

Index

- μ -law, 6, 14, 148
- .wav file format, 10, 13, 14
- A-law, 6, 14, 19, 109, 148
- A-weighting, 89, 131, 191
- absolute
 - pitch, 97
 - scaling, 20
- accelerometer, to measure pitch, 208
- accent, 303
- accuracy, 234
- ACELP, *see* codebook excited linear prediction, algebraic
- Adams, Douglas, 307
- adaptive differential PCM, 144, 148, 177
 - sub-band ADPCM, 147
- ADC, *see* analogue-to-digital converter
- ADPCM, *see* adaptive differential PCM
- affricative, 64
- ageing, effect of, 303
- allophone, 63
- allotone, 63
- AMDF, *see* average magnitude difference function
- analogue-to-digital converter, 4, 5, 13, 141
- analysis, 149, 158, 180, 196
- analysis algorithm, 33
- analysis-by-synthesis, 182
- angular frequency, 40
- animal noises, analysis of, 219
- aphonia, 311
- Apple
 - iPod, 1, 109
 - Siri, 2, 366
- articulation index, 82, 270
- ASA, *see* auditory scene analysis
- ASR, *see* automatic speech recognition
- audio
 - waveform, 17
- audio fidelity, 6, 122, 148, 149
- audioread(), 13, 229
- audiorecorder(), 10
- audiowrite(), 16
- auditory adaptation, 94, 103
- auditory cortex, 93
- auditory fatigue, 93
- auditory image, 231
- auditory scene analysis, 112
- authentication, 363
- autocorrelation, 41, 160
- automatic speech recognition, 224, 267–269, 286, 298, 301, 310
- average magnitude difference function, 199, 200, 209, 211
- Babbage, Charles, 225
- back propagation, 239
- backwards speech, 102
- bandwidth, 142
- Bark, 104, 124, 129, 131, 137
- bark2f(), 105, 106, 337
- basilar membrane, 94
- Bayes' theorem, 293
- Bayes, Thomas, 293
- Bayesian inference criteria, 309
- Bell, Alexander Graham, 1
- Bernoulli–Bernoulli RBM, 254
- big data
 - rationale, 225
 - sources of, 226
 - storage, 227
- big endian, 15, 17, 19
- binaural, 97, 98, 324
- biometric, 363
- birdsong, 219
- blind source separation, 242
- Blu-ray, 1
- bronchial tract, 55
- buffer(), 27
- cadence, of speech, 216
- calibrated reference speaker, 88
- candidate vectors, 184, 186
- cceps(), 204
- cceps(), icceps(), 42
- CD, 6, 121, 330
- CELP, *see* codebook excited linear prediction
- cepstrum, 40, 42, 202, 204, 209, 211, 299

- Chinese, 60, 61, 63, 70, 216, 272, 288, 299, 306, 308, 310, 317
- chirp
 exponential, 46
 generation of, 46
 linear, 46
- chirp(), 34, 46, 90
- chord, musical, 50, 100, 118
- classification, 230, 232, 235, 239, 255, 292
- classification error, 215
- click, 27
- closure, of sounds, 114
- clustering, 247, 250
- CNN, *see* convolutional neural network
- co-modulation masking release, 95
- code excited linear prediction, 183
- code switching, 310
- codebook, 286, 350
- codebook excited linear prediction, 140, 153, 165, 172, 174, 183–186, 188, 191, 207, 338, 341, 348
 in speech synthesis, 316
 algebraic, 186
 computational complexity of, 186
 forward–backward, 189
 latency, 189
 split codebooks, 187
 standards, 190
- coffee, 304
- combination tones, 90
- comfort noise, 116
- common fate, of sounds, 116
- compression, 228
- concatenative synthesis
 of speech, 316
- concert pitch, 45
- consonant, 63, 64
- contextual information, 80
- continuous processing, 28
- continuous speech recognition, 268
- contour plot, 213
- contour spectral plot, 214
- convergence, 230
- conversion between reflection coefficients and
 LPCs, 161
- conversion of LPC to LSP, 168
- convolutional neural network, 259
- correlogram, 40, 41
- covariance, 160
- critical band, 94, 95, 110, 123, 124, 127–129, 134, 137, 148
 filters, 104
- CVSDM, *see* continuously variable delta modulation
- DAB, *see* digital audio broadcast
- DAC, *see* digital-to-analogue converter
- data
 storage cost, 223
 storage, 227
 value of, 223
- database
 public, 228
- dBa, 62
- DBF, 309
- DBN, *see* deep bottleneck network
- DCT, *see* discrete cosine transform
- decoding, 293
- deep belief network, 257
- deep bottleneck features, 305
- deep bottleneck network, 286
- deep neural network, 253, 256, 257, 305
- DeepLearn Toolbox, 258, 261
- delta modulation, 143
 adaptive, 144
 continuously variable slope, 144
 slew rate, 143, 144
 slope overload, 143
- development platform, 3
- DFT
 discrete Fourier transform, 36
- diagnostic rhyme test, 79, 210
- dialogue, 368
- diarization, 231, 302, 308, 309
- dictionary, 299
- differential PCM, 145
- digital audio broadcast, 1
- digital compact cassette, 109
- digital-to-analogue converter, 4, 5, 13
- dimensionality, 229
- diphthong, 64
- dir(), 229
- directory name, 229
- discontinuities, 27
- discrete cosine transform, 33
- discrete digital signal processing, 142
- discrete Fourier transform, 36
- DNN, *see* deep neural network
- dog bark, 219, 220
- Dolby, 330
- DRT, *see* diagnostic rhyme test
- DTMF, *see* dual tone multiple frequency
- DTW, *see* dynamic time warping
- dual tone multiple frequency, 182
- Durbin–Levinson–Itakura method, 161
- DVD, 1
- dynamic range, 6, 142
- dynamic time warping, 303
- dysarthria, 311
- dysphonia, 311
- ear
 basilar membrane, 85
 cochlea, 85

- drum, 17, 85
- human, 85
- organs of Corti, 85
- protection, 86
- Reissner's membrane, 85
- earache, 87
- ECHELON, 368
- echoes, 101
- EGG, *see* electroglottograph
- electroencephalogram, 268
- electroglottograph, 208
- electrolarynx, 348, 350, 353, 363
- emotion, 310
- endian, 15, 19
- enhancement
 - of sounds, 110
 - of speech, 111
- equal loudness, 89, 123, 131, 133
- equal-loudness contours, 87, 103, 109

- f2bark(), 105, 106, 337
- f2mel(), 106
- fast Fourier transform, 21, 23, 32–36, 39, 40, 75, 128, 195, 199, 201, 202
- fclose(), 16
- feature transformation, 227
- features, 227, 229, 230, 232, 278, 284, 346
- Festival speech synthesis system, 323
- FFT, *see* fast Fourier transform
- fft(), 21, 34, 130, 203
- FFTW, 22
- file path, 229
- filter, 21
 - analysis, 151, 152, 184
 - continuity of, 28
 - FIR, 21, 30, 158
 - history, 31
 - IIR, 21, 151, 178
 - internal state of, 31
 - LPC, 165, 167
 - pitch, 179
 - pole-zero, 21
 - stability, 157
 - synthesis, 151, 152, 191
- filter(), 21, 29, 31, 158
- filterbank, 231
- finite impulse response, 21
- Fletcher–Munson curves, 87
- fopen(), 15
- formant strengthening, 334
- formants, 58, 152, 279, 351
- Fourier transform, 21, 36, 37, 42, 128, 211
- frame, 231
- frame power, 198, 211
- fread(), 15, 17
- freqgen(), 47, 48, 115, 119
- frequency discrimination, 96, 104, 123
- frequency resolution, 22
- frequency selectivity, 86
- freqz(), 152, 170
- fricative, 64
- front-end clipping, 281
- FS1015, 190
- FS1016, 190
- fwrite(), 16

- G.711, 148
- G.721, 148
- G.722, 147, 148
- G.723, 148, 190
- G.726, 148
- G.728, 182, 190
- G.729, 190
- Gaussian
 - process, 242
- Gaussian–Bernoulli RBM, 254
- Gaussian mixture model, 250, 283, 347
- gestures, 274
- getaudio(), 12
- getaudiodata(), 11
- Gibbs phenomena, 27
- glide, 64
- glissando, 119
- global system for mobile communications, *see* GSM
- glottis, 56, 69, 150, 162, 165, 166, 176, 208, 303, 352
- GMM, *see* Gaussian mixture model
- golden ears, 6, 7
- good continuation, of sounds, 119
- GPU, *see* graphics processing unit
- graphics processing unit, 301
- ground truth, 281
- Global System for Mobile Communications, *see* GSM
- GSM, 2, 6, 116, 177, 190
- Gulliver's Travels*, 19

- Haas effect, 100
- handphone, 2
- hang time, 280
- harmonics, 92, 100, 118, 217
- HAS, *see* human auditory system
- hearing, 86, 87
- hearing loss, 93
- Hermansky, 124, 127, 129, 132
- hidden Markov model, 207, 288, 299, 301, 303
- hidden Markov model toolkit, *see* HTK
- HMM, *see* hidden Markov model
- HTK, 105, 250, 301, 311
- human auditory system, 18, 112, 116, 148
- human ethics, 226, 227

- i-vector, 231, 285
- ICA, *see* independent component analysis
- iccps(), 204
- ifft(), 203
- iFlytek, 225, 226, 263
- IR, *see* infinite impulse response
- imagesc(), 212
- impersonation, 306
- independent component analysis, 242
- induction, auditory, 114
- infinite impulse response, 21
- infinite impulse response filter, 151
- international phonetic alphabet, 63, 320
- International Telecommunications Union, 72, 148
- IPA, *see* international phonetic alphabet
- IS, *see* Itakuro–Saito distance
- Itakura–Saito distance, 77, 78
- ITU, *see* International Telecommunications Union

- J.A.R.V.I.S, 366
- JND, *see* just-noticeable difference
- just-noticeable difference, 105

- K–L tube model, *see* Kelly–Lochbaum model
- k-means, 247, 249
- Kaldi, 301, 311
- Kelly–Lochbaum model, 162
- key-phones, 307
- keyword spotting, 268
- Kirk, Captain, 319

- language
 - n-gram model, 300
 - classification of, 305
 - grammar, 303
 - identification, 231, 305
 - learning, 310
 - model, 299
 - morphology, 306
 - phonology, 306
 - syntax, 306
- LAR, *see* log area ratios
- large vocabulary continuous speech recognition, 300
- laryngectomy, 347, 348, 350, 353
 - partial, 348
- larynx, 303
- lateral inhibition function, 124
- LDA, *see* linear discriminant analysis
- Le Roux method, 161
- learning, 235, 239
- line spectral frequencies, *see* line spectral pairs
- line spectral pairs, 159, 165, 166, 168, 170–174, 204, 208, 209, 211, 216, 217, 219, 299, 334, 338, 341
- linear discriminant analysis, 242
- linear prediction, 149, 299
- linear predictive coding, 43, 76, 150–152, 154, 157–162, 165, 168, 171–174, 177, 178, 183–191, 195, 209, 219, 350
- linear time invariant, 348
- lips, 303
- lisp, 311
- little endian, 15, 17, 19
- LLR, *see* log-likelihood ratio
- load, 16
- log area ratios, 157, 177
- log-likelihood ratio, 76, 77
- Loizou, Philip, 77
- long-play audio, 6
- long-term prediction, 177, 178, 183–190, 208, 340, 350
- long-term predictor, 60
- loudness, 103, 104
- LPC, *see* linear predictive coding, 231
- LPC cepstral distance, 76, 77
- lpc(), 151, 217
- lpc_code(), 154
- lpcsp(), 165
- lpsp(), 171
- LSF, *see* line spectral pairs
- LSF interpolation, 172
- LSP, *see* line spectral pairs
 - analysis, 204, 207
 - instantaneous analysis, 204
 - time-evolved analysis, 207
- lsp_bias(), 205
- lsp_dev(), 205
- lsp_voicer(), 345
- lspnarrow(), 336, 338
- LTI, *see* linear time invariant
- LTP, *see* long-term predictor
- ltp(), 340
- lung, 55, 56
 - excitation, 176, 184
- LVCSR, *see* large vocabulary continuous speech recognition

- machine learning, 224, 227, 230, 232, 272
- magnetic resonance imaging, 86, 268
- Mandarin, *see* Chinese, 300
- masking, 95, 103, 109, 123, 137
 - binaural, 97, 111
 - non-simultaneous, 96, 121
 - post-stimulatory, 96, 121
 - pre-stimulatory, 96
 - simultaneous, 93, 94, 148
 - temporal masking release, 111
 - two-tone suppression, 111
- max pooling, 259
- max(), 153

- McGurk effect, 86, 112
 mean opinion score, 72, 142
 mean-squared error, 73, 160, 179
 mel frequency, 43, 105, 126, 130, 134, 207
 mel frequency cepstral coefficients, 43, 123
 mel2f(), 106
 MELP, *see* mixed excitation linear prediction
 MFCC, *see* mel frequency cepstral coefficients, 231, 286, 299, 301, 304, 309, 347
 microphone
 directional, 273
 mid-speech clipping, 281
 mid-side stereo, 330
 mind control, 268
 MiniDisc, 109
 mixed excitation linear prediction, 188, 190, 348
 MNB, 73
 mobile telephone, 2
 modified rhyme test, 79
 monaural, 97
 mono-phonemes, 307
 mood, 310
 MOS, *see* mean opinion score
 MP3, 1, 14, 15, 109, 121, 122, 137, 216
 MP4, 16
 MPEG, 15
 MRI, *see* magnetic resonance imaging
 MRT, *see* modified rhyme test
 MSE, *see* mean-squared error
 multi-layer perceptron, 239, 253, 260
 multilingual, 310
 music
 analysis of, 216
 musical notes, 45

 nasal, 64, 99
 NATO phonetic alphabet, 80
 natural frequency, 40
 natural language, 267
 natural language processing, 269, 320, 368
 Newton–Raphson approximation, 167
 NIST, 271, 305, 307
 NLP, *see* natural language processing
 noise
 detected as speech, 281
 effect of correlation, 326
 background, 57, 368
 cancellation, 122
 characteristics, 61
 generation of, 46
 masking, 124
 perception, 89, 93, 94, 111
 reduction through correlation, 111
 response of speaker to, 61
 non-audible murmur, 353
 normal distribution, 242

 normalisation, 18
 Nyquist, 5, 22, 35, 46, 127, 152

 objective function, 247
 occupancy rate, of channel, 215
 Octave, 229
 Ogg Vorbis, 14–16, 109
 orthotelephonic gain, 89
 overlap, 25, 27, 231
 overlap–window–add, 27

 parallel processing, 229
 parameterisation, 157
 parameterisation of speech, 148, 150, 154
 PARCOR, *see* partial correlation
 partial correlation, 159, 160
 path, 229
 pause(), 11
 PCA, *see* principal component analysis
 PCM, *see* pulse coded modulation
 perceptron, 234, 239
 perceptual error, 184
 perceptual error weighting filter, 191
 perceptual weighting, 183, 191
 perfect pitch, 97
 PESQ, 73
 PEWF, *see* perceptual error weighting filter
 phase locking, 92
 phone, 63, 307, 323
 phoneme, 57, 63, 66, 67, 79, 208, 299, 302, 317
 phonemes, 69, 162, 350
 phonetic spelling, 317
 phonograph, 1
 phonology, 306
 pitch, 33, 231, 279, 348, 350–352
 pitch lag, fractional, 178
 pitch perception, 90, 91, 96, 100, 110
 pitch synchronous overlap and add, 339, 341
 plain old telephone service, 2
 play(), 11
 plosive, 57, 64
 PLP, 309
 Pocketsphinx, 312
 poly-phones, 307
 POTS, *see* plain old telephone service
 pre-emphasis of the speech signal, 158
 precedence effect, 100
 precision, 228
 principal component analysis, 242, 285
 private mobile radio, 6
 pronunciation, 299
 in speech synthesis, 317, 320
 in speaker classification, 303
 in speech synthesis, 303
 of phonemes, 63
 prosody, 306

- proximity, 113
- pseudo-stationarity, 32, 67, 150, 159
- pseudo-Wigner–Ville distribution, 210
- PSOLA, *see* pitch synchronous overlap and add
- PSQM, 73
- psychoacoustics, 87, 94, 109, 121–124, 129–131, 191
- puberty, 303
- pulse coded modulation, 7, 13, 14, 85, 141, 144, 145
 - standards, 148
- PWVD, *see* pseudo-Wigner–Ville distribution
- quantisation, 6, 140, 145, 146, 148, 154, 159, 171, 182, 185, 228
 - of stereo audio, 329
 - of LSPs, 173
 - of audio samples, 11, 17
 - of LPC parameters, 157, 171
 - of LSPs, 171, 172, 174
 - of speech, 150
 - split vector, 176
 - vector, 175
- quantisation error, 142–144, 146
- RBM, *see* restricted Boltzmann machine
- reassigned smoothed pseudo-Wigner–Ville distribution, 210
- recall, 232
- record(), 10
- redundancy, 80
- reflection coefficients, 159, 161
- regular pulse excitation, 177, 183, 184, 190, 348
 - standards, 190
- reinforcement learning, 232
- relative scaling, 20
- reshape(), 101
- residual, 153
- resonance conditions, 165, 166
- restricted Boltzmann machine, 253
- resume(), 11
- RPE, *see* regular pulse excitation
- RSPWVD, *see* reassigned smoothed pseudo-Wigner–Ville distribution
- sample rate, 6, 18, 142
- sampling, 5
- save, 16
- scaling of amplitude, 20
- SD, *see* spectral distortion
- segmental signal-to-noise ratio, 74, 171
- segmentation, 21, 24, 25, 27, 28, 31, 32, 57, 177, 274
 - in CELP coder, 183
- SEGSNR, *see* segmental signal-to-noise ratio
- semi-supervised learning, 232
- semilogy(), 39
- sensing, 223
- Shatner, William, 319
- short-time Fourier transform, 38, 209
- SIFT, *see* simplified inverse filtering technique
- signal processing, 93
- signal-to-noise ratio, 74
- silent speech interface, 310, 364
- simplified inverse filtering technique, 209
- size constraints, 140
- slope overload, 143
- smartphone, 2
- SNR, *see* signal-to-noise ratio
- softmax, 255
- sone, 103
- sound detection, 231
- sound perception, 92, 93
- sound pressure level, 61
- sound strengthening, 110
- sound(), 11, 18, 49
- soundsc(), 12, 18, 29, 49, 90, 92, 95, 100, 101, 114, 326
- spatial placement, 327
- speaker
 - authentication, 363
 - identification, 231, 302
 - validation, 363
 - verification, 302
- speaker identification, 43
- specgram(), 38
- spectrogram(), 38
- spectral distortion, 75, 172, 183
- spectrogram, 30, 38, 211, 214, 231
- spectrogram time–frequency distribution, 209
- spectrogram(), 34, 38, 212
- speech
 - activity detection, 275
 - concatenative synthesis, 315
 - pitch changer, 338
 - pitch period, 339
 - transformation of, 346
 - amplitude, 61, 62, 304
 - analysis of, 224
 - apraxia, 311
 - articulation, 64, 66
 - atypical, 70
 - backwards, 102
 - cadence, 216
 - characteristics, 57
 - characteristics of, 149
 - classification, 58, 208
 - codec, 142, 311
 - coding algorithms, 149
 - colloquial, 310
 - compression of, 122, 142, 150, 183, 190, 191, 311
 - disorder, 311
 - energy, 58, 65, 231
 - formants, 58, 59, 65, 66, 110, 138, 191, 200, 337

- frequency distribution, 65
- impediment, 311
- intelligibility, 58, 65, 71, 81, 82, 97, 102, 122, 138, 178
- intelligibility testing, 79
- intelligibility vs. quality, 71
- lexical, 62
- perception, 101
- pitch contour, 60
- pitch doubling, 181, 182
- pitch extraction, 178, 179, 182, 208
- pitch halving, 182
- pitch models, 176
- pitch period, 182, 202
- pitch synthesis, 150
- power, 65
- production, 55
- quality, 71, 102, 148
- quality testing, 72
- recognition, 43, 110, 267, 303
- repetition, 81
- reversed, 102
- shouting, 68
- sinewave, 54, 196
- spectrum, 199, 201
- statistics, 303
- super-audible, 362
- synthesis, 314
- to whisper, 354
- unvoiced, 64
- voiced, 64
- voicing, 215
- waveform, 199, 214
- speech recognition, 231
- Sphinx, 301, 312
- SPL, *see* sound pressure level
- split vector quantisation, 176
- spread_hz(), 125
- spreading function, 124
- SSI, *see* silent speech interface
- stammer, 311
- Star Trek*, 307, 319, 366
- stationarity, 32, 67
- statistical voice conversion, 346, 353
- statistics, of speech, 213, 303
- steganography, 122
- stepsize doubling and halving, 144
- stereo, 4, 98, 324
- stereo encoding
 - joint, 329
- Stewart, Patrick, 319
- STFD, *see* spectrogram time–frequency distribution
- STFT, *see* short-time Fourier transform
- stop(), 11
- stress, on words, 275
- stutter, 311
- subsampling, 259
- super-vector, 285
- superposition, principle of, 332
- supervised learning, 232
- support vector machine, 239, 255
- surround sound, 330
- SVM, *see* support vector machine
- swapbytes(), 17, 18
- sweet spot, 332
- syllabic rate, 67, 215, 303
- syllable, 63, 66, 67, 79, 215
- synthesis, 149
- synthesiser
 - of speech, 314
- TCR, *see* threshold crossing rate
- telephone, 1
 - mobile, 2
- temporal integration in hearing, 93
- temporary threshold shift, 93, 111
- TETRA, *see* trans-European trunked radio
- text-to-speech, 367
- TFD, *see* time–frequency distribution
- three-dimensional mesh spectral plot, 214
- threshold crossing rate, 198, 211
- timbre, 92
- time-domain waveform, 37, 200
- time–frequency distribution, 209–211
- TIMIT, 173
- toll-quality, 6
- tone
 - generation of, 44, 47
 - induction, 110
- tonegen(), 44, 48, 88, 90, 91, 95, 97, 100, 113, 326
- tongue, placement of, 303
- topic, 272, 310
- training, 230, 232
- trans-European trunked radio, 190
- transcription, 274
- transcription systems, 269
- transducer, 273
- transfer learning, 286
- tri-phone states, 299
- TTS, *see* temporary threshold shift *or* text-to-speech
- tube model, 162–166
- Turk, the, 314
- UBM, *see* universal background model
- universal background model, 283
- unsupervised learning, 232
- VAD, *see* voice activity detection
- variable tone generation, 47
- vector quantisation, 175, 286, 303
- vector sum excited linear prediction, 188
- velum, 55, 56

- violin, 217, 218
 - analysis of, 216
- visualisation, 34, 37
- Viterbi algorithm, 293, 309
- Viterbi, Andrew, 293
- vocabulary, 272
- vocal chord, *see* glottis
- vocal tract, 58, 63, 64, 176, 177, 191, 303
 - filter, 150
 - parameters, 184
 - resonances, 183, 184
- voice
 - cloud, 226
 - scrambler, 338
- voice activity detection, 231, 274, 275, 309, 362, 368
- voice operated switch, 275
- Voicebox, 164, 250
- VOS, *see* voice operated switch
- vowel, 57, 58, 63, 64, 303
- VSELP, *see* vector sum excited linear prediction
- waterfall(), 34
- wave file format, 10, 13, 14, 212
- waveread(), 13
- waverecord(), 12
- wavwrite(), 16
- weighting, 89
- whisperise(), 357
- whisperiser, 354
- whispers, 228, 348, 349
- white noise, 46, 55, 111, 123, 184
- wideband coding, 148
- Wigner–Ville distribution, 209, 210
- windowing, 27, 28, 161, 231
 - in CELP coder, 183
 - window functions, 27–29
 - window size, 32, 35
- WVD, *see* Wigner–Ville distribution
- xcorr(), 40
- ZCR, *see* zero-crossing rate
- zero-crossing rate, 196, 207, 209, 211