Subject Index

3-gram language model 307, 416

A* search algorithm 301, 308
acoustic
– feature vectors 289
– model 213, 279, 301, 331
– score caching 192
AdaBoost.MH 132
AGILE 34, 38
ajka 156, 162
Alfanum system 237
algorithm
– A* 301
– AdaBoost 132
– EM 338
– evaluation 437
– expectation-maximization 143, 202, 226
– for valency frame generation 149
– forward-backward 277, 415
– Inside-Outside 53
– naïve Bayes 99
– of automatic term clustering 151
– of clustering 157
– Porter’s 70
– TD-PSOLA 237
– Viterbi 3, 121, 277, 306
– Viterbi-I 9
– Viterbi-L 4, 8
ALISP 269
AltaVista 141
AMDF pitch-extraction 240
analysis
– lexical of Classical Chinese 117
– MelBank 299
– of sense 299
– of sppech signal 299
– of text corpora 268
– PCBF 299
– PLP 299
– syntactico-semantic 89
annotation at word-level 181
Arabic 11
ARTIC 173
artificial neural networks 181
artificial talking head 445
ASR 253, 277, 301, 341
ATIS 461
audio collection 347
audio-visual speech recognition 205
Austrian 26, 341
automatic
– dialogue systems 357
– language independent speech processing 269
– speech recognition 253
– transcription 253

backward Viterbi algorithm 306
Baldi engine 446
Basque 41
Basque corpus 397
Basurde 389
Basurde corpus 384, 390
Baum-Welch re-estimation 207
Bayes 99
– algorithm 99
– classifier 131, 227
Bayesian
– inference framework 337
– networks 365
BDI 405
bigram
– invalid 19
– language model 446
– matrix 193
– model 197, 343
bilingual French-English corpus 143
bilingual Springer vector model 111
bimodal speech recognizer 208
biphones 331
blind Web users 453
BNC 110, 112
Braille coding of texts 453
BREF French database 270, 307
BREF80 French database 288, 307
British National Corpus 110, 111
Brown corpus 115
BRULEX 278
Bulgarian 34
bunsetsu 28

C+ 133
C++ 41, 262, 406
Canadian 109
CART 215, 218, 261, 263
Catalan 41
categorization of web sites 65
CELEX 278
CER 219
CGN 277
Chinese 115
– Classical 115
– Modern 115
– morphology 115
– syntax 115
Chomsky 50, 55
CKD 246
Classical Chinese 115
classification
– of phrases 29
– of prosody 357
classification and regression tree 215
CLEF collection 60
CLIPS 405, 407
CLIR 139
clustering
– algorithm 151
– of context vectors 157
– phoneme-level 173
COMLEX 278
communication-oriented representation 34
comparison of parsing 461
compounding language 343
compression rate 30
concatenative
– text-to-speech synthesizer 237
conditional
– maximum entropy model 131, 132
– probability 50
– probability distributions 365
– responses 317
cone-shaped kernel distribution 245
confidence
– measure 181, 213
– of word recognition 181
content words 215
context pattern 152
context-independent phoneme set 322
contextual
– lexical semantics 369
– similarity 151, 368
continuous speech
– recognition system Speeral 301
continuous speech recognition 189, 197, 277, 309, 321, 329, 445
conversational system 413
Corga 5
corpus
– annotated 416
– annotation 19
– based similarity measure 151
– Basque 397
– Basurde 384, 390
– BREF 270
– British National 111
– Brown 115
– Corga 5
– DARPA-TIMIT 294
– DDAC2000 215
– dialogue 421
– DSO 131
– Dutch spoken 277, 281
– ECONOMIC 266
– error detection 19
– FAIRYTALE 266
– Hansard 109
– ILSP/ELEYTHEROTYPIA 369
– LN 257
– LOB 115
– NEGRA 19
– of German broadcast news 344
– of Kyoto University 28
– of patents 65
– of Reuter 65
– oriented morphological analyzer 422
– PoS-tagged 19
– RECIPE 265
– representativity 19
– Slovenian 329, 334
– Susanne 49, 53
– UWB_B02 257
– Verbmobil appointment-scheduling 317
– WEATHER 266
corpus-based speech synthesis 265
CRIM system 307
CRIN system 307
cross validation 101, 121, 136
cross-language information retrieval 139
cross-language retrieval 57
cross-language searching 59
CT 461

Czech 19, 34, 57, 105, 155, 173, 189, 197, 270, 446
– continuous speech recognition 197
– grammar checker 87
– isolated words 321
– language 189, 253
– morphology 148
– National Corpus 157
– nouns 159
– phonemes 191
– speech recognition 189
– speech synthesizer Demosthenes 266
– valency lexicon 147
– verbs 147
– WSD 155

DAMSL 319
DARPA-TIMIT database 294
data-driven approach to word splitting 342
database
– BRULEX 278
– CELEX 278
– COMLEX 278
– Fonilex of Flemish pronunciation 278
– M2VTS 206
– of speech segments 173
– Polyphone Dutch 205
– SNABI 333
– SpeechDat 333
– SpeechDat - Car 229
– SpeechDat-E 321
DDAC2000 corpus 215
decision tree 160, 181
decision tree learning 165
declination patterns 159
Demosthenes speech synthesizer 266
dependency
– analyzer 81
– grammar 81
– structure 29
– tree 81
detection
– of non-speech 309
– of speech 309
dialogue 413, 447
– act labelers 381
– acts 317, 367
– annotation 319
– constraints 389
– controller 377
– corpora 365
– guided 373
– information system 421
– logic 437
– manager 370, 394, 417, 424, 459
– modelling 381
– modelling language 441
– module 453
– planning 429
– representation 407
– spoken 449
– strategy 452
– – rule-based 381
– structure 390
– system 357, 377, 381, 437, 449
– – SmartKom 357
– system Basurde for Spanish 389
– system framework Elvira 437
– system in Spanish 421
– systems 365, 429
– telemedicine system 449
– understanding 389
Dillo 453
disambiguation
– word domain 131
– word sense 131
– word-sense 41, 107
discourse 73, 365, 429
– model 430
– quality improvement 365
discriminant analysis 325
distributed speech recognition 229
distribution Gaussian 321
DL ACM 101
DMOZ 66, 141
document
– categorisation 151
– selection 99
DRAFTER 34
DRT 323
DSO corpus 131
DSR 229
DTMF tone detection 414
Dutch 205, 213, 281
– Polyphone database 206
– spoken corpus 277
dynamic programming 191

ECGI 393
EIFL Direct 101

EIS 454
Elvira 437
EM 226, 337
EM algorithm 143, 310
Emacs 262
English 19, 41, 57, 109, 111, 139, 197, 341, 358, 397, 413
entropy
– cross 182
– maximum 131
– normalized 182
Ergane 61
error detection in corpus 19
EuroWordNet 41, 142, 147, 157
expectation-maximization algorithm 202, 226

facial
– animation 299
– expression 357
FAR 290
feature
– extraction 245, 333
– extraction technique 325
– reduction 227
– selection 224, 227
– vector 68, 310
– vectors 208, 289
feneme-based system 178
Festival 262
Festival architecture 261
FFT 246, 311, 328, 338
Finish 230
finite state diagrams 442
Finnish 413
Flemish 278, 281
Fonilex 278
forward-backward algorithm 277, 280
forward-backward segmentation 280
FPR 214
free phone recognition 214
French 139, 143, 270, 301, 397, 416
– BREF80 database 288
– speech corpus 288
– speech corpus BREF 270
frequency signal representation 245
FRR 290
FSD 442
Fukunaga-Koontz 225
functional generative description 147

g2p system 278
Galena 4
Galician 3
Gaussian distribution 321
Gaussian RBF kernel 328
GCFG 49
GEMINI 261, 337
GEMINI project 365
generalization of bigrams 19
German 19, 109, 173, 341, 358, 398
– corpus NEGRA 19
– orthography 342
– text-to-speech 177
Germna 413
gesture 357
GhostWriter 34
Gibbsian CFG 49
GISE 374
GIWeb 373
GNBD 124
GNU general public license 453
GoDiS 320
grammar
– Nigel 375
– of Chinese 117
– polynomial tree substitution 49
– stochastic context-free 49
– stochastic regular 389
– with errors 83
grammatical inference 389
Greece 366
Greek 261, 263, 369, 413
– speech synthesis 261

Hacoust sounding function 306
Hansard corpus 109, 143
harmonic noise model 269, 273
Hašek 447
HDddm parsing 464
Hidden Markov Model 4, 206
Hidden Markov Toolkit 205
hierarchical categories 139
HMM 4, 116, 173, 189, 205, 215, 221, 245, 279, 307, 321, 343, 383, 389, 416, 456
– state-clustered 177
– subphonetic 415
– triphone 176
HNM 173, 273
HNM synthesis 269
Homer Web browser 454

HTK 205, 250, 322
HTK toolkit 205, 250, 321, 333
HTML 443
HTML/XML 453
Hungarian 328
– sign language 73

ID3 278
ILSP/ELEYTHEROTYPIA 369
InBASE 33, 38
Infomap 107
information retrieval 47, 151
– cross language 139
– dialogues 437
information system 33
information theoretic similarity measure 369
InQuery 61
interlingual index 43
Itakura-Saito measures 340
Italian 230, 416

J++ 39
Japanese 27

k-means algorithm 157
K211d 250
Kalliope Web portal 454
Karhunen-Loeve expansion 225
Katz's discounting 259
kernel distribution 245
kernel springy discriminant analysis 326
kernel support vector machine 285
keyword spotting 285
Kittler-Young 225
KL-ONE 34
knowledge accumulation 89
knowledge acquisition 151
knowledge assimilation 94
knowledge representation 89
KPML 34, 38
KSDA 326

LAFORIA system 307
Lagrange functional 286
Lagrangian 286
language
– Arabic 11
– Basque 397
– Bulgarian 34
– Chinese 115
– Czech 19, 34
– Dutch 205
– English 19, 358
– Flemish 278
– generation 33
– German 19, 341, 358
– Greek 261
– independent Q-representation 39
– independent speech processing 269
– inflected 397, 398
– Japanese 27
– model 416
– Nenets 347
– Nganasan 347
– Russian 34, 347
– Serbian 237
– Slavic 34, 238
– Slovenian 329
– Spanish 123
– understanding 34
large vocabulary continuous speech recognition 277
latent semantic analysis 110
lattice based speech recognition 301
LDC 254
Le Monde 307
learning
– instance-based 131
– with decision tree 165
– with multilayered perceptron 165
lemmatization 447
lexical
– derivation 147
– meaning 147
– parser 123
– similarity 151
– stress marks 165
lexico-syntactic patterns 152
likelihood distance 213
LIMSI ARISE 421
LIMSI base system 307
Lin's metric 369
linear interpolation smoothing 201
Linguistic Data Consortium 254
linguistic ontology 373
Link 109
lip reading 205
LLR 216
LM score 416
LN corpus 257

LOB corpus 115
Lochbaum's Rgraph 430
LP-PSOLA 173
LPC 270, 299
LPC synthesis 261, 262, 270
LPCC 270
LVCS recognition 331
LVCSR 258, 277

M2VTS database 206
machine translation 47
MACO+ 422, 423
MALACH project 57, 253
maximum entropy 116
maximum entropy model 131
maximum likelihood 279
maximum likelihood estimate 200
Mbrola 261, 262
meaning
– analysis 107
– of sentence 127
– opposite 221
– representation 91
measure of nonprojectivity 82
medical documents 100
MEDLINE 101
Mel filter 235
MelBank 299
MeteoBayes 371
METEONEWS 366
metric
– Lin's 369
– weighted Jaccard similarity 369
MFCC 289, 299, 310
MFCC_0_D_A 322
MFCC_D_A 313
MGGI 393
minimal loss 369
minimum mean square error estimation 337
MINIPAR 43
mixture of Gaussians 337
MLE 200
MMSE 337
model
– *n*-gram 199, 341
– belief desire intention 405
– bigram 197, 446
– four-gram 185, 330
– harmonic+noise 173
– hidden Markov acoustic 245
– mumble 221
– phonetically tied-mixture 343
– probabilistic 381
– state-clustered tied-mixture 343
– stochastic 389
– tri-gram 336
– triphone acoustic 336
– triphones 331
Modern Chinese 115
monophones 331
Moore 461
morpheme prediction 398
morphological analyzer 422
morphology 148
MULTEXT-East 165
multi-modal
– access to information systems 413
– dialogue system 357
multilingual
– access 253
– access to information systems 413
– environment 41
– language generation 33
Multilingual Access to Large Spoken Archives project 253
multimodal interfaces 205
MultiNet 89
mumble model 221
mutual information 369

n-gram model 201, 259, 341, 450
n-gram tagger 20
natural language analysis 89, 447
natural language guided dialogue 373
natural language synthesis 447
NCC1 310
negotiation in dialogue 317
NEGRA corpus 19
Nenets 347
Nenets Multimedia Phrase-book 351
Nganasan 347, 353
Nganasan Audio Dictionary 350
Nigel grammar 375
NIST 207
noise 337
– estimation 229
– reduction 333
NOISEX-92 database 339

object domain model 35

ODM 35
OHSUMED collection 65
ontological hierarchies 108
ontologies 373
ontology
– of processes 34
– of things 34
orthography 278

parallel
– corpus French/English 109
– language model 418
– monolingual systems 414
parser
– Arabic 11
– efficiency 465
– MINIPAR 43
– PCFG for Classical Chinese 121
– shallow 422
– statistical 414
– SUPP 42
parsing 11, 49, 414, 461
part-of-speech tagging 19, 115
patent corpuses 65
path distance 213
PCBF 299
PCFG 49
PDT 150
PDTB 462
phoneme
– classes 278
– model 209, 322
– recognizer 325
– segmentation 293
– set 322
phoneme-level clustering 173
phonetic
– dictionary 180
– transcription 278
– transcriptions 277
phonetically tied-mixture model 343
phonetization 238
phonological
– awareness 325
– teaching system 325
phrase
– boundary 29
– dependency 27
– significance 27, 30
PLP 299, 310
PLP_D_A 313
Porter 144
Porter’s suffix removal algorithm 70
PoS 29, 42, 110, 115, 168, 360, 422, 423
– tagger 168
– tagging 3
– tagging for Chinese 115
Prague Dependency Treebank 61, 150
probabilistic
– context-free grammar 49
– context-free parsing algorithm 49
– dialogue modelling 381
probabilistic parsing techniques 461
probability
– a-posteriori 184
project MALACH 253
Prolog 42
pronunciation database 278
prosodic
– classification 357
– – rates 363
– feature vector 357
– features 361
prosody 270, 439
– classification 357
– generation 180
– generator Serbian 237
– information 270
– modeling 264, 439
– modelling of Slovene speech 459
– of Nenets 348
prune threshold 199
PSOLA 173
PT 461
pulse code modulation 262

Q-representation 34, 39
quadratic superposition principle 246
query
– expansion techniques 139
– reformulation 139
– representation 33
querying system 123
question answering 47, 131
question answering systems 89

radial basis function 285
RASTA 310
RASTA-PLP 299
RBF 290

real-time continuous speech recognition 189
recipe graph 430
recognition
– error 221
– error at word-level 181
– rate 182
– system 199
recognizer SPHINX 323
recursive transition networks 450
rejection technique 221
RELAX 423
retrieval cross-language 57
Reuters corpus 65
RFODG 82
Rgraph 430
RMR 310
robust speech recognition 229
ROC 218
Rough'n'Ready suite of technologies 343
rule-based dialogue strategies 381
Russia 347
Russian 34, 165
Russian Federation 347

salient words 68
SAM 333
Samoyedic 353
SAPI5 design specification 408
SCFG 49
Scheme 262
scoring method 181
screen reader 459
searching
– cross-language 59
– thesaurus-based 59
See5 168
segmental features 268
segmentation
– ambiguous 5
– methods 293
– of speech 277
segmentation of speech into phonemes 293
semantic
– analyzer 123, 127
– graph 110
– networks 89
– similarity 43, 109, 368
semantic discrimination 155
Semcor.LazyBoosting 132
SENSEVAL-2 132
sentence
– compaction 27
– compression 27
sentence boundary detection 132
sentence compression 27, 30
Serbian 237
short time Fourier trnasform 338
short-time spectral attenuation 337
signal representation 245
signal-to-noise ratio 229
similarity
– contextual 151
– functional 151
– lexical 151
– matrix 151
– measure 151
– semantic 109
SIRIDUS 317, 320
SISCO 124
Slavic languages 329
Slovene 247, 453
– dialogue module 453
– lexicon 165
– speech synthesis 459
Slovene speech database 250
Slovenia 454
Slovenian 168, 329, 334
– language 168
– PoS tagger 168
– text corpus 334
SMADA project 213
SMART 67, 144
SmartKom 357
SmartKom project 205
SNABI 333
SNABI speech database 331
SNR 229
Spanish 3, 41, 123, 230, 373, 384, 398, 416, 423
– dialogue corpus 421
– dialogue system Basurde 389
– language 389
– lexical parser 123
– querying system 123
– railway information dialogue system 421
– semantic analyzer 123
– syntax checker 123
– WordNet 43
speaker clustering 269
speaker-independent

– speech processing 269
– speech recognition 189, 221
spectral magnitude 337
Spectral subtraction 229
spectrogram 246
speech
– archive 57
– coding 269
– corpora 57
– corpus 177, 237
– encoding methods 293
– feature extraction 245
– processing 445
– recognition 49, 181, 197, 205, 213, 221, 229, 242, 253, 277, 309, 321, 329, 333, 341, 445
– – continuous 189
– – real-time 189
– recognition errors 213
– recognition system Speeral 301
– recognizer 234
– retrieval 57
– segment database 173
– segmentation 277, 293
– segmentation methods 277
– signal representation 245
– speaker-independent processing 269
– spectral characteristics 337
– synthesis 173, 237, 261, 269, 459
– – concatenative 173
– synthesis corpus-based 265
– synthesizer Demosthenes 266
– synthetic sample 179
– technology applications 405
– user interface 407
Speech Interfacing in the Wireless Automotive Messaging Pilot project 405
speech recognition 285
SpeechDat 229, 333
SpeechDat(II) 333
SpeechDat-E 321
SpeechVis software 239
Speeral continuous speech recognition system 301
SPHINX 323
SPINE 321, 323
SPL 38
splitting of compounds 342
spoken
– archives 253
– corpus 281
– Czech 173, 197
– dialogue 449
– dialogue system 421
– Dutch corpus 277
– German 173
Springer Link 109
SQL 128
SRILM toolkit 259
SSJČ dictionary 157
START 33
state-clustered tied-mixture model 343
statistical decision making 365
statistical inference techniques 381
STE 310
stemming 139
STFT 338
STI 310
stochastic CFG 49
STSA 337
STZCR 310
SUI 407
supervised learning 131
SUPP 42
support vector machine 285, 328
Susanne corpus 49, 53
Švejk 447
SVM 288
SWAMP 405, 406
Swedish 398
Switchboard 277
Sylvia-NLQ 124
synset 42, 147
syntactic
– analysis 49, 51
– derivation 147
– patterns 151
syntax
– analysis 81
– checker 123
– recognition 81
synthesis triphone-based 178
synthesizer Mbrola 262
system
– Alfanum 237
– conversational 413
– Elvira 437

t-test 136
TABA 421
TACAT 422

tagger
– relaxation labelling 423
– rule-based 115
– statistical 20
– transformation-based 115
– trigram-based 20
tagging
– PoS 3, 19, 115, 132
– sense 158
Taiwan corpus 119
TD-PSOLA 237
TEA text analyzer 99
telemedicine system 449
telephone dialogue system 384
telephony system 418
term
– clustering 151
– extraction 151
– similarity 151
testing of NLP systems 437
text
– analyzer TEA 99
– categorization 65
– classification 99
– documents 99
– filtering 41, 99
– summarisation 27
– unstructured 99
text-to-speech
– synthesis 445
– synthesizer 237
texts bilingual 107
TFD 246
the Netherlands 206, 353
thesauri 139
time signal representation 245
time-frequency distribution 246
TIMIT 277
tokenization 3
toolkit
– HTK 205, 250
– SRILM 259
topic
– detection 59
– tracking 59
topical importance 27
TRAINS 421
Transcriber tool 342
transcription
– automatic 253
transcription BN systems 344
transformation-based tagger 115
translation term-by-term 139
TREC 143
TreeView program 86
trellis 4
trigram model 330
TRINDI 320
triphone acoustic model 336
triphones 331, 343
TTS 173, 237, 261, 277, 407, 414, 445, 453
– Greek 261
TTS system 168

UK English 416
unification of layer attributes 89
Upper Model 34
US English 270, 294, 416
utterance
– selection 268
– verification 213
UWB_B02 corpus 257

valency
– frame 84
– lexicon of Czech 147
vector model 108
vector quantization 269
vector space model 67
vector spaces 108
Verbmobil 184, 363
Verbmobil appointment-scheduling corpus 317
Verbmobil database 341
VHF 253
ViaVoice 415
viseme 209
– class 210
– model 210
Visual History Foundation 253
visualization techniques 107
visually impaired Web users 453
Viterbi 277
– algorithm 3, 4, 116, 191, 279, 306, 394
– – iterative 3
– – one-pass 3
– algorithm on lattices 4
– alignment 207
– decoder 207
– recognition 270
– search 189, 191

Viterbi search 176
Viterbi segmentation 283
Viterbi-I 9
Viterbi-L 4
Viterbi-N 8
VLBR 269
voice chat 445
voice-driven dialogue manager 459
VoiceXML 437
VoiceXML interpreter 441
VQ based speaker clustering 269

WAM 405
WDD 131
Web Access Initiative 453
Web of Science 101
web sites categorization 65
weighted Jaccard similarity metric 369
Wiener filtering 229
Wigener-Ville distribution 246
Witten-Bell discounting 201
Witten-Bell discounting scheme 192
Witten-Bell smoothing technique 194
Wizard of Oz 357, 366, 384, 390, 394, 421
word confidence estimation 181
word decomposition 329
word order complexity 81
word-sense
– classifiers 133
– disambiguation 41, 107, 131, 155
WordNet 41, 131, 142, 155
WSD 107, 131, 155

Yahoo 66, 70, 141

ZDSSS 454
ZER 310
Zhao-Atlas-Marks distribution 245