottal pulse and their complex zeros	106
4.22	Effect of glottal zeros upon the measured spectrum of a synthetic vowel sound. (a) $\tau_0 = 4.0$ msec. (b) $\tau_0 = 2.5$ msec, (After FLANAGAN, 1961b)	108
4.23	Method for manipulating source zeros to influence vowel quality. Left column, no zeros. Middle column, left-half plane zeros. Right column, right-half plane zeros. (After FLANAGAN, 1961b)	109
4.24	Best fitting pole-zero model for the spectrum of a single pitch period of a natural vowel sound. (After MATHEWS, MILLER and DAVID, 1961b)	110
4.25	Schematic diagram of the human vocal mechanism. (After FLANAGAN et al., 1970.)	112
4.26	Network representation of the vocal system	112
4.27	Acoustic oscillator model of the vocal cords. (After FLANAGAN and LANDGRAF)	113
4.28	Simplified network of the vocal system for voiced sounds. (After FLANAGAN and LANDGRAF)	114
4.29	Glottal area and acoustic volume velocity functions computed from the vocal-cord model. Voicing is initiated at $t = 0$	114
4.30	Spectrogram of a vowel-vowel transition synthesized from the cord oscillator and vocal tract model. The output corresponds to a linear transition from the vowel /i/ to the vowel /a/. Amplitude sections are shown for the central portion of each vowel	116
4.31	Modification of network elements for simulating the properties of turbulent flow in the vocal tract. (After FLANAGAN and CHERRY)	116
4.32	Waveforms of vocal functions. The functions are calculated for a voiced fricative articulation corresponding to the constricted vowel /a/. (After FLANAGAN and CHERRY)	117
4.33	Sound spectrograms of the synthesized output for a normal vowel /a/ (left) and the constricted /a/ shown in Fig. 4.32 (right). Amplitude sections are shown for the central portion of each vowel	118
4.34	Spectrograms for the voiced-voiceless cognates /	119
4.35	Sound spectrogram for the synthesized syllable /	120
4.36	Digital operations for simulating a single formant resonance (pole-pair) (a) implementation of the standard z -transform; (b) practical implementation for unity dc gain and minimum multiplication	125
4.37	Digital operations for simulating a single anti-resonance (zero-pair)	126
4.38	Block diagram of a computer-simulated speech synthesizer. (After FLANAGAN, COKER, and BIRD)	126
4.39	Spectrograms of synthetic speech produced by a computer-simulated formant synthesizer and of the original utterance. (After FLANAGAN, COKER and BIRD)	127
4.40	Spectrograms comparing natural speech synthesized directly from printed text. (After COKER, UMEDA and BROWMAN)	130
4.41	Programmed operations for synthesis from stored formant data. (After RABfNER, SCHAFER and FLANAGAN)	130
4.42	Computer synthesis by concatenation of formant coded words. (After RABINER, SCHAFER and FLANAGAN)	130
4.43	Ladder network corresponding to a difference-equation approximation of the Webster wave equation	132
4.44	Representation of an impedance discontinuity in terms of reflection coefficients . . .	133

5.1	Schematic diagram of the human ear showing outer, middle and inner regions. The drawing is not to scale. For illustrative purposes the inner and middle ear structures are shown enlarged	136
5.2	Vibration modes of the ossicles. (a) sound intensities below threshold of feeling (b) intensities above threshold of feeling. (After BEKESY, 1960)	137
5.3	Data on middle ear transmission; effective stapes displacement for a constant sound pressure at the eardrum. (a) BÉKÉSY (1960) (one determination); (b) BÉKÉSY (1960) (another determination); (c) measured from an electrical analog circuit (after ZWISLOCKI, 1959); (d) measured from an electrical analog circuit (after MOLLER, 1961)	138
5.4	Simplified diagram of the cochlea uncoiled	139
5.5	Schematic cross section of the cochlear canal. (Adapted from Davis, 1957)	140
5.6	Amplitude and phase responses for basilar membrane displacement. The stapes is driven sinusoidally with constant amplitude of displacement. (After BÉKÉSY, 1960.) (a) Amplitude vs frequency responses for successive points along the membrane. (b) Amplitude and phase responses for the membrane place maximally responsive to 150 cps. (c) Amplitude and phase of membrane displacement as a function of distance along the membrane. Frequency is the parameter	141
5.7	Cross section of the organ of Corti. (After DAVIS, 1951)	141
5.8	Distribution of resting potentials in the cochlea. Scala tympani is taken as the zero reference. The tectorial membrane is not shown. The interiors of all cells are strongly negative. (After TASAKI, DAVIS and ELDREDGE)	144
5.9	Cochlear microphonic and dc potentials recorded by a microelectrode penetrating the organ of Corti from the scala tympani side. The cochlear microphonic is in response to a 500 cps tone. (After DAVIS, 1965)	144
5.10	A “resistance microphone” theory of cochlear transduction. (After DAVIS, 1965)	145
5.11	Schematic diagram of the ascending auditory pathways. (Adapted from drawings by NETTER)	146
5.12	Electrical firings from two auditory nerve fibers. The characteristic frequency of unit 22 is 2.3 kcps and that for unit 24 is 6.6 kcps, The stimulus is 50 msec bursts of a 2.3 kcps tone. (After KIANG et al.)	147
5.13	Frequency sensitivities for six different fibers in the auditory nerve of cat. (After KIANG et al.)	147
5.14	Electrical response of a single auditory nerve fiber (unit) to 10 successive rarefaction pulses of 100 μ sec duration. <i>RW</i> displays the cochlear microphonic response at the round window. <i>CF</i> = 540 cps. (After KIANG et al.)	149
5.15	Post stimulus time (PST) histogram for the nerve fiber shown in Fig. 5.14. <i>CF</i> = 540 cps. Stimulus pulses 10 sec ⁻¹ . (After KIANG et at.)	150
5.16	Characteristic period (1/ <i>CF</i>) for 56 different auditory nerve fibers plotted against the interpeak interval measured from PST histograms. (After KIANG et at.)	150
5.17	Responses of a single auditory neuron in the trapezoidal body of cat. The stimulus was tone bursts of 9000 cps produced at the indicated relative intensities. (After KATSUKI)	150
5.18	Relation between sound intensity and firing (spike) frequency for single neurons at four different neural stages in the auditory tract of cat. Characteristic frequencies of the single units: Nerve: 830 cps; Trapezoid: 9000 cps; Cortex: 3500 cps; Geniculate: 6000 cps.(After KATSUKI)	151
5.19	Sagittal section through theleft cochlear complex in cat. The electrode followed the track visible just above the ruled line. Frequencies of best response of neurons along thetrack are indicated. (After ROSE, GALAMBOS and HUGHES)	151

5.20	Intensity vs frequency" threshold" responses for single neurons in the cochlear nucleus of cat. The different curves represent the responses of different neurons. (a) Units with narrow response areas; (b) units with broad response areas. (After ROSE, GALAMBOS and HUGHES)	152
5.21	Schematic diagram of the peripheral ear. The quantities to be related analytically are the eardrum pressure, $p(t)$; the stapes displacement, $x(t)$; and the basilar membrane displacement at distance l from the stapes, $y_l(t)$	153
5.22	(a) Pole-zero diagram for the approximating function $F_l(s)$ (After FLANAGAN, 1962a). (b) Amplitude and phase response of the basilar membrane model $F_l(s)$. Frequency is normalized in terms of the characteristic frequency β_l	155
5.23	Response of the basilar membrane model to an impulse of stapes displacement . . .	155
5.24	Functional approximation of middle ear transmission. The solid curves are from an electrical analog by ZWISLOCKI (see Fig. 5.3c). The plotted points are amplitude and phase values of the approximating function $G(s)$. (FLANAGAN, 1962a)	156
5.25	Displacement and velocity responses of the stapes to an impulse of pressure at the eardrum	157
5.26	Displacement responses for apical, middle and basal points on the membrane to an impulse of pressure at the eardrum. The responses are computed from the inverse transform of $[G(s)F_l(s)]$	159
5.27	(a) Amplitude <i>vs</i> frequency responses for the combined model. (b) Phase <i>vs</i> frequency responses for the combined model	160
5.28	Electrical network representation of the ear model	161
5.29	(a) Impulse responses measured on the network of Fig. 5.28. (b) First difference approximations to the spatial derivative measured from the network of Fig. 5.28 . .	162
5.30	Sampled-data equivalents for the complex conjugate poles, real-axis pole, and real-axis zero	163
5.31	Functional block diagram for a digital computer simulation of basilar membrane displacement	163
5.32	Digital computer simulation of the impulse responses for 40 points along the basilar membrane. The input signal is a single rarefaction pulse, $100\mu\text{sec}$ in duration, delivered to the eardrum at time $t = 0$. (After FLANAGAN, 1962b)	165
5.33	Digital computer output for 40 simulated points along the basilar membrane. Each trace is the displacement response of a given membrane place to alternate positive and negative pressure pulses. The pulses have $100\mu\text{sec}$ duration and are produced at a rate of 200 sec^{-1} . The input signal is applied at the eardrum and is initiated at time zero. The simulated membrane points are spaced by 0.5mm . Their characteristic frequencies are indicated along the ordinate. (After FLANAGAN, 1962b)	165
5.34	Idealized schematic of the cochlea. (After PETERSON and BOGERT)	166
5.35	Instantaneous pressure difference across the cochlear partition at successive phases in one period of a 1000 cps excitation. (After PETERSON and BOGERT)	167
5.36	Electrical network section for representing an incremental length of the cochlea. (After BOGERT)	168
5.37	Comparison of the displacement response of the transmission line analog of the cochlea to physiological data for the ear. (After BOGERT)	168
5.38	Membrane displacement responses for filtered and unfiltered periodic pulses. The stimulus pulses are alternately positive and negative. The membrane displacements are simulated by the electrical networks shown in Fig. 5.28. To display the waveforms more effectively, the traces are adjusted for equal peak-to-peak amplitudes. Relative amplitudes are therefore not preserved	170

5.39	Basilar membrane responses at the 2400, 1200 and 600 cps points to a pressure-rarefaction pulse of $100\mu\text{sec}$ duration. The responses are measured on the electrical analog circuit of Fig. 5.28. Relative amplitudes are preserved	172
5.40	Experimental arrangement for measuring the interaural times that produce centered sound images. (After FLANAGAN, DAVID and WATSON)	173
5.41	Experimentally measured interaural times for lateralizing cophasic and antiphase clicks. Several conditions of masking are shown. (a) Unmasked and symmetrically masked conditions. (b) Asymmetrically masked conditions. The arrows indicate the interaural times predicted from the basilar membrane model	174
5.42	Relation between the mechanical sensitivity of the ear and the monaural minimum audible pressure threshold for pure tones	176
5.43	Average number of ganglion cells per mm length of organ of Corti. (After GUILD et al.)	176
5.44	Binaural thresholds of audibility for periodic pulses. (After FLANAGAN, 1961a) . .	177
5.45	Model of the threshold of audibility for the pulse data shown in Fig. 5.44	177
6.1	Detectability of irregularities in a broadband noise spectrum. (After MALME) . . .	186
6.2	Frequency paths and excitation pattern for a simulated time-varying formant. Rising and falling resonances are used. The epochs of the five excitation pulses are shown. (After BRADY, HOUSE and STEVENS)	187
6.3	Results of matching a nontime-varying resonance to the time-varying resonances shown in Fig. 6.2. Mean values are plotted. The vertical lines indicate the standard deviations of the matches. (After BRADY, HOUSE and STEVENS)	187
6.4	Periodic pulse stimuli for assessing the influence of amplitude and time perturbations upon perceived pitch. The left column shows the time waveforms of the experimental trains; amplitude variation (A_L), time variation (A_T), and the standard matching train (B). The second column shows the corresponding amplitude spectra, and the third column shows the complex-frequency diagram. (After FLANAGAN, GUTTMAN and WATSON; GUTTMAN and FLANAGAN, 1962)	188
6.5	Results of matching the pitch of a uniform pulse train (B) to that of: (a) a periodic train (A_L) whose alternate pulses differ in amplitude by ΔL and (b) a periodic train (A_T) whose alternate pulses are shifted in time by ΔT . In both cases the parameter is the pulse rate of the A stimulus. (After FLANAGAN, GUTTMAN and WATSON; GUTTMAN and FLANAGAN, 1962)	190
6.6	Three-parameter description of vowel articulation. r_0 is the radius of the maximum constriction; x_0 is the distance from the glottis to the maximum constriction; and A/l is the ratio of mouth area to lip rounding. (After STEVENS and HOUSE, 1955) . .	191
6.7	Listener responses to isolated synthetic vowels described by the 3-parameter technique. One value of constriction is shown. Two levels of response corresponding to 50 and 75% agreement among subjects are plotted. (After HOUSE and STEVENS, 1955) 192	
6.8	Formant frequency data of PETERSON and BARNEY for 33 men transformed into the 3-parameter description of vowel articulation. (After HOUSE and STEVENS, 1955) 193	
6.9	Stimulus patterns for determining the effect of noise-burst frequency on the perception of voiceless stop consonants: (a) frequency positions of the noise bursts, (b) formant frequencies of the two-formant vowels; (c) one of the synthetic consonant-vowel syllables formed by pairing a noise burst of (a) with a two-formant vowel of (b). (After COOPER, DELATTRE, LIBERMAN, BORST and GERSTMAN)	194
6.10	Listener responses to the synthetic consonant-vowel syllables shown in Fig. 6.9. (After COOPER <i>et al.</i>)	194
6.11	Second-formant trajectories for testing the contribution of formant transitions to the perception of voiceless stop consonants. (After COOPER et al.)	195

6.12	Median responses of 33 listeners to stop consonant and vowel syllables generated by the patterns shown in Fig. 6.11. The bars show the quartile ranges. (After COOPER et al.)	196
6.13	Listener responses in absolute identification of synthetic fricatives produced by a pole-zero filtering of noise. The frequency of the pole is indicated on the abscissa, and the frequency of the zero is approximately one octave lower. (After HEINZ and STEVENS)	197
6.14	Abstracted spectrogram showing the synthesis of a syllable with fricative consonant and vowel. The single fricative resonance is F_f . The four-formant vowel is an approximation of /	198
6.15	Absolute identifications of the initial consonant in the synthetic syllable schematized in Fig. 6.14. Two response contours are shown corresponding to 90 and 75% identification. Two consonant-to-vowel intensities (-5 and -25 db) are shown. (After HEINZ and STEVENS)	199
6.16	Median probability of correct response for frequency-coded, one-dimensional stimuli. (After HOUSE, STEVENS, SANDEL and ARNOLD)	201
6.17	Median probability of correct response for time-frequency-intensity coded three-dimensional stimuli. (After HOUSE, STEVENS, SANDEL and ARNOLD)	201
6.18	Synthetic two-formant syllables with formant transitions spanning the ranges for the voiced consonants lb, d, g/. The vowel is the same for each syllable and is representative of lei. (After LIBERMAN, HARRIS, HOFFMAN and GRIFFITH)	203
6.19	Absolute Consonant identifications of one listener for the stimuli of Fig. 6.18. (After LIBERMAN et al.)	203
6.20	ABX responses of the listener whose absolute responses are shown in Fig. 6.19. The step size between A and B stimuli was two positions in the stimulus set of Fig. 6.18. (After LIBERMAN et al.)	204
6.21	Intelligibility scores for different types of spoken material as a function of signal-to-noise ratio. (After MILLER, HEISE and LICHTEN)	205
6.22	Effects of vocabulary size upon the intelligibility of monosyllabic words. (After MILLER, HEISE and LICHTEN)	205
6.23	Block diagram model of stages in speech perception. (After BONDARKO, ZAGORUYKO, KOZHEVNIKOV, MOLCHANOV and CHISTOVICH)	208
6.24	A relation between word articulation score and sentence intelligibility. Sentences are scored for meaning conveyed. (After EGAN)	210
6.25	(a) Subject vectors obtained from a multi-dimensional scaling analysis projected onto the two most important perceptual dimensions I and III. The data are for a tone ringer experiment. (b) Preference judgments on 81 tone-ringer conditions, projected onto the two most important perceptual dimensions I and III. Direction of high preference is indicated by the vectors in Fig. 6.25a. (After BRICKER and FLANAGAN)	213
6.26	Diagram for calculating the articulation index. (After BERANEK)	214
6.27	Several experimental relations between articulation index and speech intelligibility (After KRYTER)	215
6.28	Block diagram of a tactile vocoder. (After PICKETT)	217
6.29	A frequency-dividing tactile vocoder. (After KRINGLEBOTN)	217
7.1	Weighting of an on-going signal $f(t)$ by a physically realizable time window $h(t)$. λ is a dummy integration variable for taking the Fourier transform at any instant, t	221
7.2	A method for measuring the short-time amplitude spectrum $ F(\omega, t) $	222
7.3	Alternative implementation for measuring the short-time amplitude spectrum $ F(\omega, t) $	223
7.4	Practical measurement of the short-time spectrum $ F(\omega, t) $ by means of a bandpass filter, a rectifier and a smoothing network	223

7.5	Short-time amplitude spectra of speech measured by a bank of 24 band-pass filters. A single filter channel has the configuration shown in Fig. 7.4. The spectral scans are spaced by 10 msec in time. A digital computer was used to plot the spectra and to automatically mark the formant frequencies. (After FLANAGAN, COKER and BIRD)	224
7.6	The effective time window for short-time frequency analysis by the basilar membrane in the human ear. The weighting function is deduced from the ear model discussed in Chapter IV	226
7.7	Functional diagram of the sound spectrograph	227
7.8	(a) Broadband sound spectrogram of the utterance "That you may see." (b) Amplitude vs frequency plots (amplitude sections) taken in the vowel portion of "that" and in the fricative portion of "see." (After BARNEY and DUNN)	228
7.9	Articulatory diagrams and corresponding broad-band spectrograms for the vowels li, te, a, ui as uttered by adult male and female speakers. (After POTTER, Kopp and GREEN)	229
7.10	Mean formant frequencies and relative amplitudes for 33 men uttering the English vowels in an /h-d/ environment. Relative formant amplitudes are given in db <i>re</i> the first formant of /	230
7.11	Method for the measurement of the short-time correlation function $\psi(\tau, t)$	232
7.12	Circuit for measuring the running short-time correlation function $\phi(\tau, t)$	233
7.13	Arrangement for measuring the short-time spectrum $Q(\omega, t)$. (After SCHROEDER and ATAL)	234
7.14	Circuit for measuring the long-time average power spectrum of a signal	237
7.15	Root mean square sound pressures for speech measured in -ll sec intervals 30 cm from the mouth. The analyzing filter bands are one-half octave wide below 500 cps and one octave wide above 500 cps. (After DUNN and WHITE.) The parameter is the percentage of the intervals having levels greater than the ordinate	237
7.16	Long-time power density spectrum for continuous speech measured 30 cm from the mouth. (After DUNN and WHITE)	238
7.17	Sound spectrogram showing idealized tracks for the first three speech formants . . .	238
7.18	Automatic formant measurement by zero-crossing count and adjustable prefiltering. (After CHANG)	240
7.19	Spectrum scanning method for automatic extraction of formant frequencies (After FLANAGAN, 1956a)	241
7.20	Peak-picking method for automatic tracking of speech formants. (After FLANAGAN, 1956a)	241
7.21	Formant outputs from the tracking device shown in Fig. 7.20. In this instance the boundaries of the spectral segments are fixed	242
7.22	Spectral fit computed for one pitch period of a voiced sound. (After MATHEWS, MILLER and DAVID, 1961b)	243
7.23	Tracks for the first and second formant frequencies obtained from a computer-analysis of real-time spectra. The speech samples are (a) "Hawaii" and (b) "Yowie" uttered by a man. (After HUGHES)	244
7.24	Computer procedure for formant location by the "analysis-by-synthesis" method. (After BELL et at.)	244
7.25	Idealized illustration of formant location by the "analysis-by-synthesis" method shown in Fig. 7.24	244
7.26	Computer-determined formant tracks obtained by the "analysis-by-synthesis" method. (a) Spectrogram of original speech. (b) Extracted formant tracks and square error measure. (After BELL et at.)	245
7.27	Spectrum and cepstrum analysis of voiced and unvoiced speech sounds. (After SCHAFFER and RABINER)	246

7.28	Cepstrum analysis of continuous speech. The left column shows cepstra of consecutive segments of speech separated by 20 ms. The right column shows the corresponding short-time spectra and the cepstrally-smoothed spectra	248
7.29	Enhancement of formant frequencies by the Chirp- z transform: (a) Cepstrally-smoothed spectrum in which F_2 and F_3 are not resolved. (b) Narrow-band analysis along a contour passing closer to the poles. (After SCHAFER and RABINER)	249
7.30	Automatic formant analysis and synthesis of speech. (a) and (b) Pitch period and formant frequencies analyzed from natural speech. (c) Spectrogram of the original speech. (d) Spectrogram of synthesis speech. (After SCHAFER and RABINER)	250
7.31	Pole-zero computer analysis of a speech sample using an articulatory model for the spectral fitting procedure. The (a) diagram shows the pole-zero positions calculated from the articulatory model. The (b) diagram shows the articulatory parameters which describe the vocal tract area function. (After HEINZ, 1962a)	251
7.32	Measured formant bandwidths for adult males. (After DUNN, 1961)	252
7.33	(a) Vocal-tract frequency response measured by sine-wave excitation of an external vibrator applied to the throat. The articulatory shape is for the neutral vowel and the glottis is closed. (After FUJIMURA and LINDQUIST). (b) Variation in first-formant bandwidth as a function of formant frequency. Data for men and women are shown for the closed-glottis condition. (After FUJIMURA and LINDQUIST)	253
7.34	Sagittal plane X-ray of adult male vocal tract	256
7.35	Method of estimating the vocal tract area function from X-ray data. (After FANT, 1960)	256
7.36	Typical vocal area functions deduced for several sounds produced by one man. (After FANT, 1960)	257
7.37	Typical vocal-tract area functions (solid curves) determined from impedance measurements at the mouth. The actual area functions (dashed curves) are derived from X-ray data. (After GOPINATH and SONDHI)	258
7.38	Seven-parameter articulatory model of the vocal tract. (After COKER)	258
7.39	Comparison of vocal tract area functions generated by the articulatory model of Fig. 7.38 and human area data from X-rays. (After COKER)	259
7.40	Principle of operation of a spoken digit recognizer. (After DAVIS, BIDDULPH and BALASHEK)	260
7.41	Scheme for automatic recognition of spectral patterns and spoken digits. (After DUDLEY and BALASHEK)	261
7.42	Block diagram of speech sound recognizer employing elementary linguistic constraints. (After FRY and DENES)	262
7.43	Effects of nonlinear warp in registering speech parameter patterns. The dashed curves are reference data for an individual. The solid curves are a sample utterance from the same individual. (a) Linear stretch to align end points only. (b) Nonlinear warp to maximize the correlation of the F_2 patterns. (After DODDINGTON)	268
9.1	Source-system representation of speech production	272
9.2	Diagram for computer simulation of the analytic rooter. (After SCHROEDER, FLANAGAN and LUNDRY)	275
9.3	Sound spectrograms of speech analyzed and synthesized by the analytic rooter. The transmission bandwidth is one-half the original signal bandwidth. (After SCHROEDER, FLANAGAN and LUNDRY)	277
9.4	System for transmitting speech waveforms in terms of orthogonal functions. (After MANLEY and KLEIN.) (a) Analyzer. (b) Synthesizer	279
9.5	Method for describing and synthesizing the short-time speech spectrum in terms of Fourier coefficients. (After PIROGOV)	280

9.6	Techniques for realizing the variable electrical network of Fig. 9.5	282
9.7	Expansion coefficients for the short-time auto-correlation function	283
9.8	Realization of Laguerre functions by RC networks [see Eq. (9.42)]	285
9.9	Plot of the final factor in Eq. (9.46) showing how the positive frequency range is spanned by the first several Laguerre functions. (After MANLEY)	285
9.10	A Laguerre function vocoder. (a) Analyzer. (b) Synthesizer. (After KULYA)	287
9.11	Autocorrelation vocoder. (After SCHROEDER, 1959, 1962)	288
9.12	Block diagram of the original spectrum channel vocoder. (After DUDLEY, 1939b)	289
9.13	Spectrogram of speech transmitted by a 15-channel vocoder	290
9.14	Filtering of a speech signal by contiguous band-pass filters	292
9.15	Speech synthesis from short-time amplitude and phase-derivative spectra. (After FLANAGAN and GOLDEN)	293
9.16	Programmed analysis operations for the phase vocoder. (After FLANAGAN and GOLDEN)	295
9.17	Speech transmitted by the phase vocoder. The transmission bandwidth is one-half the original signal bandwidth. Male speaker: "Should we chase those young outlaw cowboys." (After FLANAGAN and GOLDEN)	295
9.18	Phase vocoder time compression by a factor of 2. Male speaker	296
9.19	Phase vocoder time expansion by a factor of 2. Female speaker	297
9.20	Delta modulator with single integration	298
9.21	Waveforms for a delta modulator with single integration	299
9.22	Adaptive delta modulator with single integration	299
9.23	Waveform for an adaptive delta modulator with discrete control of the step size	300
9.24	Signal-to-noise ratios as a function of bit rate. Performance is shown for exponentially adaptive delta modulation (ADM) and logarithmic PCM. (After JAYANT)	301
9.25	Block diagram of linear prediction	303
9.26	Linear prediction receiver	303
9.27	Open-loop quantization of a predictor error signal	305
9.28	Predictive quantizing system. (After R. A. MCDONALD)	306
9.29	Two stage predictor for adaptive predictive coding. (After ATAL and SCHROEDER)	309
9.30	Adaptive predictive coding system. (After ATAL and SCHROEDER)	310
9.31	Analysis and synthesis operations for the homomorphic vocoder. (After OPPENHEIM)	311
9.32	Synthesis method for the maximum likelihood vocoder. Samples of voiced and voiceless excitation are supplied to a recursive digital filter of p -th order. Digital-to-analog (DIA) conversion produces the analog output. (After ITAKURA and SAITO, 1968)	313
9.33	Approximations to the speech spectrum envelope as a function of the number of poles of the recursive digital filter. The top curve, $S(f)$, is the measured short-time spectral density for the vowel / $p = 6, 8, 10, 12$	314
9.34	Automatic tracking of formant frequencies determined from the polynomial roots for $p = 10$. The utterance is the five-vowel sequence /a, o, i, u, e/. (After ITAKURA and SAITO, 1970)	316
9.35	Synthesis from a recursive digital filter employing optimum linear prediction. (After ATAL and HANAUER, 1971)	316
9.36	Formant frequencies determined from the recursive filter coefficients. The utterance is the voiced sentence "We were away a year ago" produced by a man at an average fundamental frequency of 120 cps. (After ATAL and HANAUER, 1971)	318
9.37	Phonetic pattern-matching vocoder. (After DUDLEY, 1958)	319
9.38	Parallel-connected formant vocoder. (After MUNSON and MONTGOMERY)	320
9.39	Cascade-connected formant vocoder. (After FLANAGAN and HOUSE)	321
9.40	Block diagram of voice-excited vocoder. (After DAVID, SCHROEDER, LOGAN and PRESTIGIACOMO)	323

9.41	Block diagram of the spectral flattener. (After DAVID, SCHROEDER, LOGAN and PRESTIGIACOMO)	323
9.42	Voice-excited formant vocoder. (After FLANAGAN, 1960b)	325
9.43	Block diagram of the Vobanc frequency division-multiplication system. (After BOGERT, 1956)	326
9.44	Block diagram of "harmonic compressor." (After SCHROEDER, LOGAN and PRESTIGIACOMO)	327
9.45	A "speech stretcher" using frequency multiplication to permit expansion of the time scale. (After GOULD)	327
9.46	A complete formant-vocoder system utilizing analog and digital transmission techniques. (After STEAD and JONES; STEAD and WESTON)	329
9.47	Schematic sound spectrogram illustrating the principle of the "one-man TASI." (After FLANAGAN, SCHROEDER and BIRD)	332
9.48	Block diagram of "one-man TASI" system for 2:1 band-width reduction. (After FLANAGAN, SCHROEDER and BIRD)	333
9.49	Sound spectrograms illustrating operation of the single channel speech interpolator .	333
9.50	Channel vocoder utilizing time-multiplex transmission. (After VILBIG and HAASE, 1956)	335

List of Tables

1.1	Relative frequencies of English speech sounds in standard prose. (After Dewey, 1923)	6
2.1	Vowels	19
2.2	All consonants may be divided into four broad manner classes, using the two binary features sonorant and continuant . The opposite of sonorant is obstruent ; the opposite of continuant is discontinuant	20
2.3	Fricative consonants	21
2.4	Stop consonants	22
2.5	Nasals	23
2.6	Glides and semi-vowels	24
4.1	Typical listing of control data for the computer-simulated synthesizer of Fig. 4.38 . .	128
4.2	Discrete control symbols for synthesis from printed text. (After COKER, UMEDA and BROWMAN)	129
6.1	Listener responses to synthetic and natural nasal consonants	192
9.1	Eighth-order Butterworth filter cutoff frequencies in cps	276
9.2	Impulse response durations for the Hilbert filters	276
9.3	Consonant intelligibility for a vocoder. Percent of initial consonants heard correctly in syllables (togatoms). (After HALSEY and SWAFFIELD)	291
9.4	Vocoder consonant intelligibility as a function of digital data rate. (After DAVID, 1956)	291
9.5	Quantization of formant-vocoder signals. (After STEAD and WESTON)	329
9.6	Estimated precision necessary in quantizing formant-vocoder parameters. The estimates are based upon just-discriminable changes in the parameters of synthetic vowels; amplitude parameters are considered to be logarithmic measures. (After FLANAGAN, 1957b)a	330

Chapter 1

Voice Communication

“Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals, the power of speech is intended to set forth the expedient and inexpedient, and therefore likewise the just and the unjust. And it is a characteristic of man that he alone has any sense of good and evil, of just and unjust, and the like, and the association of living beings who have this sense makes a family and a state.”

ARISTOTLE, *Politics*

Our primary method of communication is speech. Humans are unique in our ability to transmit information with his voice. Of the myriad varieties of life sharing our world, only humans have developed the vocal means for coding and conveying information beyond a rudimentary stage. It is more to our credit that we have developed the facility from apparatus designed to subserve other, more vital purposes.

Because humans evolved in an atmosphere, it is not unnatural that we should learn to communicate by causing air molecules to collide. In sustaining longitudinal vibrations, the atmosphere provides a medium. At the acoustic level, speech signals consist of rapid and significantly erratic fluctuations in air pressure. These sound pressures are generated and radiated by the vocal apparatus. At a different level of coding, the same speech information is contained in the neural signals which actuate the vocal muscles and manipulate the vocal tract. Speech sounds radiated into the air are detected by the ear and apprehended by the brain. The mechanical motions of the middle and inner ear, and the electrical pulses traversing the auditory nerve, may be thought of as still different codings of the speech information.

Acoustic transmission and reception of speech works fine, but only over very limited distances. The reasons are several. At the frequencies used by the vocal tract and ear, radiated acoustic energy spreads spatially and diminishes rapidly in intensity. Even if the source could produce great amounts of acoustic power, the medium can support only limited variations in pressure without distorting the signal. The sensitivity of the receiver—the ear—is limited by the acoustic noise of the environment and by the physiological noises of the body. The acoustic wave is not, therefore, a good means for distant transmission.

Through the ages men have striven to communicate at distances. They are, in fact, still striving. The ancient Greeks are known to have used intricate systems of signal fires which they placed on judiciously selected mountains for relaying messages between cities. One enterprising Greek, Aeneas Tacitus by name, is credited with a substantial improvement upon the discrete bonfire message. He placed water-filled earthen jars at the signal points. A rod, notched along its length and supported on a cork float, protruded from each jar. At the first signal light, water was started draining from the jar. At the second it was stopped. The notch on the rod at that level represented a previously agreed

upon message. (In terms of present day information theory, the system must have had an annoyingly low channel capacity, and an irritatingly high equivocation and vulnerability to jamming!)

History records other efforts to overcome the disadvantages of acoustic transmission. In the sixth century B.C., Cyrus the Great of Persia is supposed to have established lines of signal towers on high hilltops, radiating in several directions from his capital. On these vantage points he stationed leather-lunged men who shouted messages along, one to the other. Similar “voice towers” reportedly were used by Julius Caesar in Gaul. (Anyone who has played the party game of vocally transmitting a story from one person to another around a circle of guests cannot help but reflect upon the corruption which a message must have suffered in several miles of such transmission.)

Despite the desires and motivations to accomplish communication at distances, it was not until humans learned to generate, control and convey electrical current that telephony could be brought within the realm of possibility. As history goes, this has been exceedingly recent. Little more than a hundred years have passed since the first practical telephone was put into operation; there are now, by some accounts, more telephones than people on planet Earth.

Many early inventors and scientists labored on electrical telephones and laid foundations which facilitated the development of commercial telephony. Their biographies make interesting and humbling reading for today’s communication engineer comfortably ensconced in a well equipped laboratory.

Among the pioneers, Bell was somewhat unique for his background in physiology and phonetics. His comprehension of the mechanisms of speech and hearing was undoubtedly valuable, if not crucial, in his electrical experimentation. Similar understanding is equally important with today’s telephone researcher. It was perhaps his training that influenced Bell—according to his assistant Watson—to summarize the telephony problem by saying “If I could make a current of electricity vary in intensity precisely as the air varies in density during the production of a speech sound, I should be able to transmit speech telegraphically.” This is what he set out to do and is what he accomplished. Bell’s basic notion—namely, preservation of acoustic waveform—clearly proved to be an effective means for speech transmission. Waveform coding was the most widely used form of telephony until approximately the year 2000, when the number of digital cellular telephones began to outnumber the number of analog handsets. As we shall see, even digital telephony preserves the waveform, in the sense that only perceptually insignificant distortions are allowed.

Although the waveform principle is exceedingly satisfactory and has endured for almost a century, it is not the most efficient means for voice transmission. Communication engineers have recognized for many years that a substantial mismatch exists between the information capacity of human perception and the capacity of the “waveform” channel. Specifically, the channel is capable of transmitting information at rates much higher than those the human can assimilate.

Recent developments in communication theory have established techniques for quantifying the information in a signal and the rate at which information can be signalled over a given channel. These analytical tools have accentuated the desirability of matching the transmission channel to the information source. From their application, conventional telephony has become a much-used example of disparate source rate and channel capacity. This disparity—expressed in numbers—has provided much of the impetus toward investigating more efficient means for speech coding and for reducing the bandwidth and channel capacity used to transmit speech.

1.1 Speech as a Communication Channel

We speak to establish social bonds, and to create ideas larger than ourselves. The natural environment for speaking is noisy and complicated, with a continuously changing visual and auditory channel, as depicted, for example, in Fig. ???. In this famous painting, a group of friends relaxes on a Sunday afternoon at the restaurant *Maison Fournaise*. The image provides examples of many different kinds of conversations: flirtations, expositions, relaxed subdued conversations, and even a conversation between a woman (Aline Charigot, who would later marry Renoir) and her dog.



Figure 1.1: Conversation over lunch: Renoir's *Luncheon of the Boating Party*, 1881. (Phillips Collection, Washington D.C.)

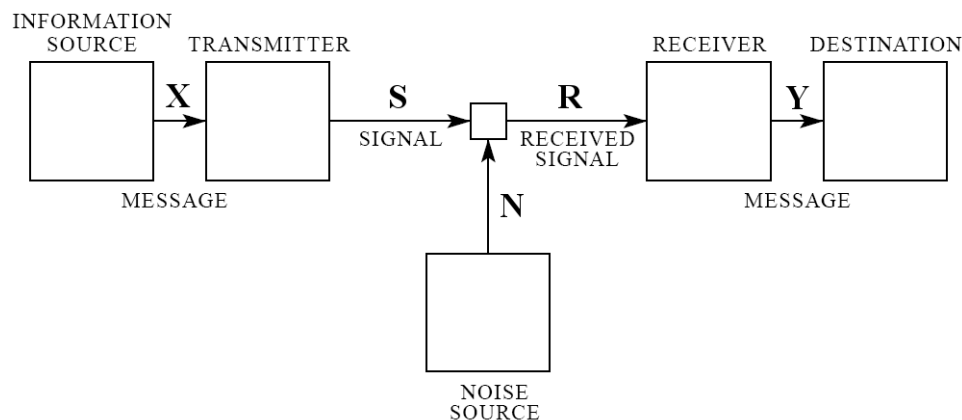


Figure 1.2: Schematic diagram of a general communication system. X =source message, Y =received message, S =transmitted signal, R =received signal, N =noise. (After Shannon and Weaver, 1949)

Before speaking, every talker conceives a message: a sequence of words, possibly annotated with subtle hints of nuance and opinion ($?$, $?$). The message is symbolic, and therefore digital: most of the content of a spoken message may be equivalently conveyed in an e-mail. In most cases, however, we find it pleasant to encode the message in an analog medium, by configuring the speech articulators (the lips, jaw, tongue, soft palate, larynx, and lungs) in order to generate an acoustic waveform. A listener measures the acoustic signal, and converts it into a neural code. The neural code passes through a series of neural circuits until, eventually, the listener has decoded the intended linguistic message—or something approximating the intended message.

The subject of this book is the encoding and decoding of the messages conveyed by speech: the digital-to-analog and analog-to-digital transformations used by humans and machines to produce and understand ordinary conversation. Before considering the analog channel in more detail, however, it's worthwhile to evaluate the end-to-end performance of the channel.

The mathematical theory of information ($?$, $?$) provides a useful mechanism for analyzing the end-to-end performance of any communications channel, independent of the details of its implementation. Fig. ?? shows the schematic of an abstract communication channel. There are six boxes in this figure. The boxes marked “information source” and “noise source” each draw a message or a noise signal, at random, from some probability distribution. The goal of the box marked “transmitter” is to encode the message, and that of the “receiver” is to decode the message, so that the received message will be as similar as possible to the transmitted message. As we shall see, the average information rate of the speech source is remarkably low. There are apparently two reasons for the low information rate of speech. First, there is evidence that human listeners are unable to process information at a rate much higher than that of the speech message; in this respect, humans are much less effective than machines. Second, low information rate allows speech transmission over extremely noisy acoustic channels. Human listeners (but not machines, yet) are able to correctly understand meaningful linguistic messages transmitted at signal to noise ratios (SNR) as low as -20dB; in this respect, humans are much more effective than machines. The low information rate of speech, and its remarkable noise robustness, are best understood as an adaptation to noisy natural environments like the outdoor lunch party in Fig. ??.

1.2 Entropy of the Speech Source

The elementary relations of information theory define the information associated with the selection of a discrete message from a specified ensemble. If the messages of the set are x_i , are independent, and have probability of occurrence $P(x_i)$, the information associated with a selection is $I = \log_2(1/P(x_i))$ bits¹. The average information associated with selections from the set is the ensemble average

$$H(X) = \sum_i P(x_i) \log_2 \left(\frac{1}{P(x_i)} \right) = - \sum_i P(x_i) \log_2 P(x_i)$$

bits, or the source entropy.

Consider, in these terms, a phonemic transcription of speech; that is, the written equivalent of the meaningfully distinctive sounds of speech. Take English for example. Table ?? shows a list of 42 English phonemes including vowels, diphthongs and consonants, and their relative frequencies of occurrence in prose (?). If the phonemes are selected for utterance with equal probability [i.e., $P(x_i) = 1/42$] the average information per phoneme would be approximately $H(X) = 5.4$ bits. If the phonemes are selected independently, but with probabilities equal to the relative frequencies shown in Table ??, then $H(X)$ falls to 4.9 bits. The sequential constraints imposed upon the selection of speech sounds by a given language reduce this average information still further². In conversational speech about 10 phonemes are uttered per second. The written equivalent of the information generated is therefore less than 50 bits/sec.

1.3 Conditional Entropy of Received Speech

Because of noise, the speech signal arriving at the receiver may be different from the signal generated by the transmitter. If the decoding algorithm is not sufficiently robust, noise in the acoustic signal may lead to errors in the received message. Perceptual errors can be characterized by the conditional probability that the receiver decodes symbol y_j , given that the transmitter encoded symbol x_i . This probability may be written as $P_{AB\gamma}(y_j|x_i)$, in order to emphasize that it is also a function of several channel characteristics, including the encoding system used by the transmitter and receiver (A), the bandwidth of the channel (B), and the SNR ($\gamma = S/N$, where S is the power of the signal coming out of the transmitter, and N is the power of the noise signal). For example, an error-free communication system is characterized by the conditional probability distribution

$$P_{AB\gamma}(y_j|x_i) = \delta_{ij} \equiv \begin{cases} 1 & y_j = x_i \\ 0 & \text{otherwise} \end{cases} \quad (1.1)$$

If $P_{AB\gamma}(y_j|x_i) \neq \delta_{ij}$, then one may say that the communication system is itself introducing “information” into the received signal. This is an undesirable behavior, because the “information” generated by the communication channel is independent of the information generated at the source; this extra “information” is usually called “error.” The average rate at which the communication

¹The base-2 logarithm is used to compute information in bits. A base-10 logarithm computes information in “digits;” a natural logarithm computes information in “nats.” All three units are commonly used in practice.

²Related data exist for the letters of printed English. Conditional constraints imposed by the language are likewise evident here. If the 26 English letters are considered equiprobable, the average information per letter is 4.7 bits. If the relative frequencies of the letters are used as estimates of $P(x_i)$, the average information per letter is 4.1 bits. If digram frequencies are considered, the information per letter, when the previous letter is known, is 3.6 bits. Taking account of trigram frequencies lowers this figure to 3.3 bits. By a limit-taking procedure, the long range statistical effects can be estimated. For sequences up to 100 letters in literary English the average information per letter is estimated to be on the order of one bit. This figure suggests a redundancy of about 75 per cent. If statistical effects extending over longer units such as paragraphs or chapters are considered, the redundancy may be still higher (?).

Table 1.1: Relative frequencies of English speech sounds in standard prose. (After Dewey, 1923)

Vowels and diphthongs			Consonants		
Pho- neme	relative frequency of occur- ence (%)	$-P(x_i) \log_2 P(x_i)$	Pho- neme	relative frequency of occur- ence (%)	$-P(x_i) \log_2 P(x_i)$
TIPAI	8.53	0.3029	TIPAn	7.24	0.2742
TIPAA	4.63	0.2052	TIPAt	7.13	0.2716
TIPAæ	3.95	0.1841	TIPAr	6.88	0.2657
TIPAE	3.44	0.1672	TIPAs	4.55	0.2028
TIPA5	2.81	0.1448	TIPAd	4.31	0.1955
TIPA2	2.33	0.1264	TIPAl	3.74	0.1773
TIPAi	2.12	0.1179	TIPAT	3.43	0.1669
TIPAE, TIPAEI	1.84	0.1061	TIPAZ	2.97	0.1507
TIPAU	1.60	0.0955	TIPAm	2.78	0.1437
TIPAAI	1.59	0.0950	TIPAk	2.71	0.1411
TIPAOU	1.30	0.0815	TIPAv	2.28	0.1244
TIPAO	1.26	0.795	TIPAw	2.08	0.1162
TIPAU	0.69	0.0495	TIPAp	2.04	0.1146
TIPAAU	0.59	0.0437	TIPAf	1.84	0.1061
TIPAA	0.49	0.0376	TIPAh	1.81	0.1048
TIPAO	0.33	0.0272	TIPAb	1.81	0.1048
TIPaju	0.31	0.0258	TIPAN	0.96	0.0644
TIPAOI	0.09	0.0091	TIPAS	0.82	0.0568
			TIPAg	0.74	0.0524
			TIPAj	0.60	0.0443
			TIPAtS	0.52	0.0395
			TIPAdZ	0.44	0.0344
			TIPAT	0.37	0.0299
			TIPAZ	0.05	0.0055
Totals	38			62	

$H(X) = -\sum_i P(x_i) \log_2 P(x_i) = 4.9$ bits. If all phonemes were equiprobable, then $H(X) = \log_2 42 = 5.4$ bits

channel introduces errors into a transmitted signal is called the *equivocation* or *conditional entropy* of Y given X , and is defined to be

$$\begin{aligned} H_{AB\gamma}(Y|X) &= - \sum_i \sum_j P_{AB\gamma}(x_i, y_j) \log_2 P_{AB\gamma}(y_j|x_i) \\ &= - \sum_i P(x_i) \sum_j P_{AB\gamma}(y_j|x_i) \log_2 P_{AB\gamma}(y_j|x_i) \end{aligned} \quad (1.2)$$

The amount of information successfully transmitted over the channel is equal to the information rate of the source, $H(X)$, minus the rate at which errors are introduced by the channel, $H_{AB\gamma}(Y|X)$. This rate is called the *mutual information* between the transmitted message and the received message:

$$\begin{aligned} I_{AB\gamma}(X, Y) &= H(X) - H_{AB\gamma}(Y|X) \\ &= \sum_i \sum_j P(x_i) P_{AB\gamma}(y_j|x_i) \left(\frac{P_{AB\gamma}(y_j|x_i)}{P(x_i)} \right) \end{aligned} \quad (1.3)$$

Human speech production is a coding algorithm, and may be evaluated just like any other coding algorithm: by computing the mutual information $I_{AB\gamma}$ that it achieves over any particular acoustic channel. Fletcher (?) found that, for SNRs of at least 30dB, phonemes in nonsense syllables are perceived correctly about 98.5% of the time, corresponding to an equivocation of roughly

$$H(Y|X) \approx 0.985 \log_2(1/0.985) + 0.015 \log_2(1/0.015) = 0.11 \text{ bits/symbol}^3. \quad (1.4)$$

In order to force listeners to make perceptual errors, Fletcher was forced to distort the acoustic channel by introducing additive noise and/or linear filtering (lowpass, highpass, or bandpass filters applied to the acoustic channel).

Eq. (??) is only an approximation of the speech channel equivocation: in order to calculate the equivocation exactly, it is necessary to know the probability $P_{AB\gamma}(y_j|x_i)$ for every (i, j) combination. Miller and Nicely (?) measured conditional probability tables under fifteen different channel conditions for a subset of the English language: specifically, for the subset $x_i \in \{p, b, t, d, k, g, f, v, s, z, m, n\}$, and y_j drawn from the same set. Each consonant was produced in a consonant vowel (CV) syllable, and the vowel was always /a/. In order to cause perceptual errors, Miller and Nicely limited the bandwidth of the acoustic channel (9 conditions), or the SNR (5 conditions). After several thousand trials, the perceptual effect of each channel was summarized in the form of a *confusion matrix*, like the one shown in Fig. ???. In a confusion matrix, entry $C(i, j)$ lists the number of times that phoneme x_i was perceived as phoneme y_j . The conditional probability $P(y_j|x_i)$ may be estimated as

$$P(y_j|x_i) \approx \frac{C(i, j)}{\sum_j C(i, j)} \quad (1.5)$$

Using the approximation in ??, the equivocation of the speech communication system, at -6 dB SNR, is 2.176 bits. Since each syllable is chosen uniformly from $2^4 = 16$ possible syllables, the source entropy is $H(X) = \log_2 16 = 4$ bits. The amount of information successfully transmitted from talker to listener, therefore, is $4 - 2.176 = 1.834$ bits. Fig. ??(a) shows the information transmitted from talker to listener, over the wideband acoustic channel, as a function of SNR. Mutual information is greater than one bit per consonant at -12dB, and the information rate only drops to zero below -18dB SNR. Fig. ??(b) shows the information transmitted over the lowpass-filtered and highpass filtered channels, as a function of the cutoff frequency.

³This approximation results from the assumption that only two events matter: the phoneme is either correctly or incorrectly recognized. The actual equivocation of a 42-phoneme communication system with a 1.5% error rate could be anywhere between 0.02 and 0.19 bits/symbol, depending on the error rates of each individual phoneme, and the distribution of errors across the various possible substitutions.

TABLE III. Confusion matrix for $S/N = -6$ db and frequency response of 200-6500 cps.

	<i>p</i>	<i>t</i>	<i>k</i>	<i>f</i>	<i>θ</i>	<i>s</i>	<i>ʃ</i>	<i>b</i>	<i>d</i>	<i>g</i>	<i>v</i>	<i>ʋ</i>	<i>z</i>	<i>ʒ</i>	<i>m</i>	<i>n</i>
<i>p</i>	80	43	64	17	14	6	2	1	1		1	1			2	
<i>t</i>	71	84	55	5	9	3	8	1				1	2		2	3
<i>k</i>	66	76	107	12	8	9	4					1			1	
<i>f</i>	18	12	9	175	48	11	1	7	2	1	2	2				
<i>θ</i>	19	17	16	104	64	32	7	5	4	5	6	4	5			
<i>s</i>	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
<i>ʃ</i>	1	6	3	4	6	29	195		3							1
<i>b</i>	1			5	4	4		136	10	9	47	16	6	1	5	4
<i>d</i>							8	5	80	45	11	20	20	26	1	
<i>g</i>					2			3	63	66	3	19	37	56		3
<i>v</i>				2		2		48	5	5	145	45	12		4	
<i>ʋ</i>					6			31	6	17	86	58	21	5	6	4
<i>z</i>					1	1	1	7	20	27	16	28	94	44		1
<i>ʒ</i>								1	26	18	3	8	45	129		2
<i>m</i>	1							4			4	1	3		177	46
<i>n</i>					4			1	5	2		7	1	6	47	163

Figure 1.3: Typical confusion matrix (6300Hz bandwidth, -6dB SNR). Entry (i, j) in the matrix lists the number of times that a talker said consonant x_i , and a listener heard consonant y_j . Each consonant was uttered as the first phoneme in a CV syllable; the vowel was always /a/. (After Miller and Nicely, 1955)

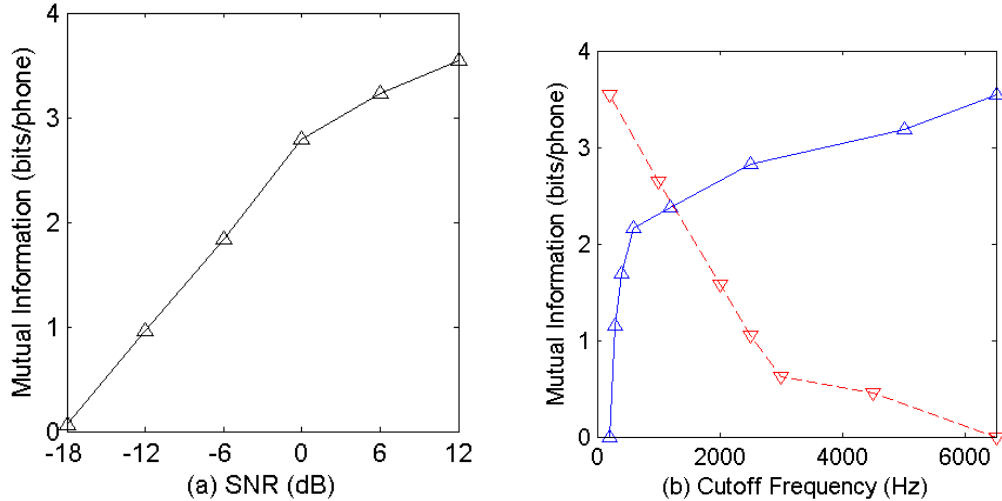


Figure 1.4: (a) Mutual information between spoken and perceived consonant labels, as a function of SNR, over an acoustic channel with 6300Hz bandwidth (200-6500Hz). (b) Mutual information between spoken and perceived consonant labels, at 12dB SNR, over lowpass and highpass acoustic channels with the specified cutoff frequencies. The lowpass channel contains information between 200Hz and the cutoff; bit rate is shown with a solid line. The highpass channel contains information between the cutoff and 6500Hz; bit rate is shown with a dashed line. (After Miller and Nicely, 1955)

1.4 Capacity of the Acoustic Channel

Mutual information is a summary of the efficiency with which algorithm A transmits information over a channel with bandwidth B and noise statistics N . Shannon has demonstrated (?) that no algorithm can transmit more information than

$$I(X, Y) \leq C\left(B, \frac{S}{N}\right), \quad (1.6)$$

where B is the bandwidth of the channel, S/N is the signal to noise ratio, and $C(B, S/N)$ is called the *channel capacity*. Shannon has shown that the channel capacity of a channel with additive Gaussian noise is given by

$$C(B, S/N) = \int_0^B \log_2 \left(1 + \frac{S(f)}{N(f)}\right) df \quad \frac{\text{bits}}{\text{second}} \quad (1.7)$$

where $S(f)$ and $N(f)$ are the power spectra of the speech and noise, respectively. Speech is transmitted over an acoustic channel with bandwidths varying between about 3000Hz (telephone transmission) to 20kHz (the audible frequency range, usable during face-to-face communication). Under very noisy listening conditions (e.g., at an SNR of -12dB or $S/N = 0.0625$), the capacity of a telephone-band acoustic channel is 188 bits/second—far greater than the information transmitted from a human talker to a human listener. In a quiet room (at an SNR of about 30dB, or $S/N \approx 1000$), the channel capacity of a 20kHz channel is 20,000 bits/second—400 times greater than the information rate achieved by a human conversationalist.

Why is speech limited to a rate of 50 bits/second? Phrased another way: why don't people talk more quickly under quiet listening conditions, or more clearly, in order to communicate at a bit rate higher than 50 bps? Is the extra information already present, in the form of subtle nuances of intonation? Is the time waveform simply an inefficient code, incapable of carrying more than 50bps? Is the human incapable of processing information at rates much higher than 50 bits/sec? Does the receiver discard much of the transmitted information? Chapter ?? will consider these questions in much greater detail; for now, let us consider some experimental studies that have tried to answer this question.

A number of experimental efforts have been made to assess the informational capacity of human listeners. The experiments necessarily concern specific, idealized perceptual tasks. In most cases it is difficult to generalize or to extrapolate the results to more complex and applied communication tasks. Even so, the results do provide quantitative indications which might reasonably be taken as order-of-magnitude estimates for human communication in general.

In one response task, for example, subjects were required to echo verbally, as fast as possible, stimuli presented visually (?, ?). The stimuli consisted of random sequences of binary digits, decimal digits, letters and words. The maximal rates achieved in this processing of information were on the order of 30 bits/sec. When the response mode was changed to manual pointing, the rate fell to about 15 bits/sec.

The same study considered the possibility for increasing the rate by using more than a single response mode, namely, by permitting manual and vocal responses. For this two-channel processing, the total rate was found to be approximately the sum of the rates for the individual response modes, namely about 45 bits/sec. In the experience of the authors this was a record figure for the unambiguous transmission of information through a human channel.

Another experiment required subjects to read lists of common monosyllables aloud (?, ?). Highest rates attained in these tests were 42 to 43 bits/sec. It was found that prose could be read faster than randomized lists of words. The limitation on the rate of reading was therefore concluded to be mental rather than muscular. When the task was changed to reading and tracking simultaneously, the rates decreased.

A different experiment measured the amount of information subjects could assimilate from audible tones coded in several stimulus dimensions (?). The coding was in terms of tone frequency, loudness, interruption rate, spatial direction of source, total duration of presentation and ratio of on-off time. In this task subjects were found capable of processing 5.3 bits per stimulus presentation. Because presentation times varied, with some as great as 17 sec, it is not possible to deduce rates from these data.

A later experiment attempted to determine the rate at which binaural auditory information could be processed (?). Listeners were required to make binary discriminations in several dimensions: specifically, vowel sound; sex of speaker; ear in which heard; and, rising or falling inflection. In this task, the best subject could receive correctly just under 6 bits/sec. Group performance was a little less than this figure.

As indicated earlier, these measures are determined according to particular tasks and criteria of performance. They consequently have significance only within the scopes of the experiments. Whether the figures are representative of the rates at which humans can perceive and apprehend speech can only be conjectured. Probably they are. None of the experiments show the human to be capable of processing information at rates greater than the order of 50 bits/sec.

Assuming this figure does in fact represent a rough upper limit to man's ability to ingest information, he might allot his capacity in various ways. For example, if a speaker were rapidly uttering random equiprobable phonemes, a listener might require all of his processing ability to receive correctly the written equivalent of the distinctive speech sounds. Little capacity might remain for perceiving other features of the speech such as stress, inflection, nasality, timing and other attributes of the particular voice. On the other hand, if the speech were idle social conversation, with far-reaching statistical constraints and high redundancy, the listener could direct more of his capacity to analyzing personal characteristics and articulatory peculiarities.

1.5 Organization of this Book

The goal of this book is to teach the science and technology of speech analysis, synthesis, and perception. The book is loosely divided into a "science" half and a "technology" half. The science and technology are unified by an information-theoretic view of speech communication, based on the theory and terminology developed by Shannon.

The first half of the book (chapters 1-5) addresses the science of speech communication. The science of speech, in our view, is the study of the speech behaviors of human beings, and includes a mathematically sophisticated treatment of ideas from both physics and psychology. Like all other communication channels, the speech communication channel is best studied by methodically elucidating the characteristics of the message, the transmitter, the receiver, and the channel. Chapter 2 describes the characteristics of the message: the alphabet of phonemes and suprasegmental speech gestures, and the probabilistic rules that govern their combination. Chapter 3 describes the speech transmitter, with a particular emphasis on the physical acoustic principles of speech production. Chapter 4 describes the speech receiver, including the results of both physiological and psychological experiments studying the transductive processes of the ear. Finally, chapter 5 describes characteristics of the channel and the receiver that relate to the perception and understanding of speech.

The second half of the book (chapters 6-9) describes technological methods that have been used to analyze, replace or augment each component of the speech communication system. Chapter 6 describes fundamental signal analysis methods that are common to the algorithms of all succeeding chapters. After a reader has finished understanding chapter 6, the rest of the book need not be read in order; each of chapters 7-9 may be studied independently as a self-contained introduction to the technology it describes. Chapter 7 describes algorithms that replace the speech transmitter by converting a text message into a natural-sounding acoustic speech signal. Chapter 8 describes

algorithms that replace the speech receiver, in the sense that they automatically convert an acoustic speech signal into a written sequence of phonemes or words. Finally, chapter 9 describes algorithms that replace the acoustic channel with a low-bit-rate digital channel, for purposes of secure, cellular, or internet telephony. All three of these areas are the subjects of active ongoing research; the goal of this book is to present fundamental concepts and derivations underlying the most effective solutions available today.

References

- Dewey, G. (1923). *Relative frequency of English speech sounds*. Cambridge, Massachusetts: Harvard University Press.
- Fletcher, H. (1922). The nature of speech and its interpretation. *Bell System Technical Journal*, 1, 129-144.
- Levelt, W. J. M. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.
- Licklider, J. C. R., Stevens, K. N., & Hayes, J. R. M. (1954). *Studies in speech, hearing and communication. final report, contract W-19122ac-1430*. Cambridge, Mass.
- Miller, G. A., & Nicely, P. E. (1955). Analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, 27, 338-352.
- Pierce, J. R., & Karlin, J. E. (1957). Information rate of a human channel. *Proc. I.R.E.*, 45, 368.
- Pollack, I., & Ficks, L. (1954). Information of elementary multidimensional auditory displays. *J. Acoust. Soc. Am.*, 26, 155-158.
- Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois.
- Webster, J. C. (1961). Information in simple multidimensional speech messages. *J. Acoust. Soc. Am.*, 33, 940-944.