## Index

abbreviations, decoding, 98 Abjab writing system, 34-5 acoustic models of speech production, 309-40 about the acoustic models, 309, 339-40 assumptions discussion, 336-40 components of the model, 309-11 models with vocal-tract losses, 335-6 nasal cavity modelling, 333-5 oral cavity sound source positions, 335 radiation models, 330 source and radiation effects, 336 see also glottis/glottal source; sound, physics of; vowel-tube models acoustic representations, 156-9 spectogram, 157-9 spectral analysis/frequency domain analysis, 156 spectral envelope, 156-7 acoustic-space formulation (ASF), 485, 493-7 acoustic theory see sound, physics of; vowel-tube model acoustic waves, 316-18 acronyms, decoding, 99 adapting systems if TTS, 50 addition paradigm, 71 Advanced Telecommunications Research (ATR), 512 - 14affective communication, 8-9 affective prosody, 17, 123-4 affricates, 153, 201 air-flow measurement (mouth air flow), 155 algorithms and features, 79-82 hand written algorithms, 80 see also text-classification algorithms allophones, 162 allophonic variation, 166-7 all-pole modelling, assumptions, 337-8 alphabetic writing, 34 alternative spellings, decoding, 98 alveolar, 154 ambiguity issues, 22 different words, same form, 54 homograph ambiguity, 22

AM model see autosegmental-metrical (AM) intonation model analogue signals, 262-78 aperiodic signals, 262 complex exponential sinusoid, 266-9 complex numbers, 268 conjugate symmetric complex amplitudes, 268-9 Euler's formula, 266-8 Fourier series/synthesis/analysis, 265-6, 269-70 frequency, 264-5 frequency domain, 270-5 frequency range, 274 fundamental frequency (F0), 148, 265 harmonic frequency, 265 periodic signals, 262-9, 305-7 phase shift, 264 quasi-periodic signals, 262 sinusoid signals, 263-9 time domain, 270 waveforms, 262 see also Fourier transform APL (Anderson, Pierrehumbert and Liberman) synthesis scheme, 246 applications, future of, 538 approximants, 154, 201 arbitraryness, 15 architectures for TTS, 71-5 addition paradigm, 71 associative arrays (maps), 72 atomic values, 72-3 autosegmental phonology of data structures, 73 Delta formulation/structure, 74 dictionaries, 72 finite partial functions, 72 heterogeneous relation graph (HRG) formalism, 72 - 5list/tree/ladder relations, 73 lookup tables, 72 overwrite paradigm, 71 utterance structure, 71 articulatory gestures, 406 articulatory phonetics see speech production/articulatory phonetics

articulatory phonology, 183, 406 articulatory physiology, 406 articulatory synthesis, 405-7 ASCII encoding, 70 aspiration, 168 assimilation effect, 167 associative arrays (maps), 72 assumed intent for prosody, 49-50 atomic values, 72-3 audio-visual speech synthesis, 527-9 about audiovisual synthesis, 406-7, 527-8 and speech control, 528-9 texture mapping, 528 visemes, 528 auditory scales, 351-2 augmentative prosody, 18, 125-6 autocorrelation function, for pitch detection, 381 autocorrelation source separation method, 360-1 automatic labelling, 521 autosegmental-metrical (AM) intonation model, 227, 237-9 analysis with, 248 APL synthesis scheme, 246 data-driven synthesis, 247-8 deterministic synthesis, 246-7 prediction of labels from text, 246 synthesis with, 245-8 and the ToBI scheme, 247, 248 autosegmental phonology, 73, 183 auxiliary generation for prosody, 49-50 bag-of-features approach, 84 Baum-Welch algorithm, 449 Bayes' rule, 86-7, 545 beam pruning, 509 bilabial constriction, 154 Blizzard Challenge testing, 526 boundary accents/tones, 121, 236, 238 braille, 27 break index concept, 115 British English MRPA phoneme inventory, 554 British intonation school, 227, 236-7 Campbell timing model, 258 canned speech, 43-4 cepstra linear-prediction cepstra, 369 mel-frequency cepstral coefficient (MFCC), 370 cepstral coefficients, synthesis from, 429-31 cepstrum speech analysis, 353-7

cepstrum coentents, synthesis noin, 429–31 cepstrum speech analysis, 353–7 as deconvolution, 355–6 definition, 353 discussion, 356–7 the magnitude spectrum as a signal, 353–5 for pitch detection, 379 chain rule, 546 channel/medium (means of conversion), 13 character character-to-phoneme conversion, 55 definition, 54 encoding schemes, 69-70 CHATR system, 513 Cholskey decomposition technique, 359 Chomskian field, 534 classical linear-prediction (LP) synthesis, 399-405 about LP synthesis, 399 a complete synthesiser, 403-4 formant synthesis comparison, 399-400 impulse/noise source model, 400-1 LP diphone-concatenative synthesis, 401-3 source modelling, 378 source problems, 404-5 classification see text-classification algorithms classifiers, F0 models, 228 clitics, 60-1 closed-phase analysis, 374-7 instants of glottal closure points, 374 pre-emphasis, 375 cluster impurity, 88 coarticulation, 168 Cocke-Younger-Kasami (CYK) algorithm, 104 collocation rule, 84 colouring effect, 167 common-form model of TTS, 5-6, 38 communication processes, 18-23 ambiguity issues, 22 common ground issues, 20 dialogue turns, 18 effectiveness factor, 19-20 efficiency factor, 19-20 encoding/decoding, 18-19, 21-2 Grice's maxims, 20 homograph ambiguity, 22 information-theoretic approach, 23 message generation, 18-21 messages, 18 semiotics, 23 speech, redundancy in, 21 text decoding/analysis, 22 understanding, 19, 22-3 communication, types of, 8-13 about communication, 8, 23-5 affective communication, 8-9 iconic communication, 9-10 interpreted communication, 8 meaning/form/signal, 12-13 signals, 13 symbolic communication, 10-12 see also human communication comparison tests, 524 competitive evaluations, 526 complex numbers, 268

Index

585

component/unit testing, 525-6 compound-noun phrases, 116-17 compound prosodic domains theory, 114 comprehension tests, 523 compression, lossless and lossy, 215 computational phonology, 184 concatenative synthesis, 401-3 issues, 431-2 macro-concatenation, 431, 497 micro-concatenation, 431 optimal coupling, 432 phase mismatch issues, 431-2 concept-to-speech systems, 42-3 future of, 537 conditional probability, 545 conjugate symmetric complex amplitudes, 268-9 consonants, 153-5 affricates, 153 alveolar, 154 approximants, 154 bilabial, 154 difficult consonants, 199-200 fricatives, 153 glides, 154-5 IPA charts, 555 labiodental, 153 nasal stops, 153 obstruent, 154 oral stops, 153 context-free grammars (CFGs), 102-4 context-orientated-clustering, 495 context-sensitive modelling, 451-4 context-sensitive rewrite rule, 83-4 context-sensitive rules, 182 context-sensitive synthesis models, 461-3 continuants, 150 contractions, 60 convolution sum, 292 correlation coefficient 544 covariance matrix, 439 covariance method, 358-60 coverage (in unit-selection synthesis), 510 cumulative density functions, 549-10 curse of dimensionality, 81, 534 databases, 517-22

automatic labelling, 521 avoiding explicit labels, 521–2 hand labelling, 519–21 and labelling, 519–21 prosody databases, 518–19 text materials, 518 unit-selection databases, 517–18 data-driven intonation models, 250–4 about data-driven models, 250–1 dynamic-system models, 252–3

functional models, 254 HMM models, 253-4 SFC model, 254 unit-selection synthesis, 251-2 data-driven synthesis, 247-8, 435, 470-1 see also hidden Markov model (HMM) data sparsity problem, 81, 193 decision lists, 85-6 decision trees, 87-8, 221, 452-5 clustering, 494-6 decoding/encoding messages, 18-19, 21 - 2text decoding/analysis, 22 see also text decoding/analysis delta formulation/structure, 72 delta/velocity coefficients, 438-9 dental constriction, 154 dependency phonology, 183 deterministic acoustic models, synthesis with, 248-50 Fujisaki superimpositional models, 249 Tilt model, 249-50 deterministic phrasing prediction, 130-1 deterministic content function (DCF), 130 deterministic content function punctuation (DCFP), 131 deterministic punctuation (DP), 130 verb-balancing rule, 131 deterministic synthesis models, 246-7 dialogue turns, 18 dictionaries, 72 digital filters, 288-94, 308 about digital filters, 288-9 convolution sum, 292, 293-4 difference equations, 289 FIR filter, 289 IIR filter, 289 impulse response, 289-91 linearity principle, 289 linear time-invariant (LTI) filter, 288 recursive filters, 289 scaling, 288-9 superposition, 289 third-order filters, 289 transfer function, 293-4 and the z-transform, 293-4 digital filters, analysis/design, 294-305 about digital filter design, 304-5 anti-resonances, 303 characteristics, 298-304 complex-conjugate pairs of poles, 300-2 polynomial analysis (poles and zeros), 294-7 practical properties, 304-6 resonance/resonators, 300 skirts of poles, 302 and the z-domain transfer function, 297-8

	586	Index				
--	-----	-------	--	--	--	--

digital signals, 278-84, 307 digital representations, 280 digital waveforms, 279 discrete Fourier transform (DFT), 281-2 discrete-time Fourier transform (DTFT), 280-1 and the frequency domain, 283-4 Laplace transform, 283 Nyquist frequency, 279 sample rate/frequency, 279 z-transform, 282-3 diphone-concatenative synthesis, 401-3 diphone inventories, 414 diphones from speech, 414-15 diphone unit-selection system, 505 diphthongs, 153, 201 discourse-neutral renderings, 116 discrete Fourier transform (DFT), 281-2 discreteness, 15-16 discrete random variables, 540-1 discrete-time Fourier transform (DTFT), 280-1 discrete-tube model, assumptions, 337 distinctiveness of speech issues, 171-2 downdrift/declination, 230-3 duality principle, 15 duration synthesis modelling, 463-4 Dutch intonation school, 237 dynamic-system synthesis models, 252-3 dynamic-time-warping (DTW) technique, 219, 469 ease of data acquisition, and synthesis with vocal-tract models, 407 egressive pulmonic air stream, 147 eigenface model, 528 electoglottography/laryngography, 155, 383 electromagnetic articulography (EMA), 156 electropalatography, 155 emotional speech synthesis, 529-31 describing emotion, 529 with HMM techniques, 531 with prosody control, 529-30 with unit selection, 531 with voice transformation, 530 emotion axes, 123 emphasis, 118-19 encoding/decoding messages, 18-19, 21-2 engineering approach to TTS, 4 engine/rule separation, 83 entropy, 546-7, 552 epoch detection, 381-4 electroglottograph, 383

electroglottograph, 383 epoch-detection algorithm (EDA), 381 instant of glottal closure (IGC), 382–3 laryngograph/laryngograph signals (Lx signals), 383–4 pitch-synchronous analysis, 381 epoch manipulation for TD-PSOLA, 417-20 equivalent rectangular bandwidth (ERB) auditory scale, 352 Euclidean distance, 486 Euler's formula, 266-8 evaluation, 522-6 about evaluation, 522-3 see also tests/testing exceptions dictionaries, 208 expressive speech see emotional speech synthesis feature geometry, 183 features and algorithms, 79-82 filter-bank speech analysis, 352-3 filters see digital filters finite-impulse-response (FIR) filter, 289 finite partial functions, 72 first-generation synthesis see vocal-tract models, synthesis with forced alignment, 468 formants (speech resonance), 159-60 formant synthesis, 388-99 about formant synthesis, 388-9 consonant synthesising, 392-4 copy synthesis technique, 394-6 Klatt synthesiser, 394-5 lumped-parameter speech generation model, 389 parallel synthesisers, 392 phonetic input, 394-7 quality issues, 397-9 serial/cascade synthesisers, 391-2 single formant synthesis, 390-1 sound sources, 389-90 formant tracking, 370-2 form/message-to-speech synthesis, 42 Fourier series/synthesis/analysis, 265-6, 269-71 Fourier transform, 275-8 discrete Fourier transform (DFT), 281-2 discrete-time Fourier transform (DTFT), 280-1 duality principle, 278 inverse Fourier transform, 277-8 scaling property, 277 sinc function, 277 frame shift in speech analysis, 346-7 frequency, 264 angular frequency, 265 frequency domain, 270-5, 307 analysis/spectral analysis, 156 for digital signals, 283-4 for pitch detection, 381 fricatives, 153 Fujisaki intonation model, 227, 239-42 Fujisaki superimpositional models analysis with, 250 synthesis with, 249

Index

587

fundamental frequency (F0), 148, 265 and pitch, 225 see also pitch detection/tracking fundamental frequency (F0) contour models, 227-9 acoustic model, 228 classifiers, 228 regression algorithms, 228 target points, 228 Gaussian mixture models, 469 Gaussian/normal distribution/bell curve, 436-8, 549 general partial-synthesis functions, 496-7 generative models, 89-90 glides, 154-5 off-glides, 155 on-glides, 155 glottis/glottal source, 148, 330-3 assumptions, 338-9 glottal-flow derivative, 333 Lijencrants-Fant model, 332 open/return/closed phases, 330-1 parameterisation of glottal-flow signals, 379 government phonology, 183 graphemes, 28 definition, 54-5 TTS models, 39 grapheme-to-phoneme (G2P) conversion, 55, 218-22 with decision trees, 221 dynamic time warping (DTW), 219 G2P algorithms, 208, 218 G2P alignment, 219 memory-based learning, 220-1 NetTalk algorithm, 219–20 neural networks, 219-20 pronunciation by analogy, 220-1 rule-based techniques, 218-19 rule ordering, 219 statistical techniques, 221-2 with support-vector machines, 221 Grice's maxims, 20 hand labelling, 519-21 hand written algorithms, 80 harmonic/noise models (HNMs), 426-9 harmonics, 148-9 Harvard sentences, 523 Haskins sentences, 523 heterogeneous relation graph (HRG) formalism, 72 - 5hidden Markov model (HMM) about the HMM, 89-91, 435, 471-3 and intonation synthesis, 253-4 and phrasing prediction, 133-5 hidden Markov model (HMM) formalism, 435-56 about HMM formalism, 435-6

acoustic representations, 439-40 backoff techniques, 444 Baum-Welch algorithm, 449 context-sensitive modelling, 451-4 covariance matrix, 439 decision trees, 452-5 delta delta/ acceleration coefficients, 439 delta/velocity coefficients, 438-9 diagonal covariance, 439 discrete state problems, 454-5 forced-alignment mode, 448 forward-backward algorithm, 449-50 as generative models, 440-3 generative nature issues, 455-6 independence of observations issues, 454 language models, 444 linearity problems, 455 recognising with HMMs, 440-3 self-transition probability, 440 smoothing techniques, 444 states of phone models, 440 training HMMs, 448-51 transition probabilities, 440 triphone models, 451 Viterbi algorithm, 444-8 see also observations for HMMs hidden Markov models (HMMs), labelling databases with, 465-8 about labelling, 465 alignments quality measurement, 470 dynamic-time-warping (DTW) technique, 469 forced alignment, 468 Gaussian mixture models, 469 phone boundaries determination, 468-70 phone sequence determination, 467-8 word sequence determination, 467 hidden Markov models (HMMs), synthesis from, 456-64, 514 about synthesis from HMMs, 456-7 acoustic representations, 460-1 context-sensitive models, 461-3 duration modelling, 463-4 example systems, 464 likeliest observations for a given state sequence, 457-60 hidden semi-Markov model (HSMM), 464 homographs, 56 abbreviation homographs, 54 accidental homographs, 54 ambiguity issues, 22, 46 decoding, 98 disambiguation, 79, 99-101 homograph disambiguation, 56 part-of-speech homographs, 54 resolution of, 53 true homographs, 54

588	Index	
	homonyms, 58	phonological versus phonetic versus acoustic,
	pure homonyms, 58	244–5
	homophones, 56–7	purpose, 244
	human communication, 13–18	superimpositional models, 242
	about human communication, 13–14	superimpositional versus linear, 245
	affective prosody, 17	Tilt model, 227, 242–4
	augmentative prosody, 18	ToBI scheme, 237
	see also linguistic levels; verbal communication	tones versus shapes, 245
	Hunt and Black algorithm, 477–9, 504	traditional model, 236–7
	iconic communication 0, 10	see also autosegmental-metrical (AM) intonation
	iconic communication, 9–10	deterministic acoustic models, surthesis with
	alassiaal L P prediction 278	interaction and tune 121 2
	classical LP prediction, 578	prediction issues 130
	independence concent 543	INTSINT intenstion model 230
	independent feature formulation (IFF) 485	inverse filtering 372
	infinite_impulse_response (IIR) filter 289	IPA see International Phonetic Association (IPA)
	information_theoretic approach 23	ISO 8859 70
	inside-outside algorithms 105	150 8859, 70
	instant of glottal closure (IGC) points 374 382-3	java speech markup language 69
	integrated systems future of 536	join functions 497–504
	intelligibility issues 3 48–9 510 523	about joining units 497–8
	International Phonetic Association (IPA)	acoustic-distance join costs, 499–500
	alphabet, 163–5	categorical and acoustic join costs, 500–1
	consonant chart, 555	ioin classifiers, 497, 502–4
	symbol set (IPA alphabet), 163–5	join costs, 497–8
	interpreted communication, 8	join detectability, 498
	interpreting characters, 69-71	join probability, 497
	intonational phonology, 121	macro-concatenation issue, 497
	intonational phrases, 114	phone-class join costs, 498-9
	intonation behaviour, 229-36	probabilistic and sequence join function, 501-2
	boundary tones, 236	sequence join classifier, 503
	downdrift/declination, 230-3	singular-value decomposition (SVD), 502
	nuclear accents, 230	splicing costs, 499
	pitch accents, 230, 234–6	
	pitch range, 233–4	Kalman filter, 252
	tune, 229–30	Klatt deterministic rules, 256–7
	intonation synthesis, 225–9	Klatt synthesiser, 394–5
	about intonation, 225, 259-61	Kullback–Leibler distance, 552
	F0 and pitch, 226	
	F0 synthesis, 229	labelling databases, 519
	intonational form, 226–7	automatic labelling, 521
	intonational synthesis, 225	avoiding explicit labels, 521–2
	micro-prosody, 229	hand labelling, 519–21
	pitch-accent languages, 227	see also hidden Markov models (HMMs),
	tone languages, 227	labelling databases with
	intonation theories and models, 236–45, 250–4	labiodental constriction, 153
	about data-driven models, 250–1	language models, 444
	autosegmental-metrical (AM) model, 237–9	<i>iv</i> -gram language model, 444
	British school, $22/$ , $236-/$	Lonloss transform, 282
	uata uriven models, 230–4	Laplace transform, $283$
	F0 contour models 227 0	aryngograph/naryngograph signais (Lx signals),
	FU contour model, 227–9 Engineeri model, 227–220, 42–250	203-4 lowwy 149
	rujisaki mouel, 227, 239–42, 230	I autrante system 513
	INTSINT model 230	Laureau System, 313
	11N I 511N I 1110UCI, 239	least moundation principle, 477

**Index** 589

letter sequences, decoding, 99 Levinson-Durbin recursion source-filter separation technique, 361-2, 367 lexemes, inflected forms, 59 lexical phonology/word formation, 179-81 lexical stress, 116, 186-9 lexicons, 63, 207-18 compression, lossless and lossy, 215 computer lexicons, 207 exceptions dictionaries, 208 formats, 210-12 grapheme-to-phoneme algorithms, 208 language lexicons, 207 memorising the data, 209 offline lexicon, 213-14 orthographic and pronunciation variants, 210 - 12orthography-pronunciation lexicons, 207 over-fitting data, 209 quality of, 215-16 as a relational database, 210-11 rules for, 208-10 simple dictionary formats, 210 speaker's lexicons, 207 system lexicon, 214-15 unknown word problems, 216-18 Lijencrants-Fant model for glottal flow, 332, 374, 376 limited-domain synthesis systems, 44 linear filters, assumptions concerning, 337 linear-prediction cepstra, 369 linear-prediction (LP) PSOLA, 423-4 linear-prediction (LP) speech analysis, 357-65 about linear prediction, 357-8 autocorrelation method, 360-1 Cholskey decomposition method, 359 covariance method for finding coefficients, 358-60 Levinson-Durbin recursion technique, 361-2 perceptual linear prediction, 370 spectra for, 362-5 Toeplitz matrix, 361 linear-prediction (LP) synthesis see classical linear-prediction (LP) synthesis; residual-excited linear prediction linear time-invariant (LTI) filters, 288, 310 for nasalised vowels, 333-4 line-spectrum frequencies (LSFs), 367-9 linguistic-analysis TTS models, 39 linguistic levels, 16-17 morphemes/morphology, 16 phonetics/phonology, 16 pragmatics, 17 semantics, 16-17 speech acoustics, 16 syntax, 16

linguistics/speech technology relationship, 533-6 future of, 537-8 log area ratios, 367 logographic writing, 34 logotomes/nonsense words, 415 log power spectrum, 343-4 lookup tables, 72 lossless tube, assumptions, 338 LP see linear-prediction (LP) ... lumped-parameter speech generation model, 389 machine-readable phonetic alphabet (MRPA) phoneme inventory, 204-5 machine translation, 1 macro-concatenation, 497 magnetic resonance imaging (MRI), 156 Manhattan distance, 486 marginal distributions, 543 markup languages, 68-9 java speech markup language, 69 speech synthesis markup language (SSML), 69 spoken text markup language, 69 VoiceXML, 69 maximal onset principle, 185 MBROLA technique, 429 meaning/form/signal, and communication, 12-13 meaning-to-speech system, 42-3 mean opinion score, 524 medium (means of conversion), 13 mel-frequency cepstral coefficients (MFCCs), 370, 429-31, 439 mel-scale, 351 memorising the data (machine learning), 209 memory-based learning, 220-1 message/form-to-speech synthesis, 42 messages, 18 message generation, 20-1 metrical phonology, 114, 120, 183 metrical stress, 188 micro-prosody, 229 minimal pair principle/analysis, 163, 197-9, 204 mis-spellings, decoding, 98 model effectiveness, and synthesis with vocal-tract models 407 models of TTS, 37-41 common-form model, 38 comparisons, 40-1 complete prosody generation, 40 full linguistic-analysis, 39-40 grapheme form, 39 phoneme form, 39 pipelined, 39 prosody from the text, 40 signal-to-signal, 39 text-as-language, 39 modified rhyme test (MRT), 523, 524

590	Index				

modified timit ascii character set, 166 modularity, and synthesis with vocal-tract models, 407 moments of a PMF, 542 monophthongs, 153 morphemes/morphology, 16 morphology, 222-3 derivational, 59, 222 inflectional, 222 morphological decomposition, 222-3 and scope, 59 MRPA phoneme inventory, 204-5 multi-band-excitation (MBE), 427 multi-centroid analysis, 371 multi-pass searching, 509 naive Bayes' classifier, 86-7 names, pronunciation, 223 nasal cavity modelling, 333-5 nasalisation colouring, 167 nasal and oral sounds, 150 nasal stops, 153 natural-language parsing, 102-5 Cocke-Younger-Kasami (CYK) algorithm, 104 context-free grammars (CFGs), 102-4 probabilistic parsers, 104-5 statistical parsing, 105 natural-language text decoding, 46-7, 97-101 about natural-language text, 97-8 acronyms, 99 homograph disambiguation, 99-101 letter sequences, 99 non-homographs, 101 naturalness issues/tests, 3, 47-8, 510, 523, 524 mean opinion score, 524 natural phonology, 183 NetTalk algorithm, 219-20 neural networks, and G2P algorithms, 219-20 neutral vowel sound, 152-3 NextGen system (AT&T), 513-14 n-gram model, 91 non-linear phonology, 183 non-linguistic issues, 32-3 non-natural-language text decoding, 92-7 about non-natural-language text, 92 parsing, 95 semiotic classification, 92-4 semiotic decoding, 95 verbalisation, 95-7 nonsense words/logotomes, 415 non-standard words (NSWs), 106 non-uniform unit synthesis, 480 nuclear accents, 230 null/neutral prosody, 18 number-communication systems, 33 Nyquist frequency, 279

observations for HMMs, 436-8 covariance matrix, 437 Gaussian/normal distribution/bell curve, 436-8 multivariate Gaussian, 437 probabilistic models, 436 probability density functions (pdfs), 436 standard deviation, 436 variance, 436 obstruent consonants, 154 offline lexicon, 213-14 open-phase analysis, 377-8 optimal coupling, 432 optimality theory, 183 oral cavity, 152 sound source positions, 335 oral and nasal sounds, 150 oral stops, 153 over-fitting data, 209 overwrite paradigm, 71 palatalisation, 180 parameterisation of glottal-flow signals, 379 parsing/parsers, 53, 95, 103 probabilistic parsers, 104-5 statistical parsing, 105 see also natural-language parsing partial-synthesis function, 493 part-of-speech (POS) tagging, 82, 88-92 generative models, 89-90 hidden Markov model (HMM), 89-91 n-gram model, 91 observation probabilities, 90 POS homographs, 88 syntactic homonyms, 88-9 transition probabilities, 90 Viterbi algorithm, 92 perceptual linear prediction, 370 perceptual substitutability principle, 485 periodic signals, 262-9, 305-7 phase mismatch issues, 431-2 phase shift, 264 phase-splicing systems, 44 phone-class join costs, 498-9 phoneme inventories, 204-5 British English MRPA, 554 modified TIMIT for General American, 553 phonemes and graphemes, 28 and verbal communication, 14, 16, 161-4 phoneme TTS models, 39 phones about phones, 162-4 definitions, 553-5 phonetic similarity principle, 197-9

Index

591

phonetics/phonology, 16 phonetic context, 167 phonetic variants, 57 see also phonological theories; phonology; phonotactics; speech production/ articulatory phonetics phonological theories, 181-4 articulatory phonology, 183 autosegmental phonology, 183 computational phonology, 184 context sensitive rules, 182 dependency phonology, 183 feature geometry, 183 government phonology, 183 metrical phonology, 183 natural phonology, 183 non-linear phonology, 183 optimality theory, 183 The Sound Pattern of English (SPE), 181-2 phonology, 172-89 about phonology, 172, 189-91 lexical stress, 186-9 maximal onset principle, 185 metrical phonology, 114 palatalisation, 180 phonological phrases, 114 syllabic consonants, 184 syllables, 184-6 word formation/lexical phonology, 179-81 see also phonological theories; phonotactics phonotactics, 172-9 distinctive features, 174 feature structure, 174-9 phonotactic grammar, 172-7, 207 primitives issues, 176 syllable structures, 176-7 phrasing prediction, 129-36 classifier approaches, 132-3 deterministic approaches, 130-1 experimental formulation, 129-30 HMM approaches, 133-5 hybrid approaches, 135-6 precision and recall scheme, 130 phrasing/prosodic phasing, 112-15 about phrasing, 112-13 phasing models, 113-15 pictographic writing, 34 pipelined TTS models, 39 pitch-accent languages, 121, 227 pitch accents, 230, 234-6 alignment factors, 235-6, 534-5 height factors, 235 pitch detection/tracking, 379-81 pitch-detection algorithms (PDAs), 379 pitch-marking, 381 pitch range, 233-4

pitch-synchronous overlap and add (PSOLA) techniques, 415-21 about PSOLA, 415-16, 421 epoch manipulation, 417-20 time-domain PSOLA (TD-PSOLA), 416-17 pitch-synchronous speech analysis, 347, 381 polynomial analysis (poles and zeros), 294-7 post-lexical processing, 223-4 pragmatics, 17 pre-processing, 52 pre-recorded prompt systems, 43 probabilistic models, 436 probabilistic parsers, 104-5 probabilistic and sequence join function, 501-2 probability density functions (pdfs), 436 probability mass functions (PMFs), 541 probability theory, continuous random variables, 547-50 cumulative density functions, 549-50 expected values, 548-9 Gaussian (normal) distribution, 549 uniform distribution, 549 probability theory, discrete probabilities, 540-42 discrete random variables, 540-1 expected values, 541-2 moments of a PMF, 542 probability mass functions (PMFs), 541 probability theory, pairs of continuous random variables, 550-2 entropy for, 552 independent versus uncorrelated, 551 Kullback-Leibler distance, 552 sum of two, 551-2 probability theory, pairs of discrete random variables, 542-7 Baye's rule, 545 chain rule, 546 conditional probability, 545 correlation, 544 entropy, 546-7 expected values, 543 higher-order moments and covariance, 544 independence, 543 marginal distributions, 543 moments of a joint distribution, 544 sum of random variables, 545-6 problems in text-to-speech, 44-50 adapting systems, 50 assumed intent for prosody, 49-50 auxiliary generation for prosody, 49-50 homograph ambiguity, 46 intelligibility issues, 48-9 natural language text decoding, 46-7 naturalness, 47-8 syntactic ambiguity, 46-7 text classification/semiotic systems, 44-6

prosody
about prosody 14
affective. 17
augmentative 18
null/neutral, 18
in reading aloud, 36–7
prosody, determination from text, 127–9
augmentative prosody control. 128
prosody and human reading, 127–9
prosody and synthesis techniques, 128–9
TTS models, 40
prosody, prediction from text, 53, 111–45
about prosody, 111–12, 144–5
affective prosody, 123–4
augmentative prosody, 125–6, 142
intonational-tune prediction, 139
intonation and tune, 121–2
labelling schemes/accuracy, 139–41
linguistic theories/prosody, 141–2
phrasing, 112–15
prosodic meaning and function, 122–7
prosodic phase structures, 113
prosodic style, 127
real dialogues, 143–4
speaker choice/variability. 142–3
suprasegmentality, 124
symbolic communication 126–7
underspecified text 142
see also phrasing prediction: prominence:
prominence prediction: prosody
determination
pruning methods, 508–9
beam pruning, 509
PSOLA see pitch-synchronous overlap and add
(PSOLA) techniques
punctuation
status markers, 65
and tokenisation. 65–6
underlying punctuation, 66
pure unit selection, 477
r ····································
quality improvements, future of, 537
1 J I I I I I I I I I I I I I I I I I I
radiation models for sound, 330
assumptions, 338
Reading aloud, 35–7
prosody in, 36–7
and silent reading, 35–6
style issues, 37
verbal content, 37
RealSpeak system, 514
reduced stress, 187
redundancy, in speech. 21
reflection coefficients. 366–7
regression algorithms. F0 contour models. 228
resequencing algorithms, 477

**Index** 593

residual analysis, for pitch detection, 381 residual-excited linear prediction, 421-4 about residual excited LP, 421-3 linear-prediction PSOLA, 423-4 residual manipulation, 423 residual manipulation, 423 residual speech signals, 372-4 error signals, 372 inverse filtering, 372 resonance, 159 formants, 159-60 resonant systems, 311-13 rhotic/non-rhotic accents, 202 rVoice system, 514 scope and morphology, 59 secondary stress, 187 second-generation synthesis systems, 412-34 about second-generation systems, 412-13, 433-4 cepstral coefficients, synthesis from, 429-31 concatenation issues, 431-2 diphone inventory creation, 414 diphones from speech, 414-15 MBROLA technique, 429 speech units in, 413-15 see also pitch-synchronous overlap and add (PSOLA) techniques; residual-excited linear prediction; sinusoidal models techniques segments, 162 semantics, 16-17 semiotic classification, 45, 79, 92-4 open-class rules, 93 specialist sub-classifiers, 93 and translation, 45-6 semiotic decoding, 95 semiotics, 23 semiotic systems, 33-4 sentences, 14, 62-3 sentence-final prosody, 67 sentence splitting, 53, 67-8 style manuals, 68 sentential stress, 186-7 sequence join classifier, 503 signal processing and unit-selection, 511 signals and communication, 13 see also analogue signals; digital signals; transforms signal-to-signal TTS models, 39 sinc function, 277 singular-value decomposition (SVD), 502 sinusoidal models techniques, 424-9 about sinusoidal models, 424-5 harmonic/noise models (HNMs), 426-9 multi-band-excitation (MBE), 427 pure sinusoidal models, 425-6

sinusoid signals, 263-5 Sound Pattern of English, The (SPE), 181-2 sound, physics of, 311-19 acoustic capacitance, 317 acoustic impedance, 317 acoustic inductance, 317 acoustic reflection, 318 acoustic resistance, 317 acoustic waves, 316-18 boundary conditions, 315 lossless tubes, 317 resonant systems, 311-13 sound propagation, 317 speed of sound, 316 standing waves, 314 travelling waves, 313-15 see also vowel-tube model sound sources see speech production/articulatory phonetics source/filter model of speech, 151 source-filter separation see cepstrum speech analysis; filter-bank speech analysis; linear-prediction (LP) speech analysis source-filter separation assumptions, 338 source signal representations, 372-9 closed-phase analysis, 374-7 impulse/noise models, 378 open-phase analysis, 377-8 parameterisation of glotta-flow signals, 379 residual signals, 372-4 speaker choice/variability, 142-3 spectral analysis/frequency domain analysis, 156 spectral-envelope, 156-7 and vocal-tract representations, 362-72 spectral representations of speech, short term, 343-5 envelopes, 345 spectrograms, 157-9 in speech analysis, 348-51 speech, disfluences in, 30 speech acoustics, 16 speech, communicative use principles, 160-72 about communicating with speech, 161-2 allophones, 162 allophonic variation, 166-7 aspiration, 168 assimilation effect, 167 coarticulation, 168 colouring effect, 167 continuous nature issues, 169-70 distinctiveness issues, 171-2 IPA alphabet, 163-5 minimal pair, 163 modified timit ascii character set, 166 nasalisation colouring, 167 phonemes, 161-4 phones, 162-4

594

Cambridge University Press 978-0-521-89927-7 - Text-to-Speech Synthesis Paul Taylor Index <u>More information</u>

Index

speech, communicative ( <i>cont.</i> )	non-linguistic contents, 32–3
segments 162	physical patures of 27.8
targets 168 0	provide and verbal contents 30, 1
transcriptions 170–1	semiotic systems 33-4
speech production/articulatory phonetics 146–56	speech spontaneity 30
about speech production 146–7	usage of each 29–30
consonants, 153–5	speed of sound, 316
continuants, 150	SPE ( <i>The Sound Pattern of English</i> ), 181–2
egressive pulmonic air stream, 147	splicing costs, 499
examining speech production, 155–6	spoken text markup language, 69
fundamental frequency (F0), 148	standard deviation, 436
harmonics, 148–9	standing waves, 314
larynx, 148	statistical parsing, 105
neutral vowel sound, 152-3	status markers, 65
oral cavity, 152	stochastic signals, 288
oral and nasal sounds, 150	stop sounds, 150
source/filter model of speech, 151	stress in speech, 116
stop sounds, 150	lexical stress, 116, 186-9
timbre, 149	metrical stress, 188
unvoiced sounds, 150	reduced stress, 187
velum, 150	secondary stress, 187
vocal folds, 148	sentential stress, 186-7
vocal organs, 147	strict layer hypothesis, 114
vocal-tract filter, 150–1	style manuals, 68
vowels, 151–3	sum of random variables, 545–6
see also acoustic models of speech production;	sums-of-products model, 257-8
glottis/glottal source	superimpositional intonation models, 242
speech recognition, 1, 22	superposition of functional contours (SFC) model,
speech, redundancy in, 21	254
speech signals analysis, 341–86	support-vector machines, 221
about speech analysis, 341, 384–6	suprasegmentality, 124
spectral-envelope and vocal-tract representations,	syllables, 184–6
362-/2	boundaries, 206
see also cepstrum speech analysis; epoch	syllabic consonants, 184
detection; filter-bank speech analysis;	syllabic writing, 34
detection, course signal correspondences	symbolic language/communication, 10–12,
detection, source signal representations	120-/
speech signals analysis, short term, 541–52	combinations of symbols, 11–12
envelopes 345	syntactic ambiguity 46.7
equivalent rectangular handwidth (ERB) scale	syntactic analysis 102
	syntactic homonyms 88_9
frame lengths and shifts 345–9	syntactic prominence patterns 116–18
malescale 351	syntactic trees 102
nich-seale, 551	syntax 16
nitch-synchronous analysis, 347	syntactic hierarchy 16
spectral representations 343–5	syntactic phrases 16
spectrograms, 348–51	synthesis
time-frequency tradeoff. 346	articulatory synthesis, 405–7
windowing, 342–5	synthesis algorithms, future of, 536–7
Speech synthesis markup language (SSML), 69	synthesis specification. 387–8
speech technology/linguistics relationship, 533–6	see also classical linear-prediction (LP) synthesis;
future of, 537–8	formant synthesis; hidden Markov models
speech/writing comparisons, 26–35	(HMMs), synthesis from; second-generation
component balance, 31–2	synthesis systems; vocal-tract models,
form comparisons, 28–9	synthesis with

595

synthesis of prosody see autosegmental-metrical (AM) intonation model; data-driven intonation models; deterministic acoustic models, synthesis with; intonation ...; timing issues system lexicon, 214-15 system testing, 523 tagging, 82-3 talking-head synthesis, 406-7, 527 see also audio-visual speech synthesis targets, 168-9 tests/testing, 523-6 Blizzard Challenge testing, 526 comparison tests, 524 competitive evaluations, 526 Harvard sentences, 523 Haskins sentences, 523 modified rhyme test (MRT), 523, 524 naturalness tests, 524 semantically unpredictable sentences, 523 system testing, 523 test data, 525 unit/component testing, 525-6 word-recognition tests, 523-4 text analysis, future of, 536 text anomalies, 105 text-as-language TTS models, 39 text-classification algorithms, 79-92 ad-hoc approaches, 83 bag-of-features approach, 84 cluster impurity, 88 collocation rule, 84 context-sensitive rewrite rule, 83-4 curse of dimensionality, 81, 534 data driven approach, 80 decision lists, 85-6 decision trees, 87-8 deterministic rule approaches, 83 engine/rule separation, 83 features and algorithms, 79-82 hidden Markov model (HMM), 89-91 naive Bayes' classifier, 86-7 part-of-speech (POS) tagging, 82, 88-92 probabilistic approach, 80 statistical approach, 80 tagging, 82-3 trigger tokens, 84-5 unsupervised approach, 80 word-sense disambiguation (WSB), 82-3 text decoding/analysis, 22, 52-3, 78-110 about text decoding, 78-9, 105-10 see also natural-language parsing; non-natural-language text decoding; text-classification algorithms text materials, 518 text normalisation, 44, 106

text segmentation and organisation, 63-8 about text segmentation, 52-3, 75-7 sentence splitting, 67-8 tokenisation, 64-7 see also architectures for TTS; processing documents; sentences; words text-to-speech (TTS) about text-to-speech, 1-2, 26, 50-1 basic principles, 41 common-form model, 5-6 development goals, 3 engineering approach, 4 intelligibility issues, 3 naturalness issues, 3 purposes, 2 see also models of TTS; problems in text-to-speech texture mapping, 528 third-generation techniques see hidden Markov model (HMM); unit-selection synthesis Tilt intonation model analysis with, 250 synthesis with, 227, 242-4, 249-50 time-domain PSOLA (TD-PSOLA), 416-17 pitch-scale modification, 416-17 time-scale modification, 416 time-frequency tradeoff, in speech analysis, 346 time invariance, assumptions concerning, 337 timing issues, 254-9 about timing, 254-5 Campbell model, 258 durations, 254 Klatt rules, 256-7 nature of timing, 255-6 phase-final lengthening, 256 sums-of-products model, 257-8 TIMIT phoneme inventory, 203-6, 553 modified timit ascii character set, 166 ToBI intonation scheme, 237, 247, 248 Toeplitz matrix, 361 token, definition, 54 tokenisation, 53, 64-7 and punctuation, 65-6 tokenisation algorithms, 66-7 tone languages, 124, 227 tonemes, 121 transcriptions, 170-1 transfer-function poles, 364-5 transforms, 284-8, 307 about transforms, 284 analytical analysis, 287 convolution, 287 duality for time and frequency, 284-5 frequency shift, 286 impulse properties, 285 Laplace transform, 283

596

Cambridge University Press 978-0-521-89927-7 - Text-to-Speech Synthesis Paul Taylor Index <u>More information</u>

Index

transforms (cont.)	original/derived features, 481
linearity, 284	partial synthesis, 482
modulation, 286	script technique, 480
numerical analysis, 287	target feature structure, 481
scaling, 285	unit-selection synthesis, searching, 504-9
stochastic signals, 288	about searching, 504–5
time delay, 286	beam pruning, 509
<i>z</i> -transform, 282–3	diphone unit-selection system, 505
see also Fourier transform	half-phone solution, 506–7
translation from semiotic classification, 45-6	Hunt and Black algorithm, 504
tree-banks, 105	multi-pass searching, 509
trigger tokens, 84–5	pre-selection, 508–9
triphone models, 451	pruning methods, 508–9
tune and intonation, 121–2	unit back-off solution, 505-6
	Viterbi algorithm/search, 504-5, 508
understanding, 19, 22-3	unit-selection synthesis, target function formulation,
uniform distribution, 549	484–93
unit back-off searching, 505-8	about the target function, 484-5
unit/component testing, 525-6	acoustic-space formulation (ASF), 485, 493-7
unit-selection databases, 517–18	context-orientated-clustering, 495
speaker choice issues, 518	decision-tree clustering, 494-6
unit-selection synthesis, 474–516	disruption issues, 485
about unit selection synthesis, 251–2, 474–9,	distance/cost issues, 484
510-11, 515-16	equal-error-rate approach to learning, 491
ATR family contribution, 512–14	Euclidean distance, 486
CHATR system, 513	feature axis scaling, 488
concatenation of units, 477	full set of candidates, 484
coverage, 510	general partial-synthesis functions, 496–7
extending from concatenative synthesis, 475–7	hand tuning, 491
features, cost and perception, 511–12	independent feature formulation (IFF), 485–8
HMM system 514	independent-feature formulation limitations
Hunt and Black algorithm 477–9	491–3
Laureate system 513	Manhattan distance 486
NextGen system ( $AT\&T$ ) 513–14	perceptual approaches 490–1
nrinciple of least modification 477	perceptual approaches, 490 T
nure unit selection 477	perceptual substitutability principle 485
RealSpeak system 514	search candidates set 484
resequencing algorithms 477	target weights setting 488-01
rVoice system 514	unknown words
signal processing issues 511	decoding 98
signal processing issues, 511	problems with 216_18
unit selection synthesis features $470$ 84	unvoiced sounds 150
has type 470 10	LITE 8 71
dimensionality reduction/acourtouv tradaoff 492	UTF 16 71
facture choosing 481-2	UTF-10, /1
feature combination structures 481	utterance structure, /1
feature temps, 482, 4	vortion og 126
head labelling technique 480 1	variance, 450
hataraganaous systems, 480	vertuil, 150
homogeneous systems, 480	verbai communication, 14–10
nomogeneous systems, 480	aroutraryness, 15
interingibility issue, 510	uiscreteness, 15–16
Join feature structure, $481$	quality, 15
lett/right join feature structure, 481	pnonemes, 14
linguistic and acoustic features, 480–1	productiveness, 15
naturalness issues, 510	sentences, 14
non-uniform unit synthesis, 480	words, 14

Index

597

verbalisation, 95-7 verb-balancing rule, 131 visemes, 528 Viterbi algorithm, 92, 444-8, 504-5, 508 vocal organs, 147 vocal-tract filter, 150-1 sound loss models, 335-6 straight tube assumptions, 337 transfer function, 310-11 vocal-tract models, synthesis with, 387-411 about synthesis with vocal-tract models, 387, 407-11 ease of data acquisition, 407 effectiveness of models, 407 modularity issues, 407 synthesis specification, 387-8 see also articulatory synthesis; classical linear-prediction (LP) synthesis; formant synthesis; residual-excited linear prediction; vowel-tube models vocal-tract and spectral-envelope representations, 362-72 voice transformation, and synthesizing emotion, 530 VoiceXML, 69 vowel sounds, 151-3 diphthongs, 153 monophthongs, 153 neutral vowel, 152 vowel-tube models, 319-30 about the vowel tube, 319 all-pole resonator model, 329-30

discrete time and distance, 320 junction special cases, 322-3 junction of two tubes, 320-2 multi-tube vocal-tract model, 327-9 reflection coefficient, 322 single-tube vocal-tract model, 325-7 transmission coefficient, 322 two-tube vocal-tract model, 323-5 windowing, 342-5 word formation/lexical phonology, 179-81 word-recognition tests, 523-4 words, 14 ambiguity issues, 54 defining in TTS, 55-9 definitions/terminology, 54-5 form issues, 53-4 hyphenated forms, 61-2 shortened forms, 61 slang forms, 61 word variants, 57 word-sense disambiguation (WSB), 82-3 writing see speech/writing comparisons writing systems, 34-5 Abjab, 34–5 alphabetic, 34 logographic, 34 pictographic, 34 syllabic, 34 z-transform, 282-3 and digital filters, 293-4, 297-8