
Index

Page numbers followed by italic *f* or *t* denote figures or tables, respectively.

A

Absolute distance, 44
Acoustic measures
 for voice quality differences, 25–28,
 26*t*, 32
 vs. phonetic features, 41
Acoustic source models, 68–69
Activation, as emotional dimension,
 177–178, 177*t*
Affect Editor, 179–181, 181*t*
Anger, in expressive speech synthesis,
 227–228, 228*t*
Articulatory modeling, 63, 83–84
 acoustic sources in, 68–69
 concatenative, 75–78
 definition, 77
 natural speech units in, 78–79
 parameters, 66–68, 67*f*
 prototype system using, 82–83
 rule-based control of, 74–75
 synthesis from parameters in, 69–74
 terminology, 76–78
 vocal tract acoustics in, 65–66
ASR. *See* Automatic speech
 recognition (ASR)
Autocorrelation coefficients
 in voice quality variation, 27
Automatic speech recognition (ASR)
 annotation tools, 117–118
 development of, 110–111
 limitations of, 121–123
 parametric synthesis and, 124–126
 search algorithms in, 114–117, 115*f*,

117*f*

 system components, 111–112, 111*f*

B

Buzziness, causes of, 10

C

CNET/Elan system, 212
Codebook, articulatory, 80–81
Cognitive theory, of emotions, 176
Coker articulatory model, 67, 67*f*, 70*f*,
 81
Coker TTS system, 77–78
Complex factorization scheme, 93–94,
 93*t*
 vs. foot factorization scheme,
 98–100, 100*t*
Concatenative synthesis. *See also*
 Corpus-based synthesis
 advantages, 182
 data collection, 186–189
 definition, 76–77
Constructivist theory, of emotions,
 176–177
Context clustering, 138–139, 139*f*
Contour control model, 159–160*t*,
 159–162, 161*f*, 162*f*
Corpus-based synthesis. *See also*
 Concatenative synthesis
 concatenation algorithm, 8–12
 definition, 76–77
 disadvantages, 90
 evaluation, 12–15

- philosophy of, 90
- for prosody, 206
- shift-only F0 smoothing, 3–8
- single-speaker. *See* Voice quality variation
- Correlation coefficients, of voice
 - quality differences, 28–31, 29*f*, 30*t*, 31*f*
- Cost minimization, in shift value calculation, 4
- D**
- Darwin, Charles
 - theory of emotions, 175–176
- Data-driven synthesis. *See* Concatenative synthesis
- Delta coefficients, of speech
 - parameters, 43–44
- Diphone synthesis, 90–91
- Diphone transplantation, 205–206
- Diphthong, spectral discontinuities in, 47
- Discontinuities. *See also* Pitch smoothing
 - calculating shifts for, 4–7, 6*f*, 7*f*
 - frequency of, 2
 - prediction of, 38–39, 41
 - reduction of, 39
 - types of, 2
- Distance measures
 - absolute, 44
 - Euclidean, 44
 - Kullback-Leiber, 37–38, 45
 - mahalanobis, 38–39, 44–45
 - perceptual scores and, 50–53, 51*f*, 52*t*, 53*t*
 - of pitch contours, 96–98
 - weighted, 46, 53–56, 98
- Distribution clustering, in automatic speech recognition, 112
- Duration corpus, 95, 100*t*
- Duration modeling, 138
- E**
- EMOSYN, 181
- Emotional speech synthesis. *See* Expressive speech synthesis
- Emotion(s)
 - activation and evaluation, 177, 178*t*
 - Plutchik’s wheel, 177, 177*f*
 - prosody of, 184–185, 184*t*
 - standards in research on, 196
 - theories of, 175–176
 - vs.* expression, 178
- Emphasis, in expressive speech synthesis, 236
- ESpeech, 189
- Euclidean distance, 44
- Evaluation. as emotional dimension, 177–178, 177*t*
- Excitation function, of vocal cord vibration, 73–74
- Expression, *vs.* emotion, 178
- Expressive speech synthesis. *See also* Speech synthesis markup language (SSML)
 - adaptation in, 233–234
 - anger ratings, 227–228, 228*t*
 - of apologetic text, 232–233, 233*t*
 - assessment of, 242–244
 - baseline system, 224
 - concatenative method, 183–185
 - data collection, 185–189, 224–225, 230
 - experimental evaluation, 189–195, 192–193*f*, 194*t*
 - expression detection in, 228–229, 229*t*
 - F0 target estimation in, 230–231
 - formant method, 179–182, 181*t*
 - liveliness ratings, 225–227, 226*t*
 - overview, 175–176, 221–223
 - prosody in, 184–185, 184*t*
 - rule-based method, 234–235
 - sadness ratings, 227, 227*t*
 - target-duration estimation in, 231–232
 - use of, 235–237

F

F0. *See* Fundamental frequency (F0)

Festival speech synthesis system

expressive speech application, 182

minimizing pitch modification in,
101–105

multilingual application, 147–148,
149*f*

Foot corpus, 95–96, 96*t*, 97*t*, 100*t*

perceptual experiment, 101–105

Foot factorization scheme, 93, 93*t*, 94*f*,

95, 96*t*, 97

vs. complex factorization scheme,
98–100, 100*t*

Formant synthesis, 77, 179–181. *See*

also Rule-based synthesis

Fundamental frequency (F0)

modeling of, 138

contour control in, 159–160*t*,

159–186, 161*f*, 162*f*

multispace probability distribution
in, 140–143, 142*f*

shift-only smoothing of, 3

in voice quality variation, 28

H

HAMLET system, 181

Hidden Markov model (HMM)

in automatic speech recognition, 112,
118–119

in expressive speech synthesis
system, 224

fundamental frequency modeling in,
140–143, 142*f*

in Japanese text to speech synthesis.

See Japanese text to speech
synthesis

in multilingual speech synthesis
system. *See* Multilingual speech
synthesis system

multispace probability distribution,
140–143, 142*f*

prosody control for, 155–156

speech-parameter generation
algorithm, 143–147, 147*f*

unit selection in, 115*f*, 148–150, 149*f*,
150*f*

HMM. *See* Hidden Markov model
(HMM)

I

Ignorance modeling, 110

J

James, William

theory of emotions, 176

Japanese text to speech synthesis,

155–157, 158*f*. *See also*

Multilingual speech synthesis
system

F0 contour control model for,

159–160*t*, 159–162, 161*f*, 162*f*

phoneme-duration-control model,

162–167, 163–165*t*, 166*f*, 167*f*

quantification theory in, 157–158

speech-rate-variable method,

168–170

Join cost(s)

comparison of, 36–37

definition, 35–36

inventory enrichment and, 39

perceptual scores and, 50–53, 51*f*,
52*t*, 53*t*

of phonetic features, 41

spectral measures in, 37–39

target cost and, 40

weighted subcosts of, 40–41, 53–56,
55*t*, 56*t*, 57–58

K

Kullback-Leibler divergence

computation, 45

in hidden Markov models, 118

studies of, 37–38

L

LIMSI NUU system, 212. *See also*

Nonuniform unit synthesis; Unit
selection synthesis

Line spectral frequency (LSF)

definition, 43

- join costs using, 53*t*, 55*t*
- Listening tests. *See* Preference tests
- Liveliness, in expressive speech synthesis, 225–227, 226*t*
- LSF. *See* Line spectral frequency (LSF)
- M**
- Mahalanobis distance measure, 38–39, 44–45
- Mass-spring model, 68
- Maximum likelihood linear regression (MLLR), 113, 121–122, 122*f*
- MBROLA. *See* Multiband resynthesis overlap add (MBROLA)
- MCA. *See* Multiple centroid analysis (MCA) coefficients
- Mel-cepstral feature extraction, 111–112
- Mel-frequency cepstral coefficient (MFCC)
 - definition, 43
 - join costs using, 52*t*, 55*t*
 - as predictor of perceived discontinuity, 38
 - in voice quality variation, 26–27
- MFCC. *See* Mel-frequency cepstral coefficient (MFCC)
- Minimization ranges, 4, 4*f*
- Model(s)
 - acoustic source, 68–69
 - articulatory. *See* Articulatory modeling
 - Coker, 67, 67*f*
 - contour control. *See* Contour control model
 - hidden Markov. *See* Hidden Markov model (HMM)
 - ignorance, 110
 - mechanical, 4
 - n*-gram. *See* *N*-gram model
 - phoneme-duration-control. *See* Phoneme-duration-control model
 - spectrum, 137
 - terminal analog, 64
 - voice source, 68
- Multiband resynthesis overlap add (MBROLA). *See also* TP-MBROLA
 - characteristics, 8–9
 - database preprocessing, 10
 - quality improvement, 9
- Multilingual speech synthesis system.
 - See also* Japanese text to speech synthesis
 - overview, 135–137, 136*f*
 - state output vector, 137
 - synthesis, 139–140, 140*f*
 - training, 137–139
- Multiple centroid analysis (MCA)
 - coefficients
 - definition, 43
 - join costs using, 54*t*, 55*t*, 57*t*
 - weighted sums of, 56–57, 57*t*, 58*t*
- Multispace probability distribution, 140–143, 142*f*
- N**
- N*-gram model
 - in automatic speech recognition, 112–113
 - for language generation, 120
 - for text normalization, 119–120
- Natural speech units, 78–79
- Nonuniform unit synthesis. *See also* Unit selection synthesis
 - background, 8–9
 - evaluation, 12–14
 - TP-MBROLA, 9–12, 10*f*, 11*f*
- O**
- OGIresLPC algorithm, 102
- P**
- Parametric synthesis, 124–126
- Perceptual listening tests. *See* Preference tests
- Phoneme-duration-control model, 162–167, 163–165*t*, 166*f*, 167*f*
- Phonemes
 - classes of, 165*t*

- sizes of words and, 210*f*
 - Phonetic features, 41
 - Pitch contours, 92–94, 96–98
 - Pitch distances, 95–98, 96*t*
 - Pitch marking, 10*f*
 - Pitch modification. *See also* Pitch smoothing; Prosody
 - overview, 91, 105–106
 - perceptual experiment, 101–105
 - speech corpus analysis in, 92–100
 - text corpus analysis in, 100–101
 - Pitch perception, 207
 - Pitch smoothing. *See also* Pitch modification; Prosody
 - boundaries of, 3–4
 - calculating shifts in, 4–7, 6*f*, 7*f*
 - preference tests, 7–8
 - shift-only F0, 3–8
 - in TP-MBROLA, 12
 - Plutchik's wheel, of emotions, 177, 177*f*
 - Power spectrum, in voice quality variation, 27
 - Preference tests
 - consistency in, 49–50, 50*t*
 - of Festival speech synthesis system, 101–105
 - for minimizing pitch modification, 101–105
 - of perceived discontinuity, 47–49
 - of rVoice system, 47–49
 - for shift-only F0 smoothing, 7–8
 - of TP-MBROLA, 12–14, 14*f*
 - Prosody. *See also* Pitch modification; Pitch smoothing
 - in automatic speech recognition, 125–126
 - in expressive speech synthesis, 90, 183–185
 - in HMM-based speech synthesis, 155–156
 - join costs of, 41
 - quality of, 89
 - in speech corpus analysis, 92–94, 93*t*
 - in speech synthesis markup language, 241*t*
 - statistical models for prediction of, 121
 - unit selection synthesis of, 206–211
- Q**
- Quantification theory
 - for phoneme duration, 163–165*t*
 - of prosody generation, 157–158
- R**
- Raw concatenation, 13
 - Reynolds number, 80
 - Rule-based synthesis. *See also* Formant synthesis
 - definition, 76–77
 - for expressive speech generation, 234–235
 - rVoice
 - preference test, 47–49
- S**
- Sadness, in expressive speech synthesis, 227, 227*t*
 - Search algorithms, 114–117, 115*f*, 117*f*
 - Shift values, 4–7, 6*f*, 7*f*
 - Signal processing, in automatic speech recognition, 122–123
 - Simple scheme, speech corpus analysis, 92–93, 93*t*
 - Sinusoidal modeling, 233–234
 - Smoothing techniques
 - for discontinuity removal, 2
 - shift-only F0, 3–8
 - Social constructivist theory, of emotions, 176–177
 - Speaker adaptation, in automatic speech recognition, 113, 119
 - Speaker recognition, 19–20. *See also* Voice quality variation
 - Spectral distance, 42–44. *See also* Distance measures
 - objective, 42*f*
 - parameterization, 43–44
 - Spectral measures, of join costs, 37–39

- Spectral tilt, in voice quality variation, 25–26
- Spectrum modeling, 137
- Speech corpus, types of, 95, 96*t*, 97*t*
- Speech corpus analysis
for minimizing pitch modification, 92–100
prosodic factors, 92–94, 93*t*
- Speech mimic, 79–82, 80*f*, 82*f*
- Speech-parameter generation
algorithm, 143–147, 147*f*
- Speech-rate-variable method, 168–170
- Speech recognition. *See* Automatic speech recognition (ASR)
- Speech synthesis, history of, 64–65
- Speech synthesis markup language (SSML)
examples, 240–242
extensibility, 239–240
multilevel structure of, 237–239, 238*f*
prosody elements of, 241*t*
specification in, 237
during synthesis, 239
- SSML. *See* Speech synthesis markup language (SSML)
- State-duration probability, 146
- T**
- Target cost
join cost and, 40–41
of unit selection synthesis, 208
- Text corpus analysis, for minimizing pitch modification, 100–101
- Text to speech synthesis
basic system for, 179, 180*f*
concatenative. *See* Concatenative synthesis
formant method. *See* Formant synthesis
history of, 64–65
prototype articulatory system for, 82–83
unit selection system. *See* Unit selection synthesis
- Time intervals, voice quality variation
and, 28
- ToBi. *See* Tones and break indices (ToBI)
- Tones and break indices (ToBI)
in expressive speech synthesis, 230–232
in speech corpus analysis, 92
in speech synthesis markup language, 240
- TP-MBROLA. *See also* Multiband resynthesis overlap add (MBROLA)
algorithms in, 11–12
database preprocessing, 9–11, 11*f*
evaluation, 12–14
pitch smoothing in, 12
vs. MBROLA, 15
- Transfer function, of tube, 71–73, 72*f*
- TTS. *See* Text to speech synthesis
- Tube, transfer function of, 71–73, 72*f*
- Two-mass model, 68
- U**
- Unit selection synthesis. *See also* Nonuniform unit synthesis
comparative evaluation, 211–217, 213*f*, 213*t*, 215*t*, 217*t*
databases for, 206–207
in HMM-based approach, 148–150, 149*f*, 150*f*
join cost calculation, 35–37
as search problem, 114–115, 115*f*
selection system architecture, 207–211, 208–210*f*, 211*t*
tuning selection, 211
- V**
- Viterbi alignment procedure, 117
- Vocal cord vibration, 73–74
- Vocal tract acoustics, 65–66, 67*f*
- Voice conversion, 119
- Voice mimic. *See* Speech mimic
- Voice quality variation, 19–20
acoustic measures of, 25–28, 26*t*, 32
factors in, 23–25, 24*f*, 25*f*

perceptual experiment, 20–23, 21*f*
prediction of, 28–31, 29*f*, 30*t*, 31*f*
Voice source models, 68
Voice transformation, 119

W

Weighted distances, 46

Weighted finite-state transducer
(WFST), 116

Weighted time average model
(WTAM), 207

WFST. *See* Weighted finite-state
transducer (WFST)

WTAM. *See* Weighted time average
model (WTAM)