Historical Development and Future Directions in Speech Recognition and Understanding

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

This report is one of five reports that were based on the MINDS workshops, led by Donna Harman (NIST) and sponsored by Heather McCallum-Bayliss of the Disruptive Technology Office of the Office of the Director of National Intelligence's Office of Science and Technology (ODNI/ADDNI/S&T/DTO). To find the rest of the reports, and an executive overview, please see http://www.itl.nist.gov/iaui/894.02/minds.html.

I. Introduction

On November 13-14, 2006, a workshop entitled "Meeting of the MINDS: Future Directions for Human Language Technology," sponsored by the U.S. Government's Disruptive Technology Office (DTO), was held in Chantilly, VA.. A second workshop was subsequently held on February 25-26, 2007 in Marina del Rey, CA.. "MINDS" is an acronym for **M**achine Translation, **I**nformation Retrieval, **N**atural Language Processing, **D**ata Resources, and **S**peech Understanding. These 5 areas were each addressed by a number of experienced researchers. The goal of these working groups was to identify and discuss especially promising future research directions, especially those which are un(der)funded. The intent is to elicit from the Human Language Technology Community itself a set of well-considered directions or "Rich Areas for Future Research" that could lead to major paradigm shifts in the field. Each group first reviewed major past developments in their respective fields and the circumstances that led to their success, and then focused on areas they deemed especially fertile for future research. The Speech Understanding working group was tasked to focus specifically on speech recognition/understanding, and did not at this time address speech synthesis, speech-based human-computer interface, or dialogue systems. Cross-disciplinary research areas were proposed as well as 3-5 year "Grand Challenges" to stimulate advanced research by dealing with realistic tasks of broad interest.

This report includes the following sections: Introduction, Historically Significant Developments in Speech Recognition and Understanding, Grand Challenges, Rich Areas for Future Research, Acknowledgments, Conclusions, and References.

II. Historically Significant Developments in Speech Recognition and Understanding

Major developments in the technology of Speech Recognition and Understanding are focused in this report primarily over the past 35 years, a period which has witnessed this multidisciplinary field proceed from its infancy to its coming of age shortly before the turn of this past century. Though far from a "solved" problem, it now has a growing number of practical applications in many sectors. Further research and development will enable increasingly more powerful systems, deployable on a world-wide basis. The scope of this discussion focuses primarily on issues relating to speech recognition technology, and does not yet address the many critical issues relating specifically to speech synthesis, dialogue, or human-computer interfaces.

This discussion will briefly review some highlights of those developments in five areas: Infrastructure, Knowledge Representation, Models and Algorithms, Search, and Metadata. Broader and deeper discussions of these areas with citations are available from a number of excellent sources (e.g. books: Furui, 2001; Gold and Morgan, 2000; Huang, Acero, and Hon, 2001; Jelinek, 1997, Jurafsky and Martin, 2000; Lee, Soong and Paliwal, 1996; Deng and O'Shaughnessy, 2003; Lee, 1988; Rabiner and Juang, 1993; Reddy, 1975; Stevens, 1998; and websites: IEEE History Center, Automatic Speech Synthesis and Recognition; Saras Institute History of Speech and Language Technology Project; Smithsonian Speech Synthesis History Project).

1. Infrastructure: Moore's Law, doubling the amount of computation for a given cost in every 12 - 18 months, in conjunction with the constantly shrinking cost of memory, has been instrumental in enabling Speech Recognition/Understanding researchers to run increasingly complex systems in short enough time (e.g. meaningful experiments in a day or less) to make significant progress and improvement over the past three decades.

The availability of common speech corpora for speech training, development, and evaluation, has been critical in creating systems of 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 world-wide community by the National Institute of Science and Technology (NIST), the Linguistic Data Consortium (LDC), and others. The character of the recorded speech has progressed from limited, constrained speech materials to masses of progressively more realistic, spontaneous and “found” speech.
The development and adoption of rigorous benchmark evaluations and standards, nurtured by the National Institute of Science and Technology 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 HTK, Sphinx, CMU LM toolkit, SRILM toolkit, etc. Extensive research support combined with workshops, task definitions, and system evaluations sponsored by DARPA and others have been essential to today's system developments.

2. Knowledge Representation: Major advances in speech signal representations have included perceptually motivated Mel-Frequency Cepstral Coefficients (MFCC) (Krishnamurthy and Childers, 1986), Perceptual Linear Prediction (PLP) coefficients (Hermansky, 1990) as well as Cepstral Mean Subtraction (CMS) (Rosenberg et al., 1994; Furui, 2001), RASTA (Hermansky and Morgan, 1994), and Vocal Tract Length Normalization (VTLN) (Eide, 1996).

Architecturally, the most important development has been the searchable unified graph representations allowing multiple sources of knowledge to be incorporated in a common probabilistic framework. Non-compositional methods include multiple speech streams, multiple probability estimators, multiple recognition systems combined at the hypothesis level (e.g. ROVER (Fiscus, 1997), etc.), and multipass systems with increasing constraints (bigram vs. 4-gram, within word dependencies vs. cross-word, etc). 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) (Kumar and Andreou, 1998), feature-space minimum phone error (fMPE) (Povey, 2005), and neural net-based features (Hermansky et al., 2000).

3. Models and Algorithms: The most significant paradigm shift has been the introduction of statistical methods, especially stochastic processing with Hidden Markov Models (HMMs) (Baker, 1975, and Jelinek, 1976) in the early 1970's (Poritz, 1988). 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 (Dempster, 1977) and the Forward-Backward or Baum-Welch algorithm (Baum, 1972) 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 (Breiman, 1984) 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 (Bahl et al, 1993) and some neural network techniques (Lippman, 1987; Beaufays et al., 2002).

Adaptation is vital to accommodating a wide range of variable conditions for the channel, environment, speaker, vocabulary, domain, etc. Popular techniques include Maximum a posteriori probability (MAP) estimation (Wilks, 1962; Poor, 1988; Gauvain and Lee, 1997) , Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995) and Eigenvoices (Kuhn, 1998). This can take place on the basis of small amounts of data from new tasks or domains for additional training material, as well as “one-shot” learning, or “unsupervised” training at test time.

4. Search: Key decoding or search strategies, originally developed in non-speech applications, have focused on stack decoding (A* search) (Jelinek, 1969) and Viterbi or N-best search (Viterbi, 1967). Derived from communications and information theory, stack decoding was subsequently applied to speech recognition systems (Jelinek, 1976; Paul, 1991). Viterbi search, broadly applied to search alternative hypotheses, derives from dynamic programming in the 1950's (Bellman, 1957), and subsequent speech applications in the 1960's and 1970's, to the 1980's from Russia and Japan to the U.S. and Europe (Vintsyuk, 1968; Velichko, 1970, Sakoe and Chiba, 1971; Baker, 1975; Bridle, Brown, and Chamberlain, 1982, Ney, 1984; Bourlard, Kamp, Ney, and Wellekens, 1988).

5. Metadata: The automatic determination for sentence and speaker-segmentation as well as punctuation (Huang and Zweig, 2002) has become a key feature in some processing systems. Starting in the early 1990's, audio indexing and mining enabled high performance automatic topic detection and tracking, as well as applications for language and speaker identification (Gillick et al, 1993).

III. Grand Challenges

Grand Challenges are ambitious but achievable 3-5 year research program initiatives that will significantly advance the state-of-the-art in speech recognition and understanding. Six (6) such programs are described below. Each proposed program has defined measurable goals and comprises a complex of important capabilities that will substantially advance the field and enable significant applications. These are rich task domains which will enable progress on multiple 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.

1. Everyday Audio

"Everyday Audio" is a term that represents a large range of speech, speaker, channel, and environmental conditions which people typically encounter and routinely adapt to in responding and recognizing speech signals. Currently automatic recognition systems are challenged, and can deliver significantly degraded performance, when they encounter audio signals that differ sometimes even slightly from the limited conditions under which they were originally developed and "trained."

This 3-5 year research area would focus 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, etc.), speaker characteristics (e.g. speaking style, non-native accents, emotional state, etc.), and language characteristics (e.g. formal/informal styles, dialects, vocabulary, topic domain, etc.).

New techniques and architectures are proposed to enable exploring these critical issues in meaningful environments as diverse as meeting room presentations to unstructured conversations. A primary focus will be exploring alternatives for automatically adapting to changing conditions in multiple dimensions, even simultaneously.

Its 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.

2. Rapid Portability to Emerging Languages

Today's state-of-the-art systems deliver top performances by building complex acoustic and language models using a large collection of domain-specific speech and text examples. This set of language resources is often not readily available for rarely-seen languages that could be emerging in the horizon.

The goal of this research program is to create rapid portability spoken language technologies.

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, namely: (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, can 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 them 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 comparable performance level similar to today's high-accuracy systems.

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

3. 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 human speech facility, which is constantly updated over a lifetime, routinely acquiring new vocabulary items and idiomatic expressions, and experience with previously unseen non-native 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, non-speech 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 will engage in automatic pattern discovery, active learning and adaptation. Research in this area must address both the learning of new models, as well as the integration of such models into pre-existing 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, a speech recognition system may encounter a new proper noun in the 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 will be necessary for such learning, perhaps including the automatic selection (by the system) of parts of the unlabeled data for manual labeling in a way that maximizes its utility.

A motivation for investing in such research is the growing activity in the allied field of Machine Learning. Success in this endeavor will extend the lifetime of deployed systems, and directly advance our ability to develop speech systems in new languages and domains without onerous demands of labeled speech, by essentially creating systems that automatically learn and improve over time. This research will benefit from cross-fertilization with Natural Language Processing, Information Retrieval, and Cognitive Science.

4. Detection of Rare, Key Events

Current speech recognition systems have difficulty in handling unexpected – and thus often the most information rich – lexical items. This is especially problematic in speech that contains interjections of 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 key to tasks such as *spoken term detection* and *information extraction* from speech. Their accurate detection is therefore of vital importance.

The goal of this program is to create systems that reliably detect when they don’t know a (correct) word.

A clue to the occurrence of such error events is the *mismatch* between an analysis of the purely sensory signal unencumbered by prior knowledge, such as unconstrained phone recognition, and the word- or phrase-level hypothesis based on higher level knowledge, often encoded in a language model. A key component of this research will therefore be to develop 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 is to transcribe them phonetically when the system is confident that its word hypothesis is unreliable, and devise error-correction schemes.

An immediate application that such detection will enable is sub-word (e.g. phonetic) indexing and search of speech regions where the system suspects the presence of errors. Phonetic transcription of the error-prone regions will also enable the development of the next generation of self-learning speech systems – the system may 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.

5. Cognition-derived Speech and Language Systems

A key cognitive characteristic of humans is their ability to learn and adapt to new patterns, and stimuli. Although this behavior is very important and relatively well-understood in humans, very little of this knowledge has found its way into automatic speech and language systems. *The focus of this 3-5 year project is 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, domains, etc., 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, etc.) can 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, are the dramatic improvements made in recent 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 enable an understanding of 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 achieving that understanding could be very far-reaching. This research area will draw on the related disciplines of brain and cognitive science, natural language processing, and information retrieval.

6. Spoken Language Comprehension (Mimicking average language skills at 1st to 3rd grade)

Today’s state-of-the-art systems are designed to transcribe spoken utterances. To achieve a broad level of speech understanding capabilities, it is critical that the community explores building language comprehension systems that can be improved by gradual accumulation of knowledge and language skills. An interesting approach is to compare a system with a child at 1st to 3rd grade for listening comprehension skill. Just like a child learning a new subject, a system can be exposed to a wide rage of study materials in the learning phase. In the testing stage, the systems and the children are given the written questions first to get some idea what kind of information to look for in the test passages. Comprehension tests given by teachers can be in both oral and written forms.

The goal of this research program is to facilitate language comprehension enabling technologies.

It is clear such evaluations will emphasize accurate detection of information-bearing elements in speech rather than raw word error rate. Natural language understanding of some limited domain knowledge is needed. Four key research topics need 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 driver to the success of this program. The outcomes of this research will provide a paradigm shift for building domain-specific language understanding systems, and impact the education and learning communities.

IV. Rich Areas for Future Research

Following lengthy discussion and debate, a number of especially rich areas for future research have been designated. These especially fertile areas are grouped according to whether they may be considered as part of Infrastructure, Knowledge Representation, or Models, Algorithms, and Search.

1. Infrastructure

Creation of high-quality annotated corpora: The single best proven way for present 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. The evolution of system capabilities has progressed hand-in-hand with the availability and use of increasingly sizeable and complex speech corpora to capture the tremendous variability inherent in the speech signal to be decoded. The dividends resulting from the creation and availability of this data have been repeatedly demonstrated. Despite all the various speech databases that have been created to date, system performance consistently improves when more is available. System performance with existing corpora has not yet reached an asymptote. This situation clearly indicates that more data is needed for capturing 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 world 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, etc.

We have barely scratched the surface in collecting and sampling the many kinds of speech, environments, and channels that people routinely recognize. Parenthetically, we observe that we currently provide to our automatic systems only a very small fraction of the amount of materials that humans utilize to acquire language. 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 also has been and continues to be the *sine qua non* for rigorous comparative system evaluations, and competitive analyses conducted by the National Institute for Science and Technology (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, etc) is important in successfully understanding and interpreting speech. Indeed, the single most popular speech database available from the LDC is TIMIT, a very compact acoustic-phonetic database, created by MIT, where the speech data also contains a subword (phonetic) transcription. Over the years, a succession of significant speech corpora, such as Wall Street Journal, Switchboard, Call Home, 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, etc. There is a serious R&D need today for understanding the basic elements of the speech signal with much larger representative sets of speech corpora, both in English and other world languages.

In order to explore important phenomena “above the word” level, labeling also needs to be available to indicate emotion, dialog acts, and semantics (e.g. Framenet (Fillmore, Baker, and Sato, 2002), Propbank (Kingsbury and Palmer, 2002)). Human speech understanding is predicated on these factors. If our systems hope to recognize these important characteristics, there must be suitably labeled speech data on 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 systems that are tolerant to labeling errors.

High-volume data sources very different from the ones we know: We now have some very exciting opportunities to collect large amounts of audio data that have not previously been available. Thanks in large part to the Internet, there are now readily accessible large quantities of "Everyday" speech, reflecting a variety of materials and environments previously unavailable. Some of it is of quite variable and often poor quality audio, such as user-posted material from YouTube (www.YouTube.com). Better quality audio materials are reflected in the diverse oral histories recorded by organizations such as StoryCorps (www.StoryCorps.net). Another rich source are university course lectures, seminars and similar material, which are progressively being put on-line. These materials all reflect a less formal, more spontaneous, and natural form of speech than present-day systems have typically been developed to recognize. Some "weak" transcripts; e.g. closed captions, subtitles, etc., are available for some of these audio materials. The benefits 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 non-English languages.

Tools to collect and process large quantities of speech data: Over the years, the availability of both open source (e.g. Carnegie Mellon University Sphinx) and commercial speech tools (e.g. Entropic Systems/Cambridge University 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 worldwide web (e.g. open source software, Wikipedia,) could generate substantial quantities of language resources very efficiently and cost-effectively. This could be especially valuable for creating significant new capabilities for resource "impoverished" languages.

New initiatives, though seriously under-funded at present, include digital library technology aiming to scan huge amounts of text (e.g. the Million Book Project (Reddy et al, 2003)) and the creation of large-scale speech corpora (e.g. the Million Hour Speech Corpus, (Baker, 2006)) aiming to collect many hours of speech in many world languages. If successful, these projects will make a significant impact in advancing the state-of-the-art in the automation of world language speech understanding and proficiency. These resources will also provide rich resources to enable strong research into the fundamental nature of speech and language itself.

2. Knowledge Representation

2.1 Fundamental Science of Human Speech Perception and Production

One principal knowledge source that we can draw to benefit machine speech recognition for long-term research is in the area of human speech perception, understanding, and cognition. This rich knowledge source has its basis in both psychological and physiological processes in humans. Physiological aspects of human speech perception of most interest include cortical processing in the auditory area as well as in the associated motor area of the brain. One important principle of auditory perception is its modular organization, and recent advances in functional neuro-imaging technologies provide a driving force motivating new studies towards developing integrated knowledge of the modularly organized auditory process in an end-to-end manner. Psychological aspects of human speech perception embody the essential psychoacoustic properties that underlie auditory masking and attention. Such key properties equip human listeners with the remarkable capability of coping with cocktail party effects that no current automatic speech recognition techniques can successfully handle. Intensive studies are needed in order for speech recognition and understanding applications to reach a new level, delivering performance comparable to 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 non-native accents and the time course over which human listeners re-acquaint themselves to a language known to them. Humans have amazing capabilities to adapt to non-native accents. Current ASR systems are extremely poor in this aspect, and the improvement is expected only after we have sufficient understanding of human speech processing mechanisms.

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

Numerous approaches have been taken over the past dozen years to address the above weaknesses of HMMs (e.g., Ostendorf et al., 1996; Deng et al. 1994; Aradilla et al., 2005; Wachter et al., 2003, 2006; Deng and O'Shaughnessey, 2003; Axelrod and Maison, 2004; Glass 2003; Morgan et al., 2005). These approaches can be broadly classified into the following two categories. The first, parametric, structure-based approach establishes mathematical models for stochastic trajectories/segments of speech utterances using various forms of parametric characterization (e.g., Ostendorf et al., 1996; Deng et al. 1994; Deng et al., 2003; Frankel and King, 2006). 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 multi-tiered 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, non-parametric and template-based approach to overcoming the HMM weaknesses involves direct exploitation of speech feature trajectories (i.e. "template") in the training data without any modeling assumptions (Aradilla et al., 2005; Wachter et al., 2003, 2006; Axelrod and Maison, 2004). This newer approach is based on episodic learning as evidenced in many recent human speech perception and recognition experiments (Hawkins, 2003; Maier and Moore, 2005). Due to the dramatic increase of speech databases and computer storage capacity available for training, as well as the exponentially expanded computational power, non-parametric methods and episodic learning provide rich areas for future research (Wachter et al., 2003, 2006; Maier and Moore, 2005; Strik, 2006). The essence of the template-based approach is that it captures strong dynamic segmental information about speech feature sequences in a way complimentary to the parametric, structure-based approach. The recent Sound-to-Sense project in Europe has been devoted to this area of research.

2.2 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 should augment the current "word error rate" measure for speech recognition performance. It is 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 speech recognition systems. For example, after the speech recognition 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.

2.3 Understanding How Cortical Speech and 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 (George et al, 1995; Dale and Halgren, 2001; Marinkovic, 2004). A combination of direct and EEG recordings with neuroimaging studies using fMRI, PET and MEG technologies have revealed substantial information about cortical processing of speech and language. Near-term, we can now hope to gain significant insights into how the human brain processes this information and to try to use that knowledge to benefit speech recognition 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, etc. 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, or the role that episodic learning plays in human speech perception and word recognition.

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

Understanding of the interplay between these opposing situations in human speech perception will provide a wealth of information enabling the construction of better models (than HMMs) that reflect 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/understanding? Answers to these questions will benefit our design of next-generation machine speech recognition models and algorithms.

2.4 Heterogeneous Knowledge Sources for Automatic Speech Recognition

Heterogeneous parallelism in both the algorithms and the 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 speech recognition, we are coming into a period where the resources are 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 sub-phone elements) even for large vocabulary recognition, while still preserving the advantage of the smaller units (Wu et al 1998); additionally, more fundamental units such as articulatory features can be considered (Sun and Deng, 2002; Frankel and King, 2001). At the level of the signal processing "front end", 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 (Bourlard and Dupont 1996, Morgan et al 2005). 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 (Kingsbury and Palmer 2002) and Framenet (Fillmore et al 2002) will become important in disambiguating between similar sounding recognition hypotheses.

The incorporation of multiple knowledge sources is a key part of what could also be called multi-stream analysis. In the cases referred to above, the 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) (Farrell et al 1992). 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) (Ng et al 2000). 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 kind of resources required to learn how to best do the integration.

2.5 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 non-verbal 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 we are strongly confident that a word is known, and also know when we are not confident of a recognition result (Jiang and Huang 1998). In many applications, "knowing when we don't know" could be as or even more important than just having a low word error rate. Additionally, we tend to have poor performance for words that are within the system vocabulary, but for which there are minimal training examples. However, such low-frequency words frequently contain critical information (for instance if it is a named entity). Learning how to deal more effectively with both interfering sounds and with information-bearing sounds that are poorly represented in our training is a critical area for future research (Ketabdar and Hermansky 2006).

2.6 Novel Computational Architectures for Knowledge-Rich Speech Recognition

For decades, Moore's law has been a dependable indicator of the increased capability for computation and storage in our computational systems. The resulting effects on systems for speech recognition and understanding have been enormous, permitting the use of larger and larger training databases and recognition systems, and the incorporation of more and more detailed models of spoken language. Many of the projections for future research in this document implicitly depend upon a continued advance in computational capabilities, an assumption that certainly seems justified given recent history. However, the fundamentals of this progression have recently changed (Asanovic et al., 2006; Olukotun 2007). 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 now focused on implementing microprocessors on multiple cores. Dual core CPUs are now very common, and 4-processor and 8-processor systems are coming out. The new road maps for the semiconductor industry reflect this trend, and future speedups will come more from parallelism than from having faster individual computing elements. For the most part, algorithm designers for speech systems have ignored investigation of such parallelism, since the advance of scalar capabilities has been so reliable.

Future progress in many of the directions discussed in this document will require significantly more computation, and 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 (Janin 2004).

3. Models, Algorithms, and Search

3.1 Adaptation and Self-Learning in Speech Recognition System

Learning: Speech recognition has traditionally been cast as a classification problem where spoken input is classified into a sequence of pre-defined categories such as words (Jelinek 1976, Rabiner and Juang, 1993). Speech recognizer development typically proceeds via a heavily supervised training phase that makes use of annotated corpora, followed by a deployment (testing) phase whereby model parameters are possibly adapted to the environment, speaker, or topic etc. while the overall structure remains static. In other words, speech recognizers 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 (Bloomfield, 193, Chomsky, 1986 Lewis, 1975). Humans are able to digest large amounts of unlabeled or at best lightly annotated speech (Crain, 1991; Goodluck, 1991; Jusczyk, 1997). From these data we are able to learn, among other things, the phonetic inventories of a language, word boundaries, and we can use these abilities to acquire new words and meanings (Jusczyk, 1995; Pinker, 1994; Saffran, 2002). 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.

Research directed towards advancing learning capabilities in speech recognizers has many potential opportunities. There is a need for learning at all levels of speech and language processing to cope with changing environments, non-speech sounds, speakers, pronunciations, dialects, accents, words, meanings, and topics, to name but a few sources of variation. Research in these areas must address both the learning of new models, as well as the integration of such models into pre-existing knowledge sources. Thus, an important aspect of learning is being able to discern when something has been learned, and how to apply the result.

There are many degrees of learning ranging from “one shot” methods, to learning from small amounts of data, to learning from partially, poorly labeled or even un-annotated data (Pereira 1992). 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 the current dependency on high-quality annotations for training. This is especially true for languages for which there are little 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 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 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 has continually increased: from isolated words to continuous speech, from speaker-dependent to independent, from read to spontaneous speech, from clean to noisy, from utterance to content-based etc.

Although the complexity of such corpora has continually increased, one common property of such tasks is that they typically have a training partition 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 to improve speech recognition performance. This strategy is quite different from the human experience however. For our entire lives, we are exposed to all kinds of speech data from uncontrolled environments, speakers, and topics, (i.e., “every day” speech). Despite this 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 variation 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, different tasks, languages etc. One way to support research in this area would be to create a large corpus of “every day” 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 generalize to smaller resource languages and domains.

Machine Learning: This is an exciting time in the machine learning community. Many new machine-learning 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 (max 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 (Bloomfield, 1933; Chomsky, 1986; Lewis,1975), the degree of variation in languages across different cultures indicates that linguistic knowledge itself is acquired through interaction with and exposure to spoken language (Jusczyk, 1995; Pinker, 1994; Saffran, 2002) . Although there has been some research in unsupervised acquisition of phones, words, and grammars (Clark, 2001; Brill, 1993; Solan, 2004; Klein, 2005; de Marcken, 2001; Brent, 1999; Park, 2006; Venkataraman, 2001),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., audio-visual) corpora exist since alternative communication channels provide additional sources of constraint (Roy, 2002).

3.2 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 automatic speech recognition (ASR), language identification (LID) and speaker verification (Jelinek, 1997; Zissman, 1996; O'Shaugnessy, 1986). 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.

The acoustic environment of the speaker (e.g. background noise, reverberation, overlapping speech), and the channel through which the speech is acquired (e.g. cellular, land-line, VoIP; call-to-call variability).

The acoustic environment in which the speech is captured and the communication channel through which the speech signal is transmitted prior to its processing represent significant causes of harmful variability that is responsible for drastic degradation of system performance. Existing techniques such as Wiener filtering and cepstral mean subtraction (Rosenberg, et al, 1994) remove variability caused by additive noise or linear distortions, while methods such as RASTA (Hermansky and Morgan, 1994) 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 (e.g. age, nonnative accent) and speaking style (e.g. speech-rate, spontaneity of speech, emotional state of the speaker).

It is well known that speech characteristics vary widely among speakers due to many factors, including speaker physiology, speaker style, and accents – both regional and non-native. 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/speech types, but fail for new types.

Current ASR systems assume a pronunciation lexicon that models native speakers of a language and, furthermore, 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, (Leggetter and Woodland, 1994; Gauvain and Lee, 1994) and hybrid systems that combine these two approaches (Wang et al, 2003). Pronunciation variants have also been tried in the lexicon to accommodate accented speech (Humphries et al, 1996). Except for small gains, the problem is largely unsolved.

Similarly, some progress has been made for automatically detecting speaking rate from the speech signal (Morgan and Fosler-Lussier, 1998), but such knowledge is not exploited in the ASR system, mainly due to the lack of any explicit mechanism to model speaking rate in the recognition process.

Language characteristics (e.g. sublanguage or dialect, vocabulary, genre or topic of conversation).

Many important aspects of speaker variability have to do with 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 like Arabic, where the spoken dialects differ from the standard form dramatically (Kirchoff et al, 2003).

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 not considered either, merely the local collocation of words.

Another reason why language model adaptation to new domains and genre 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 re-estimated 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 re-estimated mutually independently, and so on.

With a few, exceptions such as vocal tract length normalization (VTLN) (Cohen et al, 1995) and cepstral mean subtraction (CMS) (Rosenberg et al, 1994) models used in today's speech systems have few explicit mechanisms to accommodate 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. 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.

3.3 Towards Robust Speech Recognition in Everyday Environments

Developing robust speech recognition 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.

1. 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 genre, etc. Such a corpus will be needed to construct several training/test partitions such that unseen conditions of various kinds are represented.
2. Multi-stream 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 entire ensemble of analyses, and using the analysis that happens to be the 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.
3. 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 à la speaker adaptive training (SAT) (Anastasakos, 1997) of acoustic models in ASR.
4. 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 multi-class MLLR, to a detailed transformation of every Gaussian density, to permit both robust transform estimation during training and unsupervised transform estimation from test data.
5. 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 abovementioned 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 etc.

6. 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 the speech signal 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.
7. 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 influences the language, vocabulary or topic of conversation will have to be used by the system.

Discriminative criteria (Bahl et al, 1986) for parameter estimation throughout the system and multi-pass recognition strategies, both being pursued today, will also be vital and are worthy of continued pursuit. 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 (Richardson et al, 1995).

8. 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.
9. 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 isn't, 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.

3.4 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 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 rely upon alternate knowledge sources only "locally" in the search process. The incorporation of multiple tiers of units (such as articulatory feature, sub-phone state, phone, syllable, word, and multi-word phrase) could have consequences for the search process. Finally, so-called "episodic" approaches to speech recognition are being investigated (Wachter 2003). 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 Hidden Markov Models, 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 to best incorporate these approaches into a search that also uses statistical models, which have already proven their worth.

V. Conclusions

This report surveys historically significant events in speech recognition and understanding which 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 are proposed for the next 3-5 year time horizon. 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. Speech recognition and understanding research achievements have provably contributed major advances to related human language technologies as well as pattern recognition, more generally.

To enable and implement these Grand Challenges, a number of especially promising research directions are outlined and supported in the Rich Areas for Future Research section. Though these are largely unfunded at present, the pursuit of these initiatives will contribute to a substantial increase in the core technology upon which robust future applications depend. Besides the members of the Speech Understanding Group itself, a number of other highly regarded speech researchers, nationally and internationally, have contributed both directly and indirectly to this report. A set of extensive References is provided as well.

Acknowledgements

The authors acknowledge significant informative discussions with several colleagues, whose opinions and advice is reflected in the text above. We wish to thank (in alphabetic order) Andreas Andreou, James Baker, Mary Harper, Hynek Hermansky, Frederick Jelinek, Damianos Karakos, Alex Park, Raj Reddy, Richard Schwartz and James West.

References

1) T. Anastasakos, J. McDonough, J. Makhoul. “Speaker Adaptive Training: a Maximum Likelihood Approach to Speaker Normalization,” *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp.1043-1046, Apr 1997.
2) G. Aradilla, J. Vepa, and H. Bourlard, “Improving Speech Recognition Using a Data-Driven Approach,” Proc. Eurospeech, Lisbon, Sept 2005, pp. 3333-3336.
3) K. Asanovic, R. Bodik, B. Christopher 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”,Technical Report UCB/EECS-2006-183, EECS Department, University of California at Berkeley, December 2006.
4) S. Axelrod and B. Maison. “Combination of hidden Markov models with dynamic time warping for speech recognition,” *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, 2004, pp. 173-176.
5) L. Bahl, P. Brown, P. de Souza, R. Mercer, Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition,” *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 49- 52, Apr 1986.
6) L. Bahl et al, “Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy”, IEEE Trans. Speech and Audio Processing, 5(2), 1993, pp. 179-190.
7) J. K. Baker, “Spoken Language Digital Libraries: The Million Hour Speech, Project” Invited Paper, International Conference on Universal Digital Libraries, Alexandria, Egypt, 2006.
8) J. K. Baker, “Stochastic Modeling for Automatic Speech Recognition”, in *Speech Recognition*, edited by D. R. Reddy, Academic Press, 1975.

9) L. Baum, "An Inequality and Associated Maximazation Technique Occurring in Statistical Estimation for Probabilistic Functions of a Markov Process," Inequalities, Vol III, 1972, pp. 1-8.

10) F. Beaufays, H. Bourlard, H. Franco, and N. Morgan, "Speech Recognition Technology," in Handbook of Brain Theory and Neural Networks, 2nd edition, M. Arbib ed., MIT Press, 2002.

11) L. Bloomfield. Language. Holt, New York, NY, 1933.

12) E. Brill. A Corpus-based approach to language learning. PhD thesis, University of Pennsylvania, Philadelphia, PA, December 1993.

13) M. R. Brent. "An efficient probabilistically sound algorithm for segmentation and word discovery." Machine Learning, 34(1-3):71–105, February 1999.

14) H. Bourlard, Y. Kamp, H. Ney, and C. Wellekens, "Speaker Dependent Connected Speech Recognition via Dynamic Programming and Statistical Methods", *Speech and Speaker Recognition*, M. Schroeder, ed., Bibliotheka Phonetica, Vol. 12 Kargers, Basel, 1988.

15) H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands." *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pages 426--429, Philadephia, Pennsylvania, October 1996.

16) L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Wadsworth & Brooks, Pacific Grove, CA, 1984.

17) J. Bridle, M. Brown, and R. Chamberlain, "A One-Pass Algorithm for Connected Word Recognition", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982, pp. 899-902.

18) N.A. Chomsky. Knowledge of Language: Is Nature, Origin, and Use. Praeger, New York, NY, 1986.

19) A. Clark. Unsupervised Language Acquisition: Theory and Practice. PhD thesis, University of Sussex, Brighton, UK, December 2001.

20) J. Cohen, T. Kamm and A.G. Andreou, "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability," 129th Meeting of the Acoustical Society of America, *Journal of the Acoustical Society of America*, Vol. 97, No. 5, pp. 3246-3247, May 1995.

21) S. Crain. Language acquisition in the absence of experience. Behavioral and Brain Sciences, 14(4):601–699, December 1991.

22) A.M. Dale and E. Halgren. "Spatiotemporal mapping of brain activity by integration of multiple imaging modalities," Curr. Opin. Neurobiology 11(2): 202-208., 2001.

23) C.G. de Marcken. Unsupervised Language Acquisition. PhD thesis, Massachusetts Institute of Technology, 1996.

24) A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, 39(1), 1977, pp. 1-21.

25) 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 Proc., Vol. 2, 1994, pp. 507-520.

26) L. Deng, D. Yu, and A. Acero. "Structured speech modeling," IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription), Vol. 14, No. 5, Sept 2006, pp. 1492-1504.

27) L. Deng and D. O'Shaughnessy, *SPEECH PROCESSING --- A Dynamic and Optimization-Oriented Approach*, Marcel Dekker Inc., New York, NY, 2003.

28) E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1996, pp. 346--349.

29) K. Farrell, R. Mammone, and J. Flanagan, "Beamforming microphone arrays for speech enhancement", Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992, pp.285-288.

30) C.J. Fillmore, C.F. Baker and H. Sato, "The FrameNet Database and Software Tools," In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC). Las Palmas, 2002, pp.1157-1160.

31) J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)", Proc. IEEE ASRU Workshop, Santa Barbara, CA, 1997, pp. 3477-3482

32) S. Furui. Digital Speech Processing, Synthesis and Recognition (Second Edition). Marcel Dekker Inc., New York, NY, 2001.

33) J. Frankel and S. King. "Speech recognition using linear dynamic models," IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, 2006.

34) J. Frankel and S. King, "ASR - Articulatory Speech Recognition," Proc. Eurospeech, pp. 599-602, Aalborg, Denmark, September 2001.

35) J-L. Gauvain and C-H. Lee. "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. Speech and Audio Processing, No. 7, 1997, PP. 711-720.

36) J. S. George, C. J. Aine, J.C. Mosher, D.M. Schmidt, D.M. Ranken, H. A. Schlitt, "Mapping function in the brain with magnetoencephalography, anatomical magnetic resonance imaging, and functional magnetic resonance imaging," J. Clinical Neurophysiology 12(5):406-431, 1995.

37) L. Gillick, Y. Ito, and J. Young. "A probabilistic approach to confidence measure estimation and evaluation". Proc. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997, pp. 879-882.

38) L. Gillick, James Baker, Janet Baker, John Bridle, Melvyn. Hunt, Y.. Ito, S. Lowe, J. Orloff, B. Peskin, R. Roth, F. Scattone, "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification

Using Telephone Speech," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 27-30, 1993, Vol. 2: pp. 471- 474.

39) J.R. Glass, A probabilistic framework for segment-based speech recognition. In: M. Russell, J. Bilmes (Eds.) Special issue on "New Computational Paradigms for Acoustic Modeling in Speech Recognition", Computer, Speech & Language, 2003, Vol. 17(2-3): 137-152.
40) B. Gold and N. Morgan, *Speech and Audio Signal Processing*, Wiley Press, 2000.
41) H. Goodluck. Language Acquisition. Blackwell Publishers, Cambridge, MA, 1991.
42) S. Hawkins. "Contribution of fine phonetic detail to speech understanding," Proc. of the 15th Int. Congress of Phonetic Sciences (ICPhS-03), Barcelona, Spain, pp. 293-296, 2003.
43) H. Hermansky and N. Morgan. "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, 1994.
44) H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech", Journal of the Acoustical Society of America, 1990, 87(4), pp. 1738-1752.
45) H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, June 2000.
46) X. D. Huang, A. Acero, and H. Hon, *Spoken Language Processing*, Prentice Hall, New York, 2001.
47) J. Huang and G. Zweig, "Maximum Entropy Model for Punctuation Annotation from Speech", Proc. INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 2002, pp. 917-920.
48) J. J. Humphries, P. C. Woodland, D. Pearce, Using accent-specific pronunciation modeling for robust speech recognition, Proc. International Conference on Spoken Language Processing, 1996.
49) IEEE History Center, Automatic Speech Synthesis and Recognition: http://www.ieee.org/organizations/history_center/sloan/ASSR/assr_index.html
50) A. Janin, "Speech Recognition on Vector Architectures", PhD dissertation, University of California at Berkeley, December 2004.
51) F. Jelinek. Statistical Methods for Speech Recognition, MIT Press, Cambridge, MA, 1997.
52) F. Jelinek, "Continuous Speech Recognition by Statistical Methods", Proceedings of the IEEE, 64(4), 1976, pp. 532-557.
53) F. Jelinek, "A Fast Sequential Decoding Algorithm Using a Stack", IBM Journal of Research and Development, 13, 1969, pp. 675-685.
54) L. Jiang and X.D. Huang, "Vocabulary-Independent Word Confidence Measure Using Subword Features," Proc. International Conference on Spoken Language Processing, Sydney, Australia, paper 401-404.
55) D. Jurafsky and J. Martin. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall, Upper Saddle River, N.J., 2000.
56) P.W. Jusczyk. The Discovery of Spoken Language. MIT Press/Bradford Books, Cambridge MA, 1997.
57) P.W. Jusczyk and R.N. Aslin. "Infants' detection of sound patterns of words in fluent speech." Cognitive Psychology, 29(1):1–23, August 1995.
58) H. Ketabdar and H. Hermansky, "Identifying unexpected words using in-context and out-of-context phoneme posteriors," Technical Report, IDIAP-RR 06-68, 2006.
59) P. Kingsbury and M. Palmer. "From Treebank to PropBank." In Proceedings of the LREC, Las Palmas, Canary Islands, Spain, 2002.
60) K. Kirchhoff, J. Bilmes, S. Das, N. Duta, M. Egan, J. Gang, H. Feng, J. Henderson, L. Daben, M. Noamany, P. Schone, R. Schwartz, D. Vergyri, Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 344-347, April 2003.
61) D. Klein. The Unsupervised Learning of Natural Language Structure. PhD thesis, Stanford University, Palo Alto, CA, March 2005.
62) A. Krishnamurthy and D. Childers, "Two Channel Speech Analysis", IEEE Trans. Acoustics, Speech and Signal Processing, 1986, 34, pp. 730-743.
63) R. Kuhn et al. "Eigenvoices for Speaker Adaptation," Proc. International Conference on Spoken Language Processing, 1998, Sydney, Australia, PP. 1771-1774.
64) N. Kumar and A. Andreou, "Heteroscedastic Analysis and Reduced Rank HMMs for Improved Speech Recognition", Speech Communications, vol 26, 1998, pp. 283-297.
65) K.-F. Lee. Automatic Speech Recognition: The Development of the Sphinx *Recognition System*, Springer-Verlag, 1988.
66) C.-H. Lee, F. Soong, and K. Paliwal (eds.), *Automatic Speech and Speaker Recognition -- Advanced Topics*, Kluwer Academic, 1996.
67) C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol. 9, 1995, pp 171-185.
68) C. Leggetter and P. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression", *Proc. of the International Conference on Spoken Language Processing*, pp. 451-454. 1994.
69) D. Lewis. Languages and language. In K. Gunderson, editor, Language, Mind, and Knowledge. University of Minnesota Press, Minneapolis, MN, 1975.

70) R. Lippman, “An Introduction to Computing with Neural Nets”, IEEE ASSP Magazine, 4(2), April 1987, pp. 4-22.

71) V. Maier, R.K. Moore. “An investigation into a simulation of episodic memory for automatic speech recognition,” Proc. of Interspeech-2005, Lisbon, 5-9 September 2005, pp. 1245-1248.

72) K. Marinkovic. “Spatiotemporal Dynamics of Word Processing in the Human Cortex,” Neuroscientist, Vol. 10(2): 142-152, 2004.

73) 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 Magazine, Sept 2005, pp. 81-88.

74) N. Morgan, E. Fosler-Lussier, Combining Multiple Estimators of Speaking Rate,” *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.729-732, May 1998.

75) H. Ney, The use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition”, IEEE Transactions on Acoustics, Speech and Signal Processing, 32: 1984, pp. 263-271.

76) 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,” Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, June 5-9, 2000, Istanbul, Turkey, I 229-232.

77) K. Olukotun, “A Conversation with John Hennessy and David Patterson,”, ACM Queue Magazine, pp 14-22, ACM Queue Magazine, December/January, 2006-2007.

78) D. O'Shaughnessy. Speaker recognition. *IEEE Acoustics, Speech and Signal Processing Magazine*, Vol. 3, No. 4, pp. 4-17, 1986.

79) 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 Proc., Vol. 4, 1996, pp. 360-378.

80) A. Park. Unsupervised Pattern Discovery in Speech: Applications to Word Acquisition and Speaker Segmentation, PhD thesis, MIT, Cambridge, MA, September, 2006.

81) D. Paul, “Algorithms for an optimal A* Search and Linearizing the Search in a Stack Decoder, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Vol. 1, 1991, pp.693-696.

82) F. Pereira and Y. Schabes. “Inside-outside re-estimation from partially bracketed corpora.” 30th Annual Meeting of the Association for Computational Linguistics, pages 128–135, Newark, Delaware, 1992. Association for Computational Linguistics.

83) S. Pinker. The Language Instinct. William Morrow and Company, New York, NY, 1994.

84) H. Poor, An Introduction to Signal Detection and Estimation, Springer Texts in Electrical Engineering, J. Thomas, ed., Springer-Verlag, N.Y. 1988.

85) A. Poritz, “Hidden Markov Models: A Guided Tour”, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1988, Vol. 1. pp. 1-4.

86) D. Povey, B, Kingsbury, L.Mangu, G. Saon, H. Soltau, and G. Zweig, “FMPE: Discriminatively Trained Features for Speech Recognition”, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, Philadelphia, PA.

87) L. Rabiner and B. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, N.J., 1993.

88) R. Reddy (Ed.), Speech Recognition., Academic Press, NY, 1975.

89) R. Reddy, J. Carbonell, M. Shamos, and G. St. Clair, “The Million Book Digital Library Project”, Computer Science Presentation, Carnegie Mellon University, Pittsburgh, PA., Nov. 5, 2003.

90) F. Richardson, M. Ostendorf and J. R. Rohlicek, Lattice-Based Search Strategies for Large Vocabulary Speech Recognition, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 576-579, May 1995.

91) A. E. Rosenberg ,C. H. Lee and F. K. Soong, “Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification,”*Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1835-1838, 1994.

92) D. Roy and A. Pentland. “Learning words from sights and sounds: A computational model.” Cognitive Science, 26(1):113–146, January 2002.

93) J.R. Saffran. Constraints on statistical language learning. Journal of Memory and Language, 47(1):172–196, July 2002.

94) S. Sakoe and S. Chiba, “A Dynamic Programming Approach to Continuous Speech Recognition”, Proc. of the Seventh International Congress on Acoustics, Budapest, Vol. 3, 1971, pp. 65-69.

95) Saras Institute History of Speech Technology Project: http://www.sarasinstitute.org/

96) Smithsonian Speech Synthesis History Project: http://www.mindspring.com/~ssshp_cd/ss_home.htm

97) Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised context sensitive language acquisition from a large corpus. In L. Saul, editor, Advances in Neural Information Processing Systems, Vol. 16, Cambridge, MA, 2004. MIT Press.

98) K. Stevens. *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1998.

99) H. Strik. “How to handle pronunciation variation in ASR: by storing episodes in memory?” Proc. ITRW on Speech Recognition and Intrinsic Variation (SRIV2006), Toulouse, France, May 2006, pp. 33-38.

100) J. Sun and L. Deng. “An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition,” Journal of the Acoustical Society of America, Vol. 111, No. 2, February 2002, pp.1086-1101.
101) V. Velichko and N. Zagoruyko, “Automatic Recognition of 200 Words”, International Journal of Man-Machine Studies, 2, 1970, p. 223.
102) A. Venkataraman. “A statistical model for word discovery in transcribed speech,” Computational Linguistics, 27(3):352–372, September 2001.
103) T. Vintsyuk, “Speech Discrimination by Dynamic Programming”, Kibernetika, 4(2), 1968, pp. 81-88.
104) A. Viterbi. “Error bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm”, IEEE Transactions on Information Theory, IT-13(2), 1967, pp. 260-269.
105) M. Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, “Data-driven example based continuous speech recognition,” Proc. EUROSPEECH, Geneva, Sept 2003, pp. 1133-1136.
106) M. Wachter, K. Demuynck, D. Van Compernolle, “Boosting HMM performance with a memory upgrade,” Proc. INTERSPEECH, Pittsburgh, Sept. 2006. pp. 1730-1733.
107) Z. Wang, T. Schultz, A. Waibel, Comparison of Acoustic Model Adaptation Techniques on Non-native Speech, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
108) S. Wilks, *Mathematical Statistics*, John Wiley and Sons, N.Y., 1962.
109) S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg, “Performance Improvements Through Combining Phone-and Syllable-scale Information in Automatic Speech Recognition,” Proc. INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING, 1998, Sydney, Australia, pp. 854-857.
110) M. Zissman, Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 1, January 1996.