Janet M. Baker, Li Deng,
James Glass, Sanjeev Khudanpur,
Chin-Hui Lee, Nelson Morgan, and
Douglas O'Shaughnessy

Research Developments and Directions in Speech Recognition and Understanding, Part 1

To advance research, it is important to identify promising future research directions, especially those that have not been adequately pursued or funded in the past. The working group producing this article was charged to elicit from the human language technology (HLT) community a set of well-considered directions or rich areas for future research that could lead to major paradigm shifts in the field of automatic speech recognition (ASR) and understanding. ASR has been an area of great interest and activity to the signal processing and HLT communities over the past several decades. As a first step, this group reviewed major developments in the field and the circumstances that led to their success and then focused on areas it deemed especially fertile for future research. Part 1 of this article will focus on historically significant developments in the ASR area, including several major research efforts that were guided by different funding agencies, and suggest general areas in which to focus research. Part 2 (to appear in the next issue) will explore in more detail several new avenues holding promise for substantial improvements in ASR performance. These entail cross-disciplinary research and specific approaches to address three-to-five-year grand challenges aimed at stimulating advanced research by dealing with realistic tasks of broad interest.

SIGNIFICANT DEVELOPMENTS IN SPEECH RECOGNITION AND UNDERSTANDING

The period since the mid-1970s has witnessed the multidisciplinary field of ASR proceed from its infancy to its coming of age and into a quickly growing number of practical applications and commercial markets. Despite its many achievements, however, ASR still remains far from being a solved problem. As in the past, we expect that further research and development will enable us to create increasingly powerful systems, deployable on a worldwide basis.

This section briefly reviews highlights of major developments in ASR in five areas: infrastructure, knowledge representation, models and algorithms, search, and metadata. Broader and deeper discussions of these areas can be found in [12], [16], [19], [23], [24], [27], [32], [33], [41], [42], and [47]. Readers can also consult the following Web sites: the IEEE History Center's Automatic Speech Synthesis and Recognition section and the Saras Institute's History of Speech and Language Technology Project at http://www.sarasinstitute.org.

INFRASTRUCTURE

Moore's Law observes long-term progress in computer development and predicts doubling the amount of computation achievable for a given cost every 12 to 18 months, as well as a comparably shrinking cost of memory. These developments have been instrumental in enabling ASR researchers to run increasingly complex algorithms in sufficiently short time frames (e.g., meaningful experiments that can be done in less than a day) to make great progress since 1975.

The availability of common speech corpora for speech training, development, and evaluation has been critical, allowing the creation of complex systems of ever increasing capabilities. Speech is a highly variable signal, characterized by many parameters, and thus large corpora are critical in modeling it well enough for automated systems to achieve proficiency. Over the years, these corpora have been created, annotated, and distributed to the worldwide community by the National Institute of Science and Technology (NIST), the Linguistic Data Consortium (LDC), and other organizations. The character of the recorded speech has progressed from limited, constrained speech materials to huge amounts of progressively more realistic, spontaneous speech. The development and adoption of rigorous benchmark evaluations and standards, nurtured by NIST and others, have been critical in developing increasingly powerful and capable systems. Many labs and researchers have benefited from the availability of common research tools such as Carnegie-Mellon University Language Model (CMU LM) toolkit, Hidden Markov Model Toolkit (HTK), Sphinx, and Stanford Research Institute Language Modeling (SRILM). Extensive research support combined with workshops, task definitions, and system evaluations sponsored by the U.S. Department of Defense Advanced Research Projects Agency (DARPA) and others have been essential to today's system developments.

KNOWLEDGE REPRESENTATION

Major advances in speech signal representations have included perceptually motivated mel-frequency cepstral coefficients (MFCC) [10], [29] and perceptual linear prediction (PLP) coefficients [21], as well as normalizations via cepstral mean subtraction (CMS) [16], [44], relative spectral (RASTA) filtering [20], and vocal tract length normalization (VTLN) [13]. Architecturally, the most important

Digital Object Identifier 10.1109/MSP.2009.932166

development has been searchable unified graph representations that allow multiple sources of knowledge to be incorporated into a common probabilistic framework. Noncompositional methods include multiple speech streams, multiple probability estimators, multiple recognition systems combined at the hypothesis level (e.g., Recognition Output Voting Error Reduction (ROVER) [15]), and multipass systems with increasing constraints (bigram versus four-gram, within word dependencies versus cross-word, and so on). More recently, the use of multiple algorithms, applied both in parallel and sequentially, has proven fruitful, as have feature-based transformations such as heteroscedastic linear discriminant analysis (HLDA) [31], feature-space minimum phone error (fMPE) [40], and neural net-based features [22].

MODELS AND ALGORITHMS

The most significant paradigm shift for speech-recognition progress has been the introduction of statistical methods, especially stochastic processing with hidden Markov models (HMMs) [3], [25] in the early 1970s [38]. More than 30 years later, this methodology still predominates. A number of models and algorithms have been efficiently incorporated within this framework. The expectation-maximization (EM) algorithm [11] and the forward-backward or Baum-Welch algorithm [4] have been the principal means by which the HMMs are trained from data. Despite their simplicity, *N*-gram language models have proved remarkably powerful and resilient. Decision trees [8] have been widely used to categorize sets of features, such as pronunciations from training data. Statistical discriminative training techniques are typically based on utilizing maximum mutual information (MMI) and the minimum-error model parameters. Deterministic approaches include corrective training [1] and some neural network techniques [5], [35].

Adaptation is vital to accommodating a wide range of variable conditions for the channel, environment, speaker, vocabulary, topic domain, and so on. Popular techniques include maximum a posteriori probability (MAP) estimation [17], [38], [51], maximum likelihood linear regression (MLLR) [34], and eigenvoices [30]. Training can take place on the basis of small amounts of data from new tasks or domains that provide additional training material, as well as "one-shot" learning or "unsupervised" training at test time.

SEARCH

Key decoding or search strategies, originally developed in nonspeech applications, have focused on stack decoding (A* search) [26] and Viterbi or *N*-best search [50]. Derived from communications and information theory, stack decoding was subsequently applied to speech-recognition systems [25], [37]. Viterbi search, broadly applied to search alternative hypotheses, derives from dynamic programming in the 1950s [6] and was subsequently used in speech applications from the 1960s to the 1980s and beyond, from Russia and Japan to the United States and Europe [3], [7], [9], [36], [45], [46], [48], [49].

METADATA

Automatic determination for sentence and speaker segmentation as well as punctuation has become a key feature in some processing systems. Starting in the early 1990s, audio indexing and mining have enabled high-performance automatic topic detection and tracking, as well as applications for language and speaker identification [18].

GRAND CHALLENGES: MAJOR POTENTIAL PROGRAMS OF RESEARCH

Grand challenges are what our group calls ambitious but achievable three-to five-year research program initiatives that will significantly advance the state of the art in speech recognition and understanding. Previous grand challenges sponsored by national and international initiatives, agencies, and other groups have largely been responsible for today's substantial achievements in ASR and its application capabilities. Six such potential programs are described below. Each proposed program has defined, measurable goals and comprises a complex of important capabilities that should substantially advance the field and enable significant applications. These are rich task domains that could enable progress in several promising research areas at a variety of levels. As noted below, each of these program initiatives could also benefit from, or provide benefit to, multidisciplinary or cross-area research approaches.

EVERYDAY AUDIO

This is a term that represents a wide range of speech, speaker, channel, and environmental conditions that people typically encounter and routinely adapt to in responding and recognizing speech signals. Currently, ASR systems deliver significantly degraded performance when they encounter audio signals that differ from the limited conditions under which they were originally developed and trained. This is true in many cases even if the differences are slight.

This focused research area would concentrate on creating and developing systems that would be much more robust against variability and shifts in acoustic environments, reverberation, external noise sources, communication channels (e.g., far-field microphones, cellular phones), speaker characteristics (e.g., speaking style, nonnative accents, emotional state), and language characteristics (e.g., formal/informal styles, dialects, vocabulary, topic domain). New techniques and architectures are proposed to enable exploring these critical issues in environments as diverse as meeting-room presentations and unstructured conversations. A primary focus would be exploring alternatives for automatically adapting to changing conditions in multiple dimensions, even simultaneously. The goal is to deliver accurate and useful speech transcripts automatically under many more environments and diverse circumstances than is now possible, thereby enabling many more applications. This challenging problem can productively draw on expertise and knowledge from related disciplines, including natural-language processing, information retrieval, and cognitive science.

RAPID PORTABILITY TO EMERGING LANGUAGES

Today's state-of-the-art ASR systems deliver top performance by building complex acoustic and language models using a large collection of domain-specific speech and text examples. For many languages, this set of language resources is often not readily available. The goal of this research program is to create spoken-language technologies that are rapidly portable. To prepare for rapid development of such spoken-language systems, a new paradigm is needed to study speech and acoustic units that are more language-universal than language-specific phones. Three specific research issues need to be addressed: 1) cross-language acoustic modeling of speech and acoustic units for a new target language, 2) cross-lingual lexical modeling of word pronunciations for new language, and 3) cross-lingual language modeling. By exploring correlation between these emerging languages and well-studied languages, cross-language features, such as language clustering and universal acoustic modeling, could be utilized to facilitate rapid adaptation of acoustic and language models. Bootstrapping techniques are also keys to building preliminary systems from a small amount of labeled utterances first, using these systems to label more utterance examples in an unsupervised manner, incorporating new labeled data into the label set, and iterating to improve the systems until they reach a performance level comparable with today's high-accuracy systems.

Many of the research results here could be extended to designing machine translation, natural-language processing, and information-retrieval systems for emerging languages. To anticipate this growing need, some language resources and infrastructures need to be established to enable rapid portability exercises. Research is also needed to study the minimum amount of supervised label information required to create a reasonable system for bootstrapping purposes.

SELF-ADAPTIVE LANGUAGE CAPABILITIES

State-of-the-art systems for speech transcription, speaker verification, and language identification are all based on statistical models estimated from labeled training data, such as transcribed speech, and from human-supplied knowledge, such as pronunciation dictionaries. Such built-in knowledge often becomes obsolete fairly quickly after a system is deployed in a real-world application, and significant and recurring human intervention in the form of retraining is needed to sustain the utility of the system. This is in sharp contrast with the speech facility in humans, which is constantly updated over a lifetime, routinely acquiring new vocabulary items and idiomatic expressions, as well as deftly handling previously unseen nonnative accents and regional dialects of a language. In particular, humans exhibit a remarkable aptitude for learning the sublanguage of a new domain or application without explicit supervision.

The goal of this research program is to create self-adaptive (or self-learning) speech technology. There is a need for learning at all levels of speech and language processing to cope with changing environments, nonspeech sounds, speakers, pronunciations, dialects, accents, words, meanings, and topics, to name but a few sources of variation over the lifetime of a deployed system. Like its human counterpart, the system would engage in automatic pattern discovery, active learning, and adaptation. Research in this area must address both the learning of new models and the integration of such models into preexisting knowledge sources. Thus, an important aspect of learning is being able to discern when something has been learned and how to apply the result. Learning from multiple concurrent modalities, e.g., new text and video, may also be necessary. For instance, an ASR system may encounter a new proper noun in its input speech and may need to examine contemporaneous text with matching context to determine the spelling of the name. Exploitation of unlabeled or partially labeled data would be necessary for such learning.

A motivation for investing in such research is the growing activity in the allied field of machine learning. Success in this endeavor would extend the lifetime of deployed systems and directly advance our ability to develop speech systems in new languages and domains without the onerous demands of labeled speech, essentially by creating systems that automatically learn and improve over time. This research would benefit from cross-fertilization with the fields of natural-language processing, information retrieval, and cognitive science.

DETECTION OF RARE, KEY EVENTS

Current ASR systems have difficulty in handling unexpected—and thus often the most information-rich—lexical items. This is especially problematic in speech that contains interjections or foreign or out-of-vocabulary words and in languages for which there is relatively little data with which to build the system's vocabulary and pronunciation lexicon. A common outcome in this situation is that high-value terms are overconfidently misrecognized as some other common and similar-sounding word. Yet such spoken events are crucial to tasks such as spoken term detection and information extraction from speech. Their accurate registration is therefore of vital importance.

The goal of this program is to create systems that reliably detect when they do not know a valid word. A clue to the occurrence of such error events is the mismatch between an analysis of a purely sensory signal unencumbered by prior knowledge, such as unconstrained phone recognition, and a word- or phrase-level hypothesis based on higher-level knowledge, often encoded in a language model. A key component of this research would therefore be the development of novel confidence measures and accurate models of uncertainty based on the discrepancy between sensory evidence and a priori beliefs. A natural sequel to detection of such events would be to transcribe them phonetically when the system is confident that its word hypothesis is unreliable and to devise error-correction schemes.

One immediate application that such detection would enable is subword (e.g., phonetic) indexing and search of speech regions where the system suspects the presence of errors. Phonetic transcription of the error-prone regions would also enable the development of the next generation of self-learning speech systems: the system may then be able to examine new texts to determine the identity of the unknown word. This research has natural synergy with natural-language processing and information-retrieval research.

COGNITION-DERIVED SPEECH AND LANGUAGE SYSTEMS

A key human cognitive characteristic is the ability to learn and adapt to new patterns and stimuli. The focus of this project would be to understand and emulate relevant human capabilities and to incorporate these strategies into automatic speech systems. Since it is not possible to predict and collect separate data for any and all types of speech, topic domains, and so on, it is important to enable automatic systems to learn and generalize even from single instances (episodic learning) or limited samples of data, so that new or changed signals (e.g., accented speech, noise adaptation) could be correctly understood. It has been well demonstrated that adaptation in automatic speech systems is very beneficial.

An additional impetus for looking now at how the brain processes speech and language is provided by the dramatic improvements made over the last several years in the field of brain and cognitive science, especially with regard to the cortical imaging of speech and language processing. It is now possible to follow instantaneously the different paths and courses of cortical excitation as a function of differing speech and language stimuli. A major goal here is to understand how significant cortical information processing capabilities beyond signal processing are achieved and to leverage that knowledge in our automated speech and language systems. The ramifications of such an understanding could be very far-reaching. This research area would draw on the related disciplines of brain and cognitive science, natural-language processing, and information retrieval.

SPOKEN-LANGUAGE COMPREHENSION (MIMICKING AVERAGE LANGUAGE SKILLS AT A FIRST-TO-THIRD-GRADE LEVEL)

Today's state-of-the-art systems are designed to transcribe spoken utterances. To achieve a broad level of speech-understanding capabilities, it is essential that the speech research community explore building language-comprehension systems that could be improved by the gradual accumulation of knowledge and language skills. An interesting approach would be to compare an ASR system with the speech performance of children less than ten years of age in listening-comprehension skill. Just like a child learning a new subject, a system could be exposed to a wide range of study materials in a learning phase. In a testing stage, the system and the children would be given written questions first to get some idea what kind of information to look for in the test passages. Comprehension tests could be in oral and written forms.

The goal of this research program is to help develop technologies that enable language comprehension. It is clear such evaluations would emphasize the accurate detection of information-bearing elements in speech rather than basic word error rate. Natural-language understanding of some limited domain knowledge would be needed. Four key research topics need to be explored: 1) partial understanding of spoken and written materials, with a focused attention on information-bearing components; 2) sentence segmentation and name entry extraction from given test passages; 3) information retrieval from the knowledge sources acquired in the learning phase; and 4) representation and database organization of knowledge sources. Collaboration between speech and language processing communities is a key element to the potential success of such a program. The outcomes of this research could provide a paradigm shift for building domain-specific language understanding systems and significantly affect the education and learning communities.

IMPROVING INFRASTRUCTURE FOR FUTURE ASR RESEARCH

CREATION OF HIGH-QUALITY ANNOTATED CORPORA

The single simplest, best way for current state-of-the-art recognition systems to improve performance on a given task is to increase the amount of task-relevant training data from which its models are constructed. System capabilities have progressed directly along with the amount of speech corpora available to capture the tremendous variability inherent in speech. Despite all the speech databases that have been exploited so far, system performance consistently improves when more relevant data are available. This situation clearly indicates that more data are needed to capture crucial information in the speech signal. This is especially important in increasing the facility with which we can learn, understand, and subsequently automatically recognize a wide variety of languages. This capability will be a critical component in improving performance not only for transcription within any given language but also for spoken-language machine translation, cross-language information retrieval, and so on.

If we want our systems to be more powerful and to understand the nature of speech itself, we must collect and label more of it. Well-labeled speech corpora have been the cornerstone on which today's systems have been developed and evolved. The availability of common speech corpora has been and continues to be the sine qua non for rigorous comparative system evaluations and competitive analyses conducted by the U.S. NIST and others. Labeling for most speech databases is typically at the word level. However, some annotation at a finer level (e.g., syllables, phones, features, and so on) is important to understand and interpret speech successfully. Indeed, the single most popular speech database available from the Linguistic Data Consortium (LDC) is TIMIT, a very compact acoustic-phonetic database created by

MIT and Texas Instruments, where the speech is associated with a subword (phonetic) transcription. Over the years, many significant speech corpora, such as Call Home, Switchboard, *Wall Street Journal*, and, more recently, Buckeye, have been made widely available with varying degrees and types of annotation. These corpora and others have fundamentally driven much of our current understanding and growing capabilities in speech recognition, transcription, topic spotting and tracking, and so on. There is a serious need today to understand the basic elements of speech with much larger representative sets of speech corpora, both in English and other languages.

In order to explore important phenomena "above the word level," databases need to be labeled to indicate aspects of emotion, dialog acts, and semantics (e.g., Framenet [14] and Propbank [28]). Human speech understanding is predicated on these factors. For systems to be able to recognize these important characteristics, there must be suitably labeled speech data with which to train them. It is also likely that some new research may be required to explore and determine consistent conventions and practices for labeling itself and for future development and evaluation methodologies to accommodate at least minor differences in labeling techniques and practices. We must design ASR systems that are tolerant of labeling errors.

NOVEL HIGH-VOLUME DATA SOURCES

Thanks in large part to the Internet, there are now large quantities of everyday speech that are readily accessible, reflecting a variety of materials and environments only recently available. Some of it is of quite variable and often poor quality, such as user-posted material from YouTube. Better-quality audio materials are reflected in the diverse oral histories recorded by organizations such as StoryCorps (available at www.storycorps.net). University course lectures, seminars, and similar material make up another rich source, one that is being placed online in a steady stream. These materials all reflect a less formal, more spontaneous, and natural form of speech than present-day systems have typically been developed to recognize. "Weak" transcripts (such as closed-captioning and subtitles) are available for some of these audio materials. The benefit of working with materials such as this is that systems will become more capable as a consequence—an important development in increasing robustness and expanding the range of materials that can be accurately transcribed under a wide range of conditions. Much of what is learned here is also likely to be of benefit in transcribing casual everyday speech in languages other than English.

TOOLS FOR COLLECTING AND PROCESSING LARGE QUANTITIES OF SPEECH DATA

Over the years, the availability of both open-source (e.g., Carnegie Mellon University's CMU Sphinx) and commercial speech tools (e.g., Entropic Systems and Cambridge University's HTK) has been very effective in quickly bringing good-quality speech processing capabilities to many labs and researchers. New Web-based tools could be made available to collect, annotate, and then process substantial quantities of speech very cost-effectively in many languages. Mustering the assistance of interested individuals on the World Wide Web (in the manner of open-source software and Wikipedia) could generate substantial quantities of language resources very efficiently and at little cost. This could be especially valuable in creating significant new capabilities for resource-impoverished languages.

New initiatives, though seriously underfunded at present, include digital library technology aiming to scan huge amounts of text (e.g., the Million Book Project [44]) and the creation of large-scale speech corpora (e.g., the Million Hour Speech Corpus [2]) aiming to collect many hours of speech in many world languages. If successful, these projects will significantly advance the state of the art in the automation of world language speech understanding and proficiency. They will also provide rich resources for strong research into the fundamental nature of speech and language itself.

ACKNOWLEDGMENTS

This article is an updated version of the "MINDS 2006–2007 Report of the Speech Understanding Working Group," one of five reports emanating from two workshops titled "Meeting of the MINDS: Future Directions for Human Language Technology," sponsored by the U.S. Disruptive Technology Office (DTO). (MINDS is an acronym for machine translation, information retrieval, natural-language processing, data resources, and speech understanding; for more information, see www.itl.nist.gov/iaui/894.02/minds.html.) The authors acknowledge significant informative discussions with several colleagues, whose opinions and advice are reflected in the text above. We wish to thank Andreas Andreou, James Baker, Donna Harman, Mary Harper, Hynek Hermansky, Frederick Jelinek, Damianos Karakos, Alex Park, Raj Reddy, Richard Schwartz, and James West.

AUTHORS

Janet M. Baker (janet_baker@email.com) is a cofounder of Dragon Systems and founder of Saras Institute, in West Newton, Massachusetts. She lectures in academic and business venues on speech technology, strategic planning, and entrepreneurship.

Li Deng (deng@microsoft.com) is principal researcher at Microsoft Research, in Redmond, Washington, and affiliate professor at the University of Washington, Seattle. He is a Fellow of the IEEE and of the Acoustical Society of America, and a member of the Board of Governors of the IEEE Signal Processing Society.

James Glass (glass@mit.edu) is a principal research scientist at the MIT Computer Science and Artificial Intelligence Laboratory, where he heads the Spoken Language Systems Group, and a lecturer in the Harvard-MIT Division of Health Sciences and Technology.

Sanjeev Khudanpur (khudanpur@jhu.edu) is an associate professor of electrical and computer engineering in the GWC Whiting School of Engineering of

Johns Hopkins University, in Baltimore, Maryland. He works on the application of information theoretic and statistical methods to human language technologies, including automatic speech recognition, machine translation, and information retrieval.

Chin-Hui Lee (chl@ece.gatech.edu) has been a professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology since 2002. Before joining academia, he spent 20 years in industry, including 15 years at Bell Labs, Murray Hill, New Jersey, where he was the director of dialog system research.

Nelson Morgan (morgan@icsi.berkeley.edu) is the director and speech group leader at ICSI, a University of California, Berkeley-affiliated independent nonprofit research laboratory. He is also professor-in-residence in the University of California, Berkeley EECS Department, the coauthor of a textbook on speech and audio signal processing, and a Fellow of the IEEE.

Douglas O'Shaughnessy (dougo@emt.inrs.ca) is a professor at INRS-EMT (University of Quebec), a Fellow of the IEEE and of the Acoustical Society of America (ASA), and the editor-in-chief of *EURASIP Journal on Audio, Speech, and Music Processing*.

REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 77–83, 1993.

[2] J. K. Baker, "Spoken language digital libraries: The million hour speech, project," in *Proc. Int. Conf. Universal Digital Libraries*, Invited Paper, Alexandria, Egypt, 2006.

[3] J. K. Baker, "Stochastic modeling for automatic speech recognition," *Speech Recognition*, D. R. Reddy, Ed. New York: Academic, 1975.

[4] L. Baum, "An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of a Markov process," *Inequalities*, vol. 3, pp. 1–8, 1972.

[5] F. Beaufays, H. Bourlard, H. Franco, and N. Morgan, "Speech recognition technology," in *Handbook of Brain Theory and Neural Networks*, 2nd ed., M. Arbib, Ed. Cambridge, MA: MIT Press, 2002.

[6] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[7] H. Bourlard, Y. Kamp, H. Ney, and C. Wellekens, "Speaker dependent connected speech recognition via dynamic programming and statistical methods," *Speech and Speaker Recognition* (Bibliotheka Phonetica, vol. 12), M. Schroeder, Ed. Basel: Kargers, 1988.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Brooks, 1984.

[9] J. Bridle, M. Brown, and R. Chamberlain, "A one-pass algorithm for connected word recognition," in *Proc. IEEE ICASSP,* 1982, pp. 899–902.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–21, 1977.

[12] L. Deng and D. OíShaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker, 2003.

[13] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. IEEE ICASSP*, 1996, pp. 346–349.

[14] C. J. Fillmore, C. F. Baker, and H. Sato, "The FrameNet database and software tools," in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, 2002, pp. 1157–1160.

[15] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," in *Proc. IEEE ASRU Workshop*, Santa Barbara, CA, 1997, pp. 3477–3482.

[16] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, 2nd ed. New York: Marcel Dekker, 2001.

[17] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[18] L. Gillick, J. Baker, J. Baker, J. Bridle, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, R. Roth, and F. Scattone, "Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech," in *Proc. IEEE ICASSP*, Apr. 1993, vol. 2, pp. 471–474.

[19] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: Wiley, 2000.

[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[21] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[22] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, Istanbul, Turkey, June 2000, vol. 3, pp. 1635–1638.

[23] X. D. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[24] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997.

[25] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, no. 4, pp. 532–557, 1976.

[26] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Dev.*, vol. 13, pp. 675–685, Nov. 1969.

[27] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[28] P. Kingsbury and M. Palmer, "From treebank to PropBank," in *Proc. LREC*, Las Palmas, Canary Islands, Spain, 2002.

[29] A. Krishnamurthy and D. Childers, "Two channel speech analysis," *IEEE Trans. Acoustics Speech Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[30] R. Kuhn, P. Nguyen, J-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. Int. Conf. Spoken Language*, Sydney, Australia, 1998, pp. 1771–1774.

[31] N. Kumar and A. Andreou, "Heteroscedastic analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283–297, 1998.

[32] K.-F. Lee, *Automatic Speech Recognition: The Development of the Sphinx Recognition System*. New York: Springer-Verlag, 1988.

[33] C.-H. Lee, F. Soong, and K. Paliwal, Eds., *Automatic Speech and Speaker Recognition-Advanced Topics*. Norwell, MA: Kluwer, 1996.

[34] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[35] R. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4–22, Apr. 1987.

[36] H. Ney, The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 32, no. 2, pp. 263–271, 1984.

[37] D. Paul, "Algorithms for an optimal A* search and linearizing the search in a stack decoder," in *Proc. IEEE ICASSP*, 1991, vol. 1, pp. 693–696.

[38] H. Poor, *An Introduction to Signal Detection and Estimation* (Springer Texts in Electrical Engineering), J. Thomas, Ed. New York: Springer-Verlag, 1988.

[39] A. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE ICASSP*, 1988, vol. 1. pp. 1–4.

[40] D. Povey, B, Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: Discriminatively trained features for speech recognition," in *Proc. IEEE ICASSP*, Philadelphia, PA, 2005, pp. 961–964.

[41] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[42] R. Reddy, Ed., *Speech Recognition*. New York: Academic, 1975.

[43] R. Reddy, J. Carbonell, M. Shamos, and G. St. Clair, "The million book digital library project," in *Computer Science Presentation*, Carnegie Mellon Univ., Pittsburgh, PA, Nov. 5, 2003.

[44] A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. IEEE ICASSP*, 1994, pp. 1835–1838.

[45] S. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proc. 7th Int. Congr. Acoustics*, Budapest, Hungary, 1971, vol. 3, pp. 65–69.

[46] F. Soong and E.-F. Huang, "A tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition," in *Proc. HLT Conf. Workshop Speech and Natural Language*, Hidden Valley, PA, 1990, pp. 12–19.

[47] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[48] V. Velichko and N. Zagoruyko, "Automatic recognition of 200 words," *Int. J. Man-Machine Stud.*, vol. 2, no. 3, pp. 223–234, 1970.

[49] T. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 4, no. 2, pp. 81–88, 1968.

[50] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 260–269, 1967.

[51] S. Wilks, *Mathematical Statistics*. New York: Wiley, 1962. [SP]