Updated MINDS Report on Speech Recognition and Understanding, Part 2

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

Updated MINDS Report on Speech Recognition and Understanding, Part 2

This article is the second part of an updated version of the "MINDS 2006–2007 Report of the Speech Understanding Working Group," one of five reports emanating from two workshops entitled "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 further information, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

KNOWLEDGE REPRESENTATION

FUNDAMENTAL SCIENCE OF HUMAN SPEECH PERCEPTION AND PRODUCTION

For long-term research, a principal knowledge source that we can exploit to improve automatic speech recognition (ASR) lies in the area of human speech perception, understanding, and cognition. This rich area has its basis in psychological and physiological processes in humans. The physiological aspects of human speech perception that are of most interest include cortical processing in the auditory area and the associated motor area of the brain. One important principle of auditory perception is its modular organization, and recent advances in functional neuroimaging technologies provide a driving force motivating new studies geared to developing integrated knowledge of the modularly organized auditory process in an end-to-end manner. The relevant psychological aspects of human speech perception include the essential psychoacoustic properties that underlie auditory masking and attention. Such key properties equip human listeners with the remarkable capability to cope with "cocktail party" effects that no current ASR techniques can successfully handle. Intensive studies are needed in order for ASR applications to reach a new level, delivering performance comparable to that of humans.

Specific issues to be resolved in the study of how the human brain processes spoken (as well as written) language are the way human listeners adapt to nonnative accents and the time course over which human listeners reacquaint themselves with a language known to them. Humans have amazing capabilities to adapt to nonnative accents. Current ASR systems are extremely poor in this regard, and improvement is expected only after we have sufficient understanding of human speech processing mechanisms.

One specific issue related to human speech perception (and linked to human speech production) is the temporal span over which speech signals are represented and modeled. One prominent weakness in current hidden Markov models (HMMs) is inadequacy in representing long-span temporal dependency in the acoustic feature sequence of speech, which is an essential property of speech dynamics in both perception and production. The main cause of this handicap is the conditional independence assumptions inherent in the HMM formalism. The HMM framework also assumes that speech can be described as a sequence of discrete units, usually phones or phonemes. In this symbolic, invariant approach, the focus is on the linguistic/phonetic information, and the incoming speech signal is normalized during preprocessing in an attempt to remove most of the paralinguistic information. However, human speech perception experiments have shown that such paralinguistic information plays a crucial role in human speech perception.

Numerous approaches have been taken over the past dozen years to address the weaknesses of HMMs described above [2], [4], [17], [19], [26], [46], [51], [64], [65]. These approaches can be broadly grouped into two categories. The first, a parametric, structure-based approach, establishes mathematical models for stochastic trajectories/segments of speech utterances using various forms of parametric characterization [17], [19], [22], [51]. The essence of such an approach is that it exploits knowledge and mechanisms of human speech perception and production so as to provide the structure of the multitiered stochastic process models. These parametric models account for the observed speech trajectory data based on the underlying mechanisms of speech coarticulation and reduction directly relevant to human speech perception and on the relationship between speaking-rate variations and the corresponding changes in the acoustic features.

The second, nonparametric and template-based approach to overcoming the weaknesses of HMMs involves direct exploitation of speech feature trajectories (i.e., "templates") in the training data without any modeling assumptions [2], [4], [64], [65]. This newer approach is based on episodic learning as seen in many recent human speech perception and recognition experiments [28], [43]. Due to the recent dramatic increase of speech databases and computer storage capacity available for training as well as exponentially expanded computational power, nonparametric methods and episodic learning

Digital Object Identifier 10.1109/MSP.2009.932707

1053-5888/09/$25.00©2009IEEE

provide rich areas for future research [43], [61], [64], [65]. The essence of the template-based approach is that it captures strong dynamic segmental information about speech feature sequences in a way complementary to the parametric, structure-based approach. The recent Sound-to-Sense project in Europe has been devoted to this area of research.

FROM TRANSCRIPTION TO MEANING EXTRACTION

Another rich area for future research is to develop machine representations of "meaning" that capture the communicative intent of a spoken utterance. This would be a complement to "word error rate," the most prevalent criterion for ASR performance. Machine representations are unlikely to achieve universal representations of "meaning," but for specific domains of speech understanding, they should be defined in a way that is consistent with human judgment of meaning in spoken utterances. This new performance measure could provide "feedback" to the low-level components of future ASR systems. For example, if ASR systems are designed with a component that represents articulation effort, then the degree to which the correct meaning is recognized should correlate with the tolerance of a range of the articulation effort. Greater accuracy of meaning understanding or more success in communicating the intent from the speaker to the listener should allow the recognizer to tolerate a wider range of speaking efforts on the part of speaker and hence a greater degree of acoustic variability. This meaning representation may also become the output of the speech system for downstream processing in some applications, such as speech translation, in which a verbatim transcript preserving every acoustic detail is neither necessary nor desirable.

UNDERSTANDING HOW CORTICAL SPEECH/LANGUAGE PROCESSING WORKS

Major advances in high-resolution imaging technologies are now enabling brain scientists to track the spatial and temporal characteristics of how the brain processes speech and language [10], [15], [25], [44], [45]. A combination of direct and EEG recordings with neuroimaging studies using functional MRI (fMRI), positron emission tomography (PET), and magnetoencephalography (MEG) has revealed substantial information about cortical processing of speech and language. In the near term, we can hope to gain significant insights into how the human brain processes this information and try to use that knowledge to improve ASR models, processing, and technology. Many phenomena can now be directly and quantifiably observed, such as the time course and details of adaptation and facilitation, semantic dissonance, and so on. A scientific understanding of cortical processing and adaptation could help us understand how our automated systems should adapt to new acoustic environments or to accented speech as well as the role that episodic learning plays in human speech perception and word recognition.

Insights from recent linguistic, phonetic, and psychological research should be used to understand the interaction of the prior structure of speech (as the knowledge source) with the acoustic measurement of speech (data) and to inform and construct ASR models beyond the current flat-structured HMMs in ASR. Newly constructed models may need to exhibit similar behavior to that of humans when listening and responding to their native languages (accented and unaccented) and foreign languages. Here, accented speech or foreign languages represent situations where the knowledge source is weak on the part of the listener. The counterpart situation—where the information about the data or signal becomes weak—is when the listeners perform ASR under adverse acoustic environments.

Understanding of the interplay between these contrasting situations in human speech perception would provide a wealth of information enabling the construction of better models (better than HMMs) that reflect particular attributes of human auditory processing and the linguistic units used in human speech recognition. For example, to what extent may human listeners use mixed word or phrase "templates" and the constituent phonetic/phonological units in their memory to achieve relatively high performance in speech recognition for accented speech or foreign languages (weak knowledge) and for acoustically distorted speech (weak observation)? How do human listeners use episodic learning (e.g., direct memory access) and parametric learning related to smaller phonetic units (analogous to what we are currently using for HMMs in machines) in speech recognition and understanding? Answers to these questions would benefit the design of next-generation machine speech recognition models and algorithms.

HETEROGENEOUS KNOWLEDGE SOURCES FOR AUTOMATIC SPEECH RECOGNITION

Heterogeneous parallelism in both ASR algorithms and computational structure will be important for research in the next decade. While the incorporation of new types of multiple knowledge sources has been on the research agenda for decades, particularly for ASR, we are entering a period in which the resources will be available to support this strategy in a much more significant way. For instance, it is now possible to incorporate both larger sound units than the typical phone or subphone elements, even for large vocabulary recognition, while still preserving the advantage of the smaller units [67]; additionally, more fundamental units such as articulatory features can be considered [23], [62]. At the level of the signal processing "front end" of ASR, we no longer need to settle on the single best representation, as multiple representations (differentiated by differing tie scales or decompositions of the time-frequency plane) have been shown to be helpful [7], [46]. At the other end of the process, the incorporation of syntactic and semantic cues into the recognition process is still in its infancy. It is possible that deeper semantic representations like Propbank [38] and Framenet [21] could become important in disambiguating similar-sounding recognition hypotheses.

The incorporation of multiple knowledge sources is a key part of what could

also be called multistream analysis. In the cases referred to above, streams correspond to information in quite heterogeneous forms. However, the streams can consist of more homogeneous elements, such as the signals from multiple sensors (e.g., microphone arrays) [20]. On the other hand, the streams can be even more heterogeneous, for instance coming from different modalities (bone-conducted vibration, cameras, or low-power radar) [48]. In all of these cases, architectures are required that can aggregate all of the modules' responses. Various approaches for this have been tried for some time, but we are only now beginning to tackle the task of integrating so many different kinds of sources, due to the emerging availability of the kinds of resources required to learn how best to do the integration.

FOCUSING ON INFORMATION-BEARING ELEMENTS OF THE SPEECH SIGNAL

While speech recognition is often viewed as a classification task, any real system must contend with input that does not correspond to any of the desired classes. These unexpected inputs can take the form of complete words that are not in the recognition vocabulary (including words in a foreign language), word fragments, environmental noises, or nonverbal vocal output (such as laughter). Thus, in addition to the closed-set classification task, speech recognition systems must also reject sounds that do not correspond to members of the desired set. Equivalently, we need to know when ASR may be strongly confident that a word is known, and we must also know when there is low confidence in an ASR result [34]. In many applications, "knowing when we don't know" could be as or even more important than merely having a low word-error rate. Additionally, ASR tends to have poor performance for words within the system vocabulary for which there are few training examples. However, such low-frequency words often contain critical information (for instance, if it is a named entity). Learning how to deal more effectively with both interfering sounds and information-bearing sounds that are poorly represented in our training is a critical area for future research [37].

NOVEL COMPUTATIONAL ARCHITECTURES FOR KNOWLEDGE-RICH SPEECH RECOGNITION

For decades, Moore's law has been a dependable indicator of the increasing capability for calculation and storage in our computational systems. The resulting effects on systems for speech recognition and understanding have been enormous, permitting the use of ever larger training databases and recognition systems and the incorporation of increasingly detailed models of spoken language. Many of the projections for future research implicitly depend on a continued advance in computational capabilities, an assumption that certainly seems justified given recent history. However, the fundamentals of this progression have recently changed [3], [49]. As Intel and others have noted recently, the power density on microprocessors has increased to the point that higher clock rates would begin to melt the silicon die. Consequently, at this point industry development is focused on implementing microprocessors on multiple cores. Dual-core CPUs are now very common, and four- and eight-processor systems are coming out. The new road maps for the semiconductor industry reflect this trend, and future speed increases will come more from parallelism than from having faster individual computing elements. For the most part, algorithm designers for speech systems have ignored the investigation of such parallelism, since the advance of scalar capabilities has been so reliable.

Future progress in many of the directions we discuss here will require significantly more computation; consequently, researchers concerned with implementation will need to consider parallelism explicitly in their designs. This will be a significant change from the status quo. In particular, tasks such as decoding, for which extremely clever schemes to speed up single-processor performance have been developed, will require a complete rethinking of the algorithms [31].

MODELS, ALGORITHMS, AND SEARCH

ADAPTATION AND SELF-LEARNING IN SPEECH RECOGNITION SYSTEMS

Learning Speech recognition has traditionally been cast as a task in which spoken input is classified into a sequence of predefined categories, such as words [33], [55]. ASR development typically proceeds via a heavily supervised training phase that makes use of annotated corpora, followed by a deployment (testing) phase during which model parameters may be adapted to the environment, speaker, topic, and so on while the overall structure remains static. In other words, ASR systems typically do not learn; they undergo supervised training and are relatively static thereafter.

Such an approach stands in stark contrast to human processing of speech and language, where learning is an intrinsic capability [6], [11], [42]. Humans can integrate large amounts of unlabeled (or, at best, lightly annotated) speech [14], [27], [35]. From these data we can learn, among other things, the phonetic inventories of a language and word boundaries, and we can use these abilities to acquire new words and meanings [36], [54], [59]. (In humans, learning and the application of learned knowledge are not separated—they are intertwined.) However, for the most part, speech recognizers are not inherently designed to learn from the data they are meant to classify.

There are many degrees of learning, ranging from "one shot" methods to learning from small amounts of data to learning from partially or poorly labeled or even unannotated, data [53]. Research in this latter area would enable systems to benefit from the enormous quantities of data becoming available online and could reduce the expense and delay associated with our current dependency on high-quality annotations for training. This is especially true for languages for which there are few or no existing large annotated corpora. Finally, research

directed towards self-learning, such as unsupervised pattern-discovery methods, could ultimately prove useful for the general problem of language acquisition—a long-standing "grand challenge" problem in the research community.

GENERALIZATION

Over the past three decades, the speech research community has developed and refined an experimental methodology that has helped to foster steady improvements in speech technology. The approach that has worked well, and has been adopted in other research communities, is to develop shared corpora, software tools, and guidelines that can be used to reduce differences between experimental setups down to the basic algorithms, so that it becomes easier to quantify fundamental improvements. Typically, these corpora are focused on a particular task. As speech technology has become more sophisticated, the scope and difficulty of these tasks have continually increased: from isolated words to continuous speech, from speaker-dependent to speaker-independent, from read to spontaneous, from clean to noisy, from utterance to content-based, and so on.

Although the complexity of such corpora has continually increased, one common property of such tasks is that they typically have a training portion that is quite similar in nature to the test data. Indeed, obtaining large quantities of training data that is closely matched to the test is perhaps the single most reliable method for improving ASR performance. This strategy is quite different from the human experience, however. Over our entire lifetimes, we are exposed to all kinds of speech data from uncontrolled environments, speakers, and topics (i.e., "everyday" speech). Despite this great variation in our own personal training data, we are all able to create internal models of speech and language that are remarkably adept at dealing with variations in the speech chain. This ability to generalize is a key aspect of human speech processing that has not yet found its way into modern speech recognizers. Research on this topic should produce technology that will operate more effectively in novel circumstances and that can generalize better from smaller amounts of data. Examples include moving from one acoustic environment to another and among different tasks and languages. One way to support research in this area would be to create a large corpus of "everyday" speech and a variety of test sets drawn from different conditions. Another research area could explore how well information gleaned from large resource languages and/or domains generalizes to smaller resource languages and domains.

MACHINE LEARNING

This is an exciting time in the machine learning community. Many new algorithms are being explored and are achieving impressive results on a wide variety of tasks. Recent examples include graphical models, conditional random fields, partially observable Markov decision processes, reinforcement-based learning, and discriminative methods such as large-margin or log-linear (maximum entropy) models. Recent developments in effective training of these models make them worthy of further exploration. The speech community would do well to explore common ground with the machine learning community in these areas.

LANGUAGE ACQUISITION

The acquisition of spoken language capability by machine through unsupervised or lightly supervised human intervention remains one of the "grand challenges" of artificial intelligence. While the amount of innate language ability possessed by humans is open to debate [6], [10], [42], the degree of variation in languages across different cultures indicates that linguistic knowledge itself is acquired through interaction with and exposure to spoken language [36], [54], [59]. Although there has been some research in unsupervised acquisition of phones, words, and grammars [8], [9], [12], [16], [40], [52], [60], [63], there remains much opportunity for research in pattern discovery, generalization, and active learning. A research program in language acquisition could have many quantifiable components, based on either speech- or text-based inputs. Particular opportunities exist where natural parallel (e.g., multilingual) or multimodal (e.g., audiovisual) corpora exist, since alternative communication channels provide additional sources of constraint [58].

ROBUSTNESS AND CONTEXT-AWARENESS IN ACOUSTIC MODELS FOR SPEECH RECOGNITION

Probabilistic models, with parameters estimated from sample speech data, pervade state-of-the-art speech technology, including ASR, language identification (LID) and speaker verification [32], [50], [68]. The models seek to recover linguistic information, such as the words uttered, the language spoken, or the identity of the speaker, from the received signal. Many factors unrelated to the information being sought by the models also significantly influence the signal presented to the system.

SPEAKER'S ACOUSTIC ENVIRONMENT AND THE SPEECH ACQUISITION CHANNEL

The acoustic environment in which speech is captured (e.g., background noise, reverberation, overlapping speech) and the communication channel through which speech is transmitted prior to its processing (e.g., cellular, land-line telephone, or VoIP connection, along with call-to-call variability) represent significant causes of harmful variability responsible for drastic degradation of system performance. Existing techniques such as Wiener filtering and cepstral mean subtraction [57] remove variability caused by additive noise or linear distortions, while methods such as RASTA [29] compensate for slowly varying linear channels. However, more complex channel distortions such as reverberation or variable noise (along with the Lombard effect) present a significant challenge.

SPEAKER CHARACTERISTICS AND STYLE

It is well known that speech characteristics (e.g., age, nonnative accent) vary widely among speakers due to many

factors, including speaker physiology, speaker style (e.g., speech rate, spontaneity of speech, emotional state of the speaker), and accents (both regional and nonnative). The primary method currently used for making ASR systems more robust to variations in speaker characteristics is to include a wide range of speakers in the training. Speaker adaptation mildly alleviates problems with new speakers within the "span" of known speaker and speech types but usually fails for new types.

Current ASR systems assume a pronunciation lexicon that models native speakers of a language. Furthermore, they train on large amounts of speech data from various native speakers of the language. A number of modeling approaches have been explored in modeling accented speech, including explicit modeling of accented speech, adaptation of native acoustic models via accented speech data [24], [41] and hybrid systems that combine these two approaches [66]. Pronunciation variants have also been tried in the lexicon to accommodate accented speech [30]. Except for small gains, the problem is largely unsolved.

Similarly, some progress has been made for automatically detecting speaking rate from the speech signal [47], but such knowledge is not exploited in ASR systems, mainly due to the lack of any explicit mechanism to model speaking rate in the recognition process.

LANGUAGE CHARACTERISTICS: DIALECT, VOCABULARY, GENRE

Many important aspects of speaker variability derive from nonstandard dialects. Dialectal differences in a language can occur in all linguistic aspects: lexicon, grammar (syntax and morphology), and phonology. This is particularly damaging in languages where spoken dialects differ dramatically from the standard form, e.g., Arabic [39]. The vocabulary and language use in an ASR task change significantly from task to task, necessitating estimation of new language models for each case. A primary reason language models in current ASR systems are not portable across tasks even within the same language or dialect is that they lack linguistic sophistication: they cannot consistently distinguish meaningful sentences from meaningless ones, nor grammatical from ungrammatical ones. Discourse structure is also rarely considered—merely the local collocation of words.

Another reason why language model adaptation to new domains and genres is very data-intensive is the "nonparametric" nature of the current models. When the genre changes, each vocabulary-sized conditional probability distribution in the model must be reestimated, essentially independently of all the others. Several contexts may share a "backing off" or lower-order distribution, but even those in turn need to be reestimated independently, and so on.

With a few exceptions, such as vocal tract length normalization (VTLN) [13] and cepstral mean subtraction (CMS) [57], models used in today's speech systems have few explicit mechanisms for accommodating most of the uninformative causes of variability listed above. The stochastic components of the model, usually Gaussian mixtures, are instead burdened with implicitly modeling the variability in a frame-by-frame manner. Consequently, when the speech presented to a system deviates along one of these axes from the speech used for parameter estimation, predictions by the models become highly suspect. The performance of the technology degrades catastrophically, even when the deviations are such that the intended human listener exhibits little or no difficulty in extracting the same information.

TOWARDS ROBUST SPEECH RECOGNITION IN EVERYDAY ENVIRONMENTS

Developing robust ASR requires going away from the matched training and test paradigm along one or more of the axes mentioned above. To do so, a thorough understanding of the underlying causes of variability in speech and, subsequently, accurate and parsimonious parameterization of such understanding in the models will be needed. The following issues, however, transcend specific methodologies and will play a key role in any solution in the future.

- A large corpus of diverse speech will have to be compiled, containing speech that carries information of the kind targeted for extraction by the technology and exhibits large (but calibrated) extraneous deviations of the kind against which robustness is sought, such as a diverse speaker population with varying degrees of nonnative accents or different local dialects, widely varying channels and acoustic environments, diverse genres, and so on. Such a corpus will be needed to construct several training and test partitions such that unseen conditions of various kinds are represented.
- Multistream and multiple-module strategies will have to be developed. Any robust method will have to identify reliable elements of the speech spectrum in a data-driven manner by employing an ensemble of analyses and using the analysis that is most reliable in that instance. A multiple-module approach will also entail a new search strategy that treats the reliability of a module or stream in any instance as another hidden variable over which to optimize and seeks the most likely hypothesis over all configurations of these hidden variables.
- New, robust training methods for estimating models from diverse (labeled) data will be required. To adequately train a model from diverse data, either the data will have to be normalized to reduce extraneous variability or training-condition-adaptive transformations will have to be estimated jointly with a condition-independent model, e.g., speaker-adaptive training (SAT) [1] of acoustic models in ASR.
- Detailed, unsupervised adaptation will become even more important in unseen test conditions than it is today. In case of adaptive model transformations, a hierarchical parameterization of the transforms will have to be developed, e.g., from parsimonious ones like VTLN or CMS through multiclass maximum likelihood linear regression (MLLR) to a detailed transformation of every Gaussian density,

in order to permit both robust transform estimation during training and unsupervised transform estimation from test data.

- Exploitation of unlabeled or partially labeled data will be necessary to train the models and to automatically select parts of the unlabeled data for manual labeling in a way that maximizes its utility. This need is partly related to the above-mentioned compilation of diverse training data. The range of possible combinations of channel, speaker, environment, speaking style, and domain is so large that it is unrealistic to expect transcribed or labeled speech in every configuration of conditions for training the models. However, it is feasible to simply collect raw speech in all conditions of interest. Another important reason for unsupervised training will be that the systems, like their human "baseline," will have to undergo lifelong learning, adjusting to evolving vocabulary, channels, language use, and so on.
- Substantial linguistic knowledge will need to be injected into structural design and parameterization of the systems, particularly the statistical language models. There are numerous studies indicating that short segments of speech are locally ambiguous even to human listeners, permitting multiple plausible interpretations. Linguistically guided resolution of ambiguity using cues from a very wide context will be needed to arrive at the "correct" interpretation. Some form of semantics, or representation of meaning, in addition to syntactic structure will have to be used in the system.
- All available metadata and context-dependent priors will have to be exploited by the systems. In a telephony application, for instance, geospatial information about the origin and destination of the call, known priors about the calling and called parties, and knowledge of world events that influence the language, vocabulary, or topic of conversation will have to be used by the system.

Discriminative criteria [5] for parameter estimation throughout the system and multipass recognition strategies, both being pursued today, will also be vital. The former yield more robust models by focusing on categorization rather than description of the training data, while the latter lead to more robust search by quickly eliminating implausible regions of the search space and applying detailed models to a small set of hypotheses likely to contain the correct answer [56].

- Language-universal speech technology is a significant research challenge in its own right, with obvious rewards for "resource-impoverished" languages, and exploiting language universals could yield additional robustness even in resource-rich languages.
- Human performance on actual test data will have to be measured and used (1) for evaluation of robustness, giving systems greater latitude where there is genuine ambiguity and insisting on meeting the "gold standard" where there is no ambiguity and (2) for gaining insights from specific instances in which humans are robust and those in which they are not, leading eventually to new technological solutions.

A research program that emphasizes the accurate transcription of "everyday speech"—by which we mean speech acquired in realistic everyday situations with commonly used microphones from native and nonnative speakers in various speaking styles on a diversity of topics and tasks—will advance the robustness of speech recognition systems along one or more of the axes of variability mentioned above.

NOVEL SEARCH PROCEDURES FOR KNOWLEDGE-RICH SPEECH RECOGNITION

As noted above, search methods that explicitly exploit parallelism may be an important research direction for speech understanding systems. Additionally, as innovative recognition algorithms are added, there will be an impact on the search component. For instance, rather than the left-to-right (and sometimes right-to-left) recognition passes that are used today, there could be advantages to either identifying islands of reliability or islands of uncertainty and relying on alternate knowledge sources only "locally" in the search process. The incorporation of multiple tiers of units (such as articulatory feature, subphone state, phone, syllable, word, and multiword phrase) could have consequences for the search process. Finally, so-called "episodic" approaches to ASR are being investigated [64]. These rely on examples of phrases, words, or other units directly, as opposed to statistical models of speech. While this seems to be a throwback to the days before the prominence of HMMs, the idea is gaining new prominence due to the availability of larger and larger speech databases and thus more and more examples for each modeled speech unit. It could well be that an important future direction would be to learn how best to incorporate these approaches into a search that also uses statistical models, which have already proven their worth.

CONCLUSIONS

We have surveyed historically significant events in speech recognition and understanding that have enabled this technology to become progressively more capable and cost-effective in a growing number of everyday applications. With additional research and development, significantly more valuable applications are within reach.

A set of six ambitious, achievable, and testable "grand challenge" tasks has been proposed. Successful achievement of these would lay the groundwork for bringing a number of high-utility applications to reality. Each of these challenge tasks should benefit and be benefited by collaboration and cross-fertilization with related human-language technologies, especially machine translation, information retrieval, and natural-language processing, as well as brain and cognitive science. Research achievements in speech recognition and understanding have

demonstrably led to major advances in related human-language technologies as well as more general areas such as pattern recognition.

To enable and implement these grand challenges, a number of especially promising research directions were outlined and supported. Though these have been largely unfunded so far, the pursuit of these initiatives would contribute to a substantial increase in the core technology on which robust future applications depend.

ACKNOWLEDGMENTS

The authors acknowledge significant informative discussions with several colleagues, whose opinions and advice is reflected in the text above. We wish to thank Andreas Andreou, James Baker, 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 (ASA) and a member of the Board of Governors of the IEEE Signal Processing Society.

Sanjeev Khudanpur (khudanpur@jhu.edu) is an associate professor of electrical and computer engineering in the GWC Whiting School of Engineering of the Johns Hopkins University, in Baltimore, Maryland. He works on the application of information theoretic and statistical methods to human-language technologies, including ASR, machine translation, and information retrieval.

Chin-Hui Lee (chl@ece.gatech.edu) has been a professor since 2002 at the School of ECE, Georgia Institute of Technology, in Atlanta. 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.

James R. 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 is a lecturer in the Harvard-MIT Division of Health Sciences and Technology.

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 Electrical Engineering and Computer Science Department at the University of California, Berkeley, 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 ASA, and the editor-in-chief of *EURASIP Journal on Audio, Speech, and Music Processing*.

REFERENCES

[1] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1997, pp. 1043–1046.

[2] G. Aradilla, J. Vepa, and H. Bourlard, "Improving speech recognition using a data-driven approach," in *Proc. Eurospeech*, pp. 3333–3336, Sept. 2005.

[3] K. Asanovic, R. Bodik, B. C. Catanzaro, J. Gebis, P. Husbands, K. Keutzer, D. Patterson, W. Plishker, J. Shalf, S. Williams, and K. Yelick, "The landscape of parallel computing research: A view from Berkeley," EECS Dept., Univ.California at Berkeley, Tech. Rep. UCB/EECS-2006-183, Dec. 2006.

[4] S. Axelrod and B. Maison, "Combination of hidden Markov models with dynamic time warping for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2004, vol. 1, pp. 173–176.

[5] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1986, pp. 49–52.

[6] L. Bloomfield, *Language*. New York: Holt, 1933.

[7] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Oct. 1996, vol. 1, pp. 426–429.

[8] M. R. Brent, "An efficient probabilistically sound algorithm for segmentation and word discovery," *Mach. Learn.*, vol. 34, no. 1–3, pp. 71–105, Feb. 1999.

[9] E. Brill, "A corpus-based approach to language learning," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA, 1993.

[10] A. Chan, S. Cash, E. Eskandar, J. M. Baker, C. Carlson, O. Devinsky, W. Doyle, R. Kuzniecky, T. Thesen, C. Wang, K. Marinkovic, and E. Halgren, "Decoding semantic category from MEG and intracranial EEG in humans," in *Proc. Neuroscience 2008 Conf.*, Washington, DC.

[11] N. A. Chomsky, *Knowledge of Language: Is Nature, Origin, and Use*. New York: Praeger, 1986.

[12] A. Clark, "Unsupervised language acquisition: Theory and practice," Ph.D. dissertation, Univ. Sussex, Brighton, U.K., 2001.

[13 J. Cohen, T. Kamm, and A. G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, pp. 3246–3247, May 1995.

[14] S. Crain, "Language acquisition in the absence of experience," *Behav. Brain Sci.*, vol. 14, no. 4, pp. 601–699, Dec. 1991.

[15] A. M. Dale and E. Halgren, "Spatiotemporal mapping of brain activity by integration of multiple imaging modalities," *Curr. Opin. Neurobiol.*, vol. 11, no. 2, pp. 202–208, 2001.

[16] C. G. de Marcken, "Unsupervised language acquisition," Ph.D. dissertation, MIT, Cambridge, MA, 1996.

[17] L. Deng, M. Aksmanovic, D. Sun, and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 507–520, 1994.

[18] L. Deng, D. Yu, and A. Acero, "Structured speech modeling," *IEEE Trans. Audio, Speech Lang. Process.* (Special Issue on Rich Transcription), vol. 14, no. 5, pp. 1492–1504, Sept. 2006.

[19] L. Deng and D. O'Shaughnessy, *Speech Processing—A Dynamic and Optimization-oriented Approach*. New York: Marcel Dekker, 2003.

[20] K. Farrell, R. Mammone, and J. Flanagan, "Beamforming microphone arrays for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1992, pp. 285–288.

[21] C. J. Fillmore, C. F. Baker, and H. Sato, "The frame net database and software tools," in *Proc. 3rd Int. Conf. Language Resources and Evaluation (LREC)*, Las Palmas, 2002, pp. 1157–1160.

[22] J. Frankel and S. King, "Speech recognition using linear dynamic models," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 246–256, 2007.

[23] J. Frankel and S. King, "ASR—Articulatory Speech Recognition," in *Proc. Eurospeech, Aalborg*, Denmark, 2001, pp. 599–602.

[24] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, no. 7, pp. 711–720, 1997.

[25] J. S. George, C. J. Aine, J. C. Mosher, D. M. Schmidt, D. M. Ranken, and H. A. Schlitt, "Mapping function in the brain with magnetoencephalography, anatomical magnetic resonance imaging, and functional magnetic resonance imaging," *J. Clin. Neurophysiol.*, vol. 12, no. 5, pp. 406–431, 1995.

[26] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Comput., Speech Lang.*, vol. 17, no. 2–3, pp. 137–152, 2003 (Eds.: M. Russell and J. Bilmes, Special Issue).

[27] H. Goodluck, *Language Acquisition*. Cambridge, MA: Blackwell Publishers, 1991.

[28] S. Hawkins, "Contribution of fine phonetic detail to speech understanding," in *Proc. 15th Int. Congress of Phonetic Sciences (ICPhS-03)*, Barcelona, Spain, 2003, pp. 293–296.

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

[30] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modeling for robust speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 2324–2327.

[31] A. Janin, "Speech recognition on vector architectures," Ph.D. dissertation, Univ. California, Berkeley, 2004.

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

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

[34] L. Jiang and X. D. Huang, "Vocabulary independent word confidence measure using subword features," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 401–404.

[35] P. W. Jusczyk, *The Discovery of Spoken Language*. Cambridge MA: MIT Press/Bradford Books, 1997.

[36] P. W. Jusczyk and R. N. Aslin, "Infants' detection of sound patterns of words in fluent speech," *Cogn. Psychol.*, vol. 29, no. 1, pp. 1–23, Aug. 1995.

[37] H. Ketabdar and H. Hermansky, "Identifying unexpected words using in-context and out-of-context phoneme posteriors," Tech. Rep., IDIAP-RR 06-68, 2006.

[38] P. Kingsbury and M. Palmer, "From tree bank to prop bank," in *Proc. LREC, Las Palmas*, Canary Islands, Spain, 2002.

[39] K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, J. Gang, H. Feng, J. Henderson, L. Daben, M. Noamany, P. Schone, R. Schwartz, and D. Vergyri, "Novel approaches to Arabic speech recognition: Report from the 2002 Johns-Hopkins summer workshop," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 2003, pp. 344–347.

[40] D. Klein, "The unsupervised learning of natural language structure," Ph.D. dissertation, Stanford Univ., Palo Alto, CA, 2005.

[41] C. Leggetter and P. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," in *Proc. Int. Conf. Spoken Language Processing*, 1994, pp. 451–454.

[42] D. Lewis, "Languages and language," in *Language, Mind, and Knowledge*, K. Gunderson, Ed. Minneapolis, MN: Univ. Minnesota Press, 1975, pp. 3–35.

[43] V. Maier and R. K. Moore, "An investigation into a simulation of episodic memory for automatic speech recognition," in *Proc. Interspeech 2005*, Lisbon, Portugal, 5–9 Sept. 2005, pp. 1245–1248.

[44] K. Marinkovic, "Spatiotemporal dynamics of word processing in the human cortex," *Neuroscientist*, vol. 10, no. 2, pp. 142–152, 2004.

[45] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[46] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos, "Pushing the envelope-aside," *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 81–88, Sept. 2005.

[47] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1998, pp. 729–732.

[48] L. C. Ng, G. C. Burnett, J. F. Holzrichter, and T. J. Gable, "Denoising of human speech using combined acoustic and EM sensor signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 5–9 June 2000, Istanbul, Turkey, pp. 229–232.

[49] K. Olukotun, "A conversation with John Hennessy and David Patterson," *ACM Queue Mag.*, vol. 4, no. 10, pp. 14–22, Dec./Jan. 2006–2007.

[50] D. O'Shaughnessy, "Speaker recognition," *IEEE Acoust. Speech Signal Process. Mag.*, vol. 3, no. 4, pp. 4–17, 1986.

[51] M. Ostendorf, V. Digalakis, and J. Rohlicek, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, 1996.

[52] A. Park, "Unsupervised pattern discovery in speech: Applications to word acquisition and speaker segmentation," Ph.D. dissertation, MIT, Cambridge, MA, 2006.

[53] F. Pereira and Y. Schabes, "Inside-outside re-estimation from partially bracketed corpora," in *30th Annu. Meeting of the Association for Computational Linguistics*, Newark, DE, 1992, pp. 128–135.

[54] S. Pinker. *The Language Instinct*. New York: William Morrow and Co., 1994.

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

[56] F. Richardson, M. Ostendorf, and J. R. Rohlicek, "Lattice-based search strategies for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, May 1995, pp. 576–579.

[57] A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1994, pp. 1835–1838.

[58] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cogn. Sci.*, vol. 26, no. 1, pp. 113–146, Jan. 2002.

[59] J. R. Saffran, "Constraints on statistical language learning," *J. Mem. Lang.*, vol. 47, no. 1, pp. 172–196, July 2002.

[60] Z. Solan, D. Horn, E. Ruppin, and S. Edelman, "Unsupervised context sensitive language acquisition from a large corpus," in *Advances in Neural Information Processing Systems*, L. Saul, Ed. Cambridge, MA: MIT Press, vol. 16, 2004.

[61] H. Strik, "How to handle pronunciation variation in ASR: By storing episodes in memory?," in *Proc. ITRW on Speech Recognition and Intrinsic Variation (SRIV2006)*, Toulouse, France, May 2006, pp. 33–38.

[62] J. Sun and L. Deng, "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoust. Soc. Amer.*, vol. 111, no. 2, pp. 1086–1101, Feb. 2002.

[63] A. Venkataraman, "A statistical model for word discovery in transcribed speech," *Comput. Linguist.*, vol. 27, no. 3, pp. 352–372, Sept. 2001.

[64] M. Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data-driven example based continuous speech recognition," in *Proc. EUROSPEECH*, Geneva, Sept. 2003, pp. 1133–1136.

[65] M. Wachter, K. Demuynck, and D. Van Compernolle, "Boosting HMM performance with a memory upgrade," in *Proc. Interspeech*, Pittsburgh, PA, Sept. 2006, pp. 1730–1733.

[66] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2003, pp. 540–543.

[67] S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg, "Performance improvements through combining phone-and syllable-scale information in automatic speech recognition," in *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998, pp. 854–857.

[68] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 31–44, Jan. 1996. [SP]