# Eighty Challenges Facing Speech Input/Output Technologies

Victor Zue

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA
zue@csail.mit.edu

## ABSTRACT

During the past three decades, we have witnessed remarkable progress in the development of speech input/output technologies. Despite these successes, we are far from reaching human capabilities of recognizing nearly perfectly the speech spoken by many speakers, under varying acoustic environments, with essentially unrestricted vocabulary. Synthetic speech still sounds stilted and robot-like, lacking in real personality and emotion. There are many challenges that will remain unmet unless we can advance our fundamental understanding of human communication – how speech is produced and perceived, utilizing our innate linguistic competence. This paper outlines some of these challenges, ranging from signal presentation and lexical access to language understanding and multimodal integration, and speculates on how these challenges could be met.

## 1. INTRODUCTION

During the past three decades, we have witnessed remarkable progress in the development of speech input/output technologies. The ability to speak and listen to computers, as if they were human, no longer exists only in Hollywood fantasies and advanced research laboratories. Speech recognition error rates continue to fall steadily as task complexity increases, and the quality and intelligibility of computer-generated speech continue to improve. Today, our lives are touched almost daily by systems that can allow us to dial phone numbers, issue verbal commands, perform transactions, or even dictate a letter, all using the devices we are born with. Despite these successes, we are far from reaching human capabilities of recognizing nearly perfectly the speech spoken by many speakers, under varying acoustic environments, with essentially unrestricted vocabulary. Synthetic speech still sounds stilted and robot-like, lacking in real personality and emotion. How are we going to reach nirvana – enjoying truly anthropomorphic interfaces that can deal with us on our terms, using human language technologies? While mathematical formalisms, data collection from humans, and rigorous performance evaluations are key ingredients, there are many challenges that will remain unmet unless we can advance our fundamental understanding of human communication – how speech is produced and perceived, utilizing our innate linguistic competence. In this paper, I will outline some of these challenges, ranging from signal presentation and lexical access to language understanding and multimodal integration, and speculate on how these challenges could be met.

## 2. FUNDAMENTAL CHALLENGES

Many challenges confront us in our quest to achieve natural and effective speech-based human computer interfaces. These challenges easily can reach eighty in number. In the interest of brevity, however, I will outline only eight, one representing each decade in the life of the person

we have gathered at this conference to honor. The choice is entirely a personal one, influenced by my own experiences and biases.

**2.1 Signal Representation**

State-of-the-art speech recognition systems can often give good performance when the acoustic conditions are satisfactory, for example, when one uses a noise-canceling, head-mounted microphone in a quiet room. Remarkable recognition accuracy has also been achieved over the telephone for systems with a working vocabulary of several thousand words (Glass et al., 1999). However, these systems can break down dramatically in the presence of ambient noise, or when the user changes orientation to or distance from the microphone, as is often the case when a speakerphone is used.

To alleviate the problem of user movement in an un-tethered environment, one can resort to wireless microphones. But the solution is unwieldy, especially when multiple users are involved, as is the case for meeting transcription. A possible solution may lie in the use of microphone arrays using beam-forming techniques to capture the desired signal (e.g., Flanagan & Jan, 1996). As an example, Figure 1 shows the performance improvement as the number of microphones in the array increases from 1 to 1,000 (Weinstein et al., 2004).

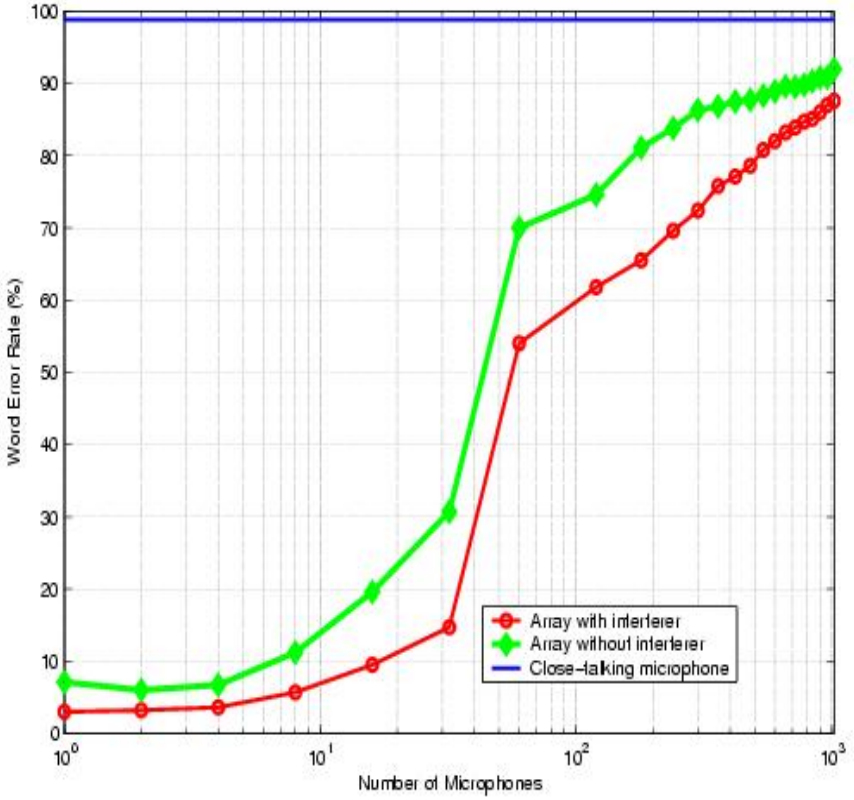


Figure 1. Recognition accuracy as a function of the number of microphones in a microphone array (Weinstein et al., 2004). There is a slight degradation when multiple competing speakers are present.

While the use of microphone arrays is a promising direction for achieving robust data capture, one can not help noticing that, in this instance at least, the number of transducers is a couple of orders of magnitude greater than what we are born with. Humans have a remarkable ability to recognize speech under extremely noisy conditions, a performance unmatched by current-day speech recognition systems (Lippmann, 1997). This observation has prompted some researchers to explore the use of auditory models as a recognition front-end (Seneff, 1988; Ghitza, 1995; Allen, 1995). These auditory front-ends typically yield similar performance to Fourier-based representations for clean speech (e.g., Meng, 1991). However, the auditory-based representations do achieve better performance when the speech signal has been corrupted by additive noise. Partly due to the computational costs, today's speech recognition systems have for the most part abandoned the auditory models in favor of a Mel-frequency cepstral representation that attempts to mimic some of the known properties of the human auditory system (Rabiner & Juang, 1993).

Continued research into the use of auditory models is essential if systems are to achieve human-level performance under varying acoustic conditions. However, these models must be extended to include binaural hearing, so that the system can better handle sound localization and cocktail party effects. As we acquire more knowledge about the decoding of linguistic information beyond the auditory periphery, we should be in a better position to increase the level of sophistication of the auditory models, leading to a better understanding of what attributes to extract, and how to utilize them for recognition.

### 2.2 Acoustic Modeling

By far the most popular approach for speech recognition is hidden Markov modeling, or HMM (Rabiner & Juang, 1993). From its humble beginning of modeling context-independent phonemes using discrete observations (Baker, 1975; Jelinek, 1976), HMM has grown in sophistication over the years. Today, context-dependent phones are routinely used to capture local contextual dependencies, and continuous density models are invariably the mechanism of choice to characterize the observations.

HMM-based speech recognition systems are frame-based, i.e., the acoustic observations are computed at a fixed rate, typically 10 to 20 times a second. It is well known that phonetic information is not encoded in the speech signal uniformly, and that acoustic landmarks contain most of the information for phonetic distinction (Stevens, 1998). Thus an alternative approach to speech recognition may be to first identify these landmarks (Glass & Zue, 1988), and then make measurements based on these landmarks. This approach, while different from the prevailing HMM-based approach, has been shown to be mathematically tractable, and to achieve competitive results (Glass, 2003).

Despite all these efforts, there is increasing evidence within the research community that the dominating factor for good performance is not acoustic modeling, but statistical language modeling. This has led to an explosion of research on how to derive good language models, at the sacrifice of research on how to better model the acoustics of the sub-word units. But reliance on the language models to "bail out" the recognition system has its perils. Since language models are highly dependent on the corpus from which they are trained, good recognition performance can only be achieved with a sufficient amount of text data to train the

models, and as such they are very task dependent. This has led to a costly development cycle, in which the need for a large corpus must first be satisfied.

It is quite possible that one can derive long term payoffs by increasing the level of sophistication in acoustic modeling from phonemes to larger units such as syllables. As I will argue, syllables can better capture the phonological constraints of a language than phonemes. As a side benefit, since the syllable inventory is quite limited in size, the amount of data one needs to develop acoustic and language models would be significantly less than that for word-based models. As a consequence, there will be less dependency on the specific corpus used for training, resulting in models that can yield robust performance across many recognition tasks.

Proper acoustic modeling means that one must implicitly or explicitly capture the contextual dependency of phonemes in words and sentences. We have known for quite some time that this context dependency can exist at various levels of the phonological hierarchy. Much of the contextual variation that one readily observes in American English is highly dependent on the syllable structure. For example, plosives have significantly reduced aspiration in a syllable-initial cluster with /s/, as in “speech,” but not in syllable-final position, as in “cusp.” Similarly, the phoneme /t/ is heavily aspirated and retroflexed in a syllable onset position followed by an /r/ (as in “nitrate,” but not in syllable-final position followed by an /r/, as in “night rate.” To properly capture acoustic manifestation of such context dependencies, the models must accommodate such a hierarchical structure. Pushed one step further, it is conceivable that one can develop a domain-independent recognition kernel, whose job it is to transcribe the speech signal into a set of syllables. Since the number of syllables in a given language is relatively small compared to the number of words, such a recognition kernel would not require a great deal of data to train. The resulting kernel could serve as a generic first-stage of a two stage recognizer, in which word-dependent (and thus task dependent) knowledge could be used to derive the final word sequences.

Recent research by Seneff and her colleagues (Chung & Seneff, 1998) has pursued this line of inquiry and achieved some promising results. In their system, syllable-level dependencies are captured in probabilistic context-free rules, and decoding is achieved through efficient parsing algorithms. Their two-stage recognition system achieved slightly better recognition results than a one-stage recognition system without explicit syllable-level constraints (Chung et al., 1999). However, a lot more work remains.

### 2.3 Lexical Access

Lexical access is the process of matching the continuous acoustic signal to the discrete lexical entries. Two important issues concerning lexical access are lexical representation and search strategy, the former being closely related to acoustic modeling. Traditionally, words in the lexicon are represented as phoneme sequences. Tri-phones, quardri-phones, or even quint-phones are employed for phonemes to capture context dependencies. However, this approach is inefficient for many of the kinds of coarticulation which can regularly appear in spoken language. For example, the potential rounding in /s/ due to the vowel in the word “strawberry" would require a *heptaphone* unit in current frameworks!

An alternative framework that may be more appropriate for describing such coarticulatory effects is one based on characterizing phonemes in terms of features (manner, place, voicing,

etc.) and feature bundles (Stevens, 1995). Such a framework might allow us to express context-dependencies in a succinct way. For example, if the manner and place contextual effects of a phone are treated separately, then far fewer models would be required, since there would be sharing of models among common classes. This framework could be especially powerful for expressing phonological transformations. In the previous example, the /s/ would acquire the "retroflexed" and "round'" features from the following syllable-internal sonorants due to anticipatory coarticulation. Such a feature-based framework is especially suitable to our landmark-based formulation, since we do not process the speech signal on a uniform frame-by-frame basis.

Words in the lexicon can be pronounced differently by different people. A word like "California" can be pronounced in five, four, or even three syllables (e.g., "Cal-for-nia.") At a word boundary, significant modifications could occur depending on the adjacent words. For example, the word-final /s/ in "gas" can be geminated (as in "gas station") or palatalized (as in "gas shortage,") depending on the context. Most speech recognition systems today do not explicitly model such phonological variations, but instead rely on context-dependent phone models to capture them. In a few systems (Zue et al., 1990; Hazen et al., 2002), phonological rules are used to expand the phonemic baseforms into pronunciation graphs, which are then searched during recognition. However, the resulting graph might be very bushy, thus increasing the number of hypotheses that must be examined, and consequently the likelihood of recognition errors. An appropriate probabilistic formulation of the phonological variations can improve this situation (Seneff & Wang, 2004). By utilizing partial feature representations, i.e., leaving some of the less reliable features unspecified, a simpler, albeit under-specified representation can be derived, which could be sufficient for lexical decoding. Alternatively, one can use such a broad class representation, together with phonotactic constraints, to initially whittle down the list of word candidates. These more similar words can then be distinguished using a more detailed analysis (Zue, 1983). Recently, this line of investigation has been revived and extended to continuous speech with promising results (Tang et al., 2003).

When examining the pronunciation graphs of words in a typical lexicon, as illustrated in Figure 2, one is often struck by the fact that the bushiest parts of the graphs typically involve reduced syllables. A possible interpretation of this observation is that the unstressed and reduced syllables are not produced with as much precision as stressed ones. As a consequence, there exists a lot of variability surrounding these syllables, as manifested by the many ways these syllables can be pronounced. If this were the case, then it makes little sense for a system to explicitly account for the variabilities by enumerating all the alternate pronunciations. It is possible that, to access a word like "California," the system should focus on the acoustic-phonetic properties of the first and third syllables, where the information is most reliable and thus least variable. The second and fourth syllable, on the other hand, may serve only as place holders, whose phonetic forms only need to be specified partially. This notion of *islands of reliability* suggests an island-driven lexical access strategy, in which the search is accomplished by anchoring on the stressed syllables. In this strategy, lexical decoding is not accomplished in a strict, left-to-right manner, as is the case with Viterbi or *A** algorithms (Rabiner & Juang, 1993). How such a search strategy can be formulated formally and implemented efficiently should be a topic of further research.

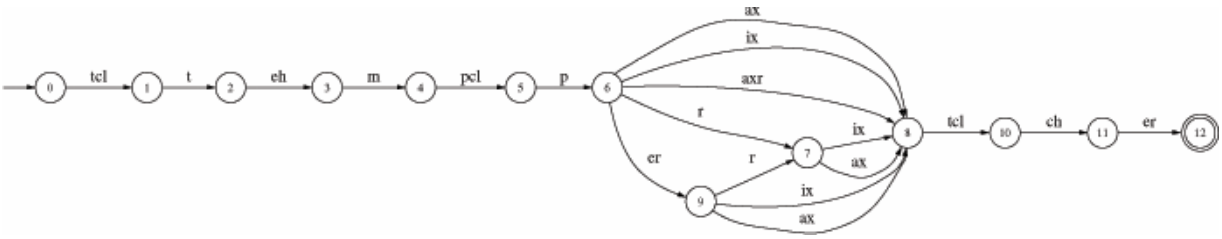


Figure 2. A pronunciation graph for the .word “temperature,” after phonological expansion has been performed on the lexical baseform.

### 2.4 Speech Understanding

Up until now, I have largely been focused on the problem of speech recognition, i.e., the conversion of the acoustic signal to a set of words. It is highly debatable whether we humans actually perform speech recognition *per se*, or we simply move directly from sound to meaning. After all, speech is mostly used to facilitate communication, and as such, understanding the meaning of an utterance should always be the final goal. If this is the case, then speech understanding, as opposed to speech recognition, should be the real task of interest. One might even argue that speech recognition only exists in the minds of the researchers, and is at best an interim problem to solve!

To achieve understanding, speech recognition must be coupled with language understanding. When the notion of developing speech understanding systems was first introduced, researchers did the obvious thing by cascading a speech recognition module and a natural language understanding module. However, they soon discovered that such an approach did not work well. Most of the natural language systems had been developed with text input in mind; it was usually assumed that the entire word string was known with certainty. This assumption is clearly false for speech input, where many words are competing for the same time span (e.g., "euthanasia" and "youth in Asia,") and some words are more reliable than others because of varying signal robustness (e.g., conjunctions and articles are often significantly reduced acoustically in conversational speech). Furthermore, spoken language contains no explicit punctuation and is often agrammatical, containing fragments, partial words, and disfluencies.

Thus, language understanding systems designed for text input needed to be modified in fundamental ways to accommodate spoken input. Natural language analysis has traditionally been predominantly syntax-driven -- a complete syntactic analysis is performed, in an attempt to account for *all* words in an utterance. However, when working with spoken material, researchers quickly came to realize that such an approach (Bobrow et al., 1990; Seneff, 1992a; Dowding et al, 1993) can break down dramatically in the presence of unknown words, novel linguistic constructs, recognition errors, and spontaneous speech events such as false starts. Furthermore, spoken language interaction is typically restricted to specific domains that make the notion of encoding semantics explicitly in the parse tree not only feasible but attractive, due to the additional knowledge that can be derived directly from the parse tree constituents.

Due to these considerations, many researchers have adopted semantic-driven approaches, at least for spoken language tasks in constrained domains. In such approaches, a meaning representation is derived by focusing on key words and phrases in the utterance, allowing words unimportant to meaning to be either skipped or handled through a traditional statistical language model. While this approach loses the constraint provided by syntax, and may not be able to adequately interpret complex linguistic constructs, the need to accommodate spontaneous speech input has outweighed these potential shortcomings. Examples of early systems that incorporate such *stochastic* modeling techniques can be found in (Jackson et al., 1991; Seneff, 1992b; Stallard & Bobrow, 1992; Gorin et al., 1997; Miller et al., 1994).

How should the speech recognition component interact with the natural language component in order to obtain the correct meaning representation? One of the most popular strategies is the so-called *N*-best interface (Chow & Schwartz, 1989), in which the recognizer proposes its best *N* complete sentence hypotheses one by one, stopping with the first sentence that is successfully analyzed by the natural language component. In this case, the natural language component acts as a filter on *whole sentence* hypotheses. Alternatively, competing recognition hypotheses can be represented in the form of a word graph (Hetherington, et al., 1993), which is more compact than an *N*-best list, thus permitting a deeper search if desired.

A competing, and seemingly more attractive, strategy is for the speech recognition and natural language components to be tightly coupled, so that only the acoustically promising hypotheses that are linguistically meaningful are advanced. For example, partial theories can be arranged on a stack, prioritized by score. The most promising partial theories are extended using the natural language component as a predictor of all possible next-word candidates; none of the other word hypotheses are allowed to proceed. Therefore, any theory that completes is guaranteed to parse. Researchers have found that such a tightly coupled integration strategy can achieve higher performance than an *N*-best interface, often with a considerably smaller stack size (Goodine et al., 1992; Goddeau, 1992; Ward, 1994; Moore et al., 1995). An attractive solution is to introduce tight-coupling with natural language understanding in the *second* stage, after the broad-class first stage system has significantly reduced the acoustic search space.

Compared to speech recognition, natural language understanding is still quite a knowledge intensive process. Most language understanding components require a great deal of human effort devoted to writing grammar rules. However, stochastic approaches that can learn the linguistic regularities automatically currently depend on a large domain-dependent corpus, often properly annotated with syntactic and semantic tags (Gorin et al., 1997; Miller et al., 1994; Papineni, 1998; He & Young, 2003). The necessity of a corpus for training leads to the chicken-and-egg problem of acquiring a domain-dependent corpus before the system is functional. Wizard-of-oz experiments are extremely costly in terms of human resources and are not really a practical option. Finally, one of the major challenges facing us is to formalize processes that can automate the discovery of linguistic facts.

Grammar induction, a machine-learning technique, has tremendous potential as a means for furthering the development and use of conversational systems. Most work in grammar induction has occurred outside of the spoken language research community. In (Starkie et al., 2002) and Wang & Acero, 2001), the focus is on applying grammar induction to create dialogue systems.

Assisted grammar induction is used to build the bridge between sample sentences and an application's ontology. In (Gavalda, 2000), users are allowed to interactively extend a basic hand-developed kernel grammar as extra-grammatical sentences are detected during use. (Meng & Siu, 2002) have investigated the use of grammar induction in unsupervised training of natural language parsers. Both of these investigations used typed text as input, although speech applications were envisioned. (Galescu et al., 1998) induce training material for a new domain by transducing from word-class mappings in an old domain to those appropriate for the new domain.

There is a symmetry between spoken language understanding and spoken language generation which leads to the parallel assumption that language generation and speech synthesis, on the output side of conversational systems, can benefit from a more closely coupled solution as well. For example, a research strategy that requires the language generation component to produce a surface form string representation of a sentence, and then subjects this text to redundant linguistic analysis in the speech synthesizer appears to be less than ideal. To extend these ideas even further, language generation should be intimately coupled with *dialogue modeling* as well, especially for applications over the phone. For example, if the amount of information is too much to be delivered verbally to the user, clarification sub-dialogue may be necessary to help the system narrow down the choices before uttering them. Finally, planning what to say is highly dependent on factors such as whether the user has access to a graphical interface where information can be presented as text or displayed on a map or in a table.

**2.5 Dealing with New Words**

The traditional approach to spoken language recognition and understanding research and development is to define the working vocabulary based on domain-specific corpora. However, experience has shown that, no matter how large the size of the training corpora, the system will invariably encounter previously unseen words (Hetherington & Zue, 1991). This is illustrated in Figure 3. For the Air Travel Information System, or ATIS, task, for example, a 100,000-word training corpus will yield a vocabulary of about 1,000 words. However, the probability of the system encountering an unknown word, is about 0.002. Assuming that an average sentence contains 10 words, this would mean that approximately one in 50 sentences will contain an unknown word.

In a *real* applications such as Electronic Yellow Pages, a much larger fraction of the words uttered by users will not be in the system's working vocabulary. This is unavoidable partly because it is not possible to anticipate all the words that all users are likely to use, and partly because the database is usually changing with time (e.g., new restaurants opening up). In the past, researchers have not paid much attention to the unknown word problem because the tasks we have chosen assume a closed vocabulary. In the limited cases where the vocabulary has been open, unknown words have accounted for a small fraction of the word tokens in the test corpus. Thus researchers could either construct generic “trash word" models and hope for the best, or ignore the unknown word problem altogether and accept a small penalty on word error rate. In real applications, however, the system must be able to cope with unknown words simply because they will always be present, and ignoring them will not satisfy the user's needs – if a person wants to know how to go from the train station to a restaurant whose name is unknown to the system, they will not settle for a response such as, “I am sorry I don't understand you. Please rephrase the question."

For a system to be truly helpful, it system must be able not only to *detect* new words, taking into account acoustic, phonological, and linguistic evidence, but also to adaptively *acquire* them, both in terms of their orthography and linguistic properties. In some cases, fundamental changes in the problem formulation and search strategy may be necessary.

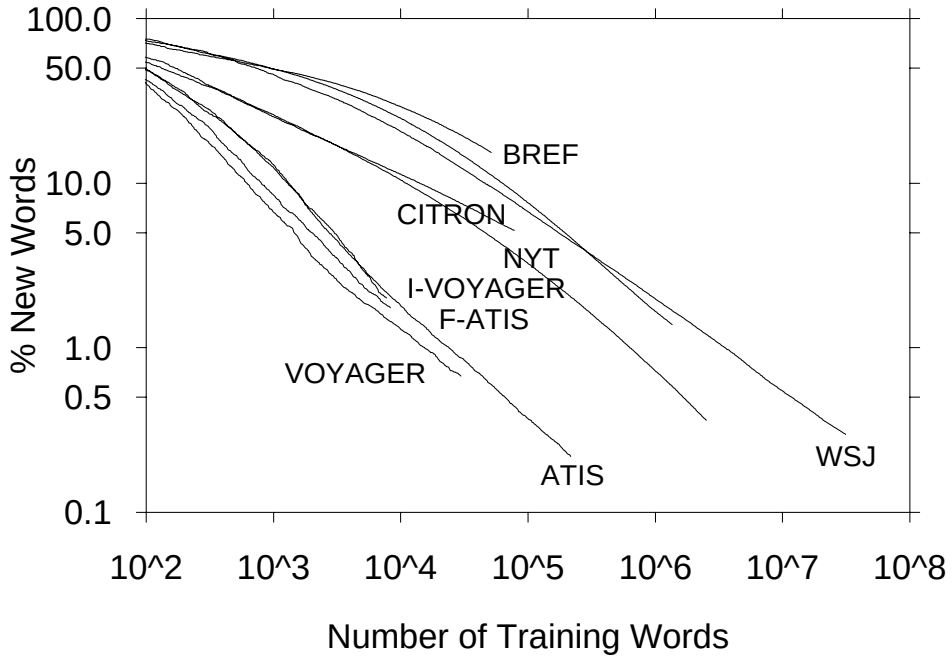


Figure 3. The percentage of unknown words in previously unseen data as a function of the size of the training corpora used to determine the vocabulary empirically. The sources of the data are: F-ATIS = French ATIS; I-VOYAGER = Italian VOYAGER; BREF = French La Monde; NYT = New York Times; WSJ = Wall Street Journal; and CITRON=Directory Assistance.

What is needed, then, is a generic capability to handle unknown or poorly recognized known words, beginning with detection, continuing on to disambiguation sub-dialogue, and terminating with an automatic update of the system such that it now knows the new word explicitly and understands its usage. (Chung et al., 2003) have recently demonstrated the ability to automatically enroll a new user's name through a speak-and-spell mode, where the orally spoken and spelled information are efficiently combined via a letter-to-sound model. The resulting spelling and pronunciation are then automatically incorporated in real time into the system's working vocabulary.

Further research is required in the integration of new word acquisition into the situational context of uncertainty in understanding. It is a research problem to be able to judge when it is appropriate to *distrust* proposed words, and therefore launch a sub-dialogue to disambiguate

and finally resolve confusion. (Filisko & Seneff, 2004) presents an initial framework to support such a sub-dialogue interaction to detect unknown words and solicit spellings from the user.

**2.6 Dialogue Interactions**

The dialogue modeling[1] component of a conversational system manages the interaction between the user and the computer. The technology for building this component is one of the least developed and most under-appreciated aspect of human language technology research, especially for mixed-initiative dialogue systems considered in this paper. Although there has been some theoretical work on the structure of human-human dialogue (Grosz & Sidner, 1986), this has not led to effective insights for building human-machine interactive systems. There is also considerable debate in the research communities about whether modeling human-machine interactions after human-human dialogues is necessary or appropriate (e.g., Thomson & Wisowaty, 1999; Sadek, 1999; Boves & den Os, 1999)

In the early stages of the conversation, the role of the dialogue manager might be to gather information from the user, possibly clarifying ambiguous input along the way, so that, for example, a complete query can be produced for the application database. The dialogue manager must be able to resolve ambiguities that arise due to recognition error (e.g., “Did you say Boston or Austin?") or incomplete specification (e.g., “On what day would you like to travel?"). In later stages of the conversation, after information has been accessed from the database, the dialogue manager might be involved in some negotiation with the user to help narrow down the number of choices to digestible chunks (e.g., “I found ten flights, do you have a preferred airline or connecting city?"). In addition, the dialogue manager must also inform and guide the user by suggesting subsequent sub-goals (e.g., “Would you like me to price your itinerary?"), offer assistance upon request, help relax constraints or provide plausible alternatives when the requested information is not available (e.g., “I don't have sunrise information for Oakland, but in San Francisco ..."), and initiate clarification sub-dialogues for confirmation.

Many current systems use a type of scripting language as a general mechanism to describe dialogue flow (e.g., Carlson & Hunnicutt, 1999; Lau et al., 1997; Souvignier, 2000). Other systems represent dialogue flow by a graph of dialogue objects or modules (e.g., Bernard et al., 1999; Sutton et al, 1999). In nearly all cases, the design of the dialogue strategy is typically hand-crafted by the system developers, and as such is largely based on their intuition about the proper dialogue flow. This can be a time-consuming process, especially for mixed-initiative dialogues, whose result may not generalize to different domains. A major challenge is how to develop *generic* dialogue modeling techniques so that the human effort invested into the development of one application domain can easily be reused in other domains (Glass & Seneff, 2003). There has also been some research recently exploring the use of machine learning techniques to automatically determine dialogue strategy (Levin et al., 1998). Regardless of the approach, however, there is the need to develop the necessary infrastructure for dialogue research. This includes the collection of dialogue data, both human-human and human-

[1] In the context of this paper, we define dialogue modeling as the preparation, for each turn, of the system's side of the conversation, including verbal, tabular, and graphical response, as well as any clarification requests.

machine. These data will need to be annotated, after developing annotation tools and establishing proper annotation conventions. In the last decade, speech recognition and language understanding communities have benefited from the availability of large, annotated corpora. Similar efforts are clearly needed for dialogue modeling.

Many research issues concerning dialogue modeling remain active areas of future research. These include 1) the ability for systems to offer context-dependent help mechanism in order to assist users to stay within the capabilities of the system 2) the recovery from the inevitable misunderstandings that a system will make, 3) in the introduction of back channel communication in spoken dialogue responses, in order to make the interaction more natural, and 4) the handling of interruptions by allowing the user to "barge in'' over the system response (Zue & Glass, 2000).

A critical roadblock to widespread use of spoken dialogue systems today is the fact that systems are usually configured with a static lexicon and fixed language model. We must develop a capability for an existing system to dynamically reconfigure itself over the course of a single dialogue interaction, by constantly readjusting its vocabulary and language model to reflect the situational context, based on (1) the topic of conversation, (2) specific entities retrieved from the database, and (3) direct interaction with the user to acquire new knowledge.

In spite of the progress made thus far, much challenging work still remains in building a system that can demonstrate the ability to dynamically learn and adopt *any* new words that appear at *any time* at the spoken input. Some of the research issues that need addressing include: (1) high quality letter-to-sound modeling, both for lexicon development and new word acquisition, (2) development of a dialogue interaction module to guide the user and the system through the new word acquisition process, and (3) development of a recognizer framework to support flexible vocabulary and dynamically introduce dialogue context dependent vocabularies and language models.

Spoken dialogue systems can behave quite differently depending on what input and output modalities are available to the user. In a telephone conversation where display is not available, it might be necessary to tailor the dialogue so as not to overwhelm the user with information. When displays are available however, it may be more desirable to simply summarize the information to the user, and to show them a table or image etc. Similarly, the nature of the interaction will change if alternative input modalities, such as pen or gesture, are available to the user. Which modality is most effective will depend, among other things, on the environment, user preference, and perhaps dialogue state (Oviatt, 1996).

### 2.7 Multimodal Integration

While speech is the most natural, effortless, and efficient way for humans to communicate, it is not the only way. In daily interactions, we often rely on pointing, gesture, and writing to augment speech. There are certainly occasions when speech would not be appropriate, as when we attempt to take notes during a meeting. To provide a full range of interactions, modalities such as pen and gesture should be included to augment and complement speech.

Interpreting multimodal inputs poses several challenges. First, the multiple inputs need to be understood in the proper context. When someone says, "What about that one?" while pointing

at an item on the shelf, the system must interpret the indirect referencing in the speech signal using information in the visual channel. In some cases, timing information may be crucial. As illustrated in Figure 4, proper interpretation of the object and the target location may depend on the system's ability to correlate the information in the acoustic and the visual channels. In addition, the system must be able to handle uncertainties, since object recognition can be error prone.

Past research on multi-modal understanding has focused primarily on the integration of speech and pen-based gesture, and as such is event driven, i.e., the pen activity is registered by clicking. By continuously tracking speech, gesture, and gaze activities, maintaining relative timing information on each channel, and using context to resolve conflicts, one can hopefully achieve robust multi-modal understanding. Proper modality selection can significantly improve information presentation. For example, a presentation might choose a graphical modality for numeric data, a speech synthesizer to deliver breaking news; and textual summaries for more detailed descriptions.

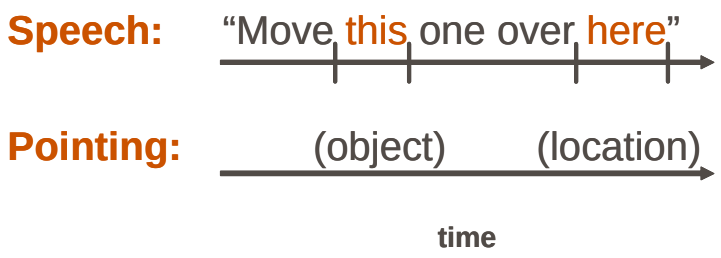


Figure 4. A schematic plot, as a function of time, of the sequence of words determined by a speech recognizer, and the interpretation of the object and its target location determined from the visual signal.

On the output side, a multimodal interface must be able to generate natural speech and integrate it in real-time with facial animation, in the context of a larger conversation. For intuitive dialogs, the system should support user interruption, back-channel, and other cues that are common in human dialogs.

For the multi-modal interaction to be effective, we must develop a unifying linguistic formalism that can describe multi-modal interactions (e.g., "Move it from here to here"), along with integration and delivery strategies. For output generation, the system must decide when to use which modality, a decision that could be based on the user's cognitive load. The system will also need to handle trans-modal interactions, in which one mode is transformed into another (verbally summarizing a weather map, for example).

### 2.8 Paralinguistics

The speech signal is primarily used for communicating linguistic information. However, it also conveys extra-linguistic information such as the identity of the speaker, and the physiological and psychological states of the user – whether (s)he is stressed, tired, sick, or agitated). Such

paralinguistic information is often useful for verbal communication. For example, knowing the identity, or even just the gender, of the speaker can lead to better recognition performance, since the system can then adopt gender-specific or speaker-specific acoustic models. Knowing that the user is agitated, perhaps due to repeated system errors, can potentially allow the system to be a little more accommodating in its management of the spoken dialogue. For spoken language generation, it is important for the system to use extra-linguistic information to convey the emotion that is appropriate for the condition at hand.

Technology for speaker identification and verification has been pursued by many researchers in the past, and impressive results have been achieved (Reynolds, 1995; Reynold et al., 2000; Doddington et al., 2000). Recent research suggests that, when speaker identification is conducted within the context of speech recognition, one can exploit knowledge about the identity and location of the phonemes to further improve speaker identification results (Park & Hazen, 2002). Finally, speaker recognition error rate can be reduced nearly ten-fold when the system is augmented with a face recognition module (Hazen et al., 2003). It is likely that our ability to identify a person from the speech signal can be further improved if other sources of information, such as gait, are also included.

In recent years, there has been increasing interest in the manifestation of emotion in the speech signal, ranging from corpus creation (Campbell, 2000) and theoretical foundations (Cornelius, 2000) to acoustic analysis (Cowie, 2000, Jackson et al., 2003) and actual recognition (Moriyama & Ozawa, 1999). While some encouraging results have been obtained, we are far from understanding the acoustic encoding of a full range of emotions, and do not yet have the ability to reliably detect and utilize them. Continuing research in this fertile area is necessary

## 3. CONCLUDING REMARKS

This paper has outlined some of the major challenges facing the development of speech input and output technologies that can one day lead to the realization of a natural and effective interface between humans and machines. While it is always important to have the right mathematical formalism for algorithm development, the available data for analysis, system development, and training, and the mechanism for rigorous evaluation, I believe it is also important for system developers to have a deep understanding of the ways humans communicate: how speech is produced, perceived, and eventually understood. Only through such understanding can we expect to realize systems with capabilities approaching those of humans.

## 4. ACKNOWLEDGEMENTS

The primary goal of this conference is to bring together scientists and engineers with diverse background to review progress made in all aspects of spoken language research over the past five decades, in the hope of promoting cross-fertilization that will lead to benefits for future work. A not-so-hidden secondary agenda is to pay tribute to Professor Kenneth Stevens, who has had an incredibly illustrious career over the past fifty years at MIT, conducting research in all the sub-disciplines covered in this conference.

I came to MIT in 1970 as a graduate student, which has led thus far to an association with Ken of more than 34 years, a journey far from over. The personal rewards in the intervening years have been many. Like many others before and after me, I learned the beauty and the diverse

nature of human language. He taught us the importance of relating speech sounds to articulatory gestures, and ultimately to the underlying linguistic features. It was under his tutelage that I took on the “hobby” of spectrogram reading, which has led to many insights on acoustic phonetics. We as Ken's students learn, through his personal example, to set high goals, to constantly expand our horizons, to approach problems with rigor and from multiple perspectives, to confront success and failure with humility, and above all, to devote ourselves to the education of students. I would like to acknowledge the impact Ken has made on my career, and the career of my students and their students.

Much of my work over the past fifteen years, and the insights gained, have benefited from my association with past and present members of the Spoken Language Systems Group, first at the Laboratory for Computer Science and now at the Computer Science and Artificial Intelligence Laboratory, especially Stephanie Seneff and Jim Glass. Stephanie Seneff also read an earlier version of this paper and offered valuable comments.

This work is supported by an industrial consortium of the MIT Oxygen Project.

**5. REFERENCES**

Allen, J. (1995) How do humans process and recognize speech, in Ramachandran & Mammone, eds., *Modern Methods of Speech Proceessing*. 251-275, Kluwer.

Baker, J. K. (1975) The Dragon system – an overview*, IEEE Trans. ASSP*, ASSP-23(1): 24-29.

Bobrow, R., Ingria, R., & Stallard, D. (1990) Syntactic and semantic knowledge in the DELPHI unification grammar, *Proc. DARPA Speech and Natural Language Workshop*, 230-236.

Boves L. & den Os, E. (1999) Applications of Speech Technology: Designing for Usability, *Proc. IEEE Worshop on ASR and Understanding*, 353-362.

Campbell, N. (2000) Databases of emotional speech, Proc. ESCA Workshop on Speech and Emotion.

Carlson R. & Hunnicutt, S. (1996) Generic and domain-specific aspects of the Waxholm NLP and dialogue modules, *Proc. ICSLP*, 677-680.

Chow, Y. & Schwartz, R. (1988) The N-Best algorithm: an efficient procedure for finding top N sentence hypotheses, *Proc. ARPA Workshop on Speech and Natural Language*, 199-202.

Chung, G. & Seneff, S. (1998) Improvements in speech understanding accuracy through the integration of hierarchical linguistic, prosodic, and phonological constraints in the jupiter domain*, Proc. ICSLP* 935-939.

Chung, G., Seneff, S. & Hetherington, I. (1999) Towards multi-domain speech understanding using a two-stage recognizer, *Proc. Eurospeech*, 2655-2658.

Chung, G. Seneff, S. & Wang, C (2003) Automatic acquisition of names using speak and spell mode in spoken dialogue systems, *Proc. HLT-NAACL*, 32-39.

Cornelius, R. (2000) Theoretical approaches to emotion, Proc. ESCA Workshop on Speech and Emotion.

Cowie, R. (2000) Theoretical Describing the emotional states expressed in speech, *Proc. ESCA Workshop on Speech and Emotion.*

Doddington, G.R., Pryzbocki, M, Martin, A.F., & Reynolds, D.A. (2000) NIST speaker recognition evaluation: overview, methodology, systems, results, perspective, (invited paper) *Speech Communication*.

Dowding, J., Gawron, J., Appelt, D., Bear, J., Cherny, L., Moore, R., & Moran, D. (1993) Gemini: a natural language system for spoken language understanding, *Proc. ARPA Workshop on Human Language Technology*, 21-24.

Filisko, E. & Seneff, S. (2004) Error detection and recovery in spoken dialogue systems, *Proc. Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Knowledge for Speech Processing.*

Flanagan, J. L., & and Jan, E. E. (1996) Sound capture form spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors, Proc, ICASSP.

Galescu L., Ringger E. & Allen, J. (1998) Rapid Language Model Development for New Task Domains, *Proc. LREC '98*.

Gavalda, M. (2001*) Growing Semantic Grammars*, PhD Thesis, Carnegie Mellon University.

Ghitza, O. (1995) Auditory models and human performance in tasks related to speech coding and speech recognition, in Ramachandran & Mammone, eds., *Modern Methods of Speech Proceessing*. Kluwer.

Glass, J.R., & Zue, V. (1988) Multilevel Acoustic Segmentation of Continuous Speech, *Proc. ICASSP,* 429- 432.

Glass, J. R., Hazen, T. J., & Hetherington, I. L. (1999) Real-time telephone-based speech recognition in the JUPITER domain, *Proc. ICASSP*.

Glass, J. (2003) A Probabilistic Framework for Segment-Based Speech Recognition, *Computer Speech and Language* 17, pp. 137-152

Glass, J. & Seneff, S. (2003) Flexible and personalizable mixed-initiative dialogue systems, *Proc. HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*.

Goddeau, D. (1992) Using probabilistic shift-reduce parsing in speech recognition systems, *Proc. ICSLP*, 321-324.

Goodine, D., Seneff, S., Hirschman, L., & Phillips, M. (1991) Full integration of speech and language understanding in the MIT spoken language system, *Proc. Eurospeech*, 845--848.

Gorin, A., Riccardi, G., & Wright, J. (1997) How may I help you?, Speech Communication, 23, 113-127.

Grosz B. & Sidner, C. (1990) Plans for discourse, in *Intentions in Communication*. MIT Press.

Hazen, T. J., Hetherington, I.L., Shu, H., & Livescu, K, (2002) Pronunciation modeling using a finite-state transducer representation, *Proc. PMLA Workshop*, 99-104.

Hazen, T. J., Weinstein, E., Kabir, R., Park A. & Heisele, B. (2003) Multi-modal face and speaker identification on a handheld device, *Proc. of the Workshop on Multimodal User Authentication*, 113-120.

He Y. & Young, S. (2003) A Data-driven spoken language understanding system," *Proc. ASRU '03*, 583-588.

Hetherington L. & Zue, V. (1991) new words: implications for continuous speech recognition, *Proc. Eurospeech*, 475-931.

Hetherington, L., Phillips, M., Glass, J., & Zue, V. (1993) A" word network search for continuous speech recognition, *Proc. Eurospeech*, 1533--1536.

Jackson, E., Appelt, D., Bear, J., Moore, R., & Podlozny, A. (1991) A template matcher for robust NL interpretation, *Proc. DARPA Speech and Natural Language Workshop*, 190-194.

Jelinek, F. (1976) Continuous speech recognition by statistical methods, *Proc. IEEE*, 24: 532-536.

Lau, R., Flammia, G., Pao, C., & Zue, V. (1997) WebGalaxy - Integrating Spoken Language and Hypertext Navigation, *Proc. Eurospeech*, 883-886.

Levin, E., Pieraccini, R., & Eckert, W. (1998) Using Markov decision process for learning dialogue strategies, *Proc. ICASSP*, 201-204.

Lippmann, R. (1997) Speech perception by humans and machines, *Speech Communication*, 22(1), 1-15.

Meng, H. (1991) *A Comparison of Auditory-based Representations of Speech and the Use of Distinctive Features as an Intermediate Representation for Automatic Speech Recognition*, S.M. Thesis, Dept. EE & CS, MIT.

Meng M. & Siu, K. (2002) Semiautomatic acquisition of semantic structures for understanding domain-specific natural language queries, *IEEE Trans. Knowledge and Data Engineering*, Vol. 14, No. 1, 172-181.

Miller, S., Schwartz, R., Bobrow, R., & Ingria, R. (1994) Statistical language processing using hidden understanding models," *Proc. ARPA Speech and Natural Language Workshop*, 278-282.

Moore, R., Appelt, D., Dowding, J., Gawron, J., & Moran, D. (1995) Combining linguistic and statistical knowledge sources in natural-language processing for ATIS, Proc. ARPA Spoken Language Systems Workshop, 261-264.

Moriyama, T. & Ozawa S. (1999) Emotion recognition and systhesis system on speech, Proc. IEEE ICMCS.

Papineni, K., Roukos, S. & Ward, R. (1998) Maximum likelihood and discriminative training of direct translation models*, Proc. ICASSP*, 189-192.

Park A. & Hazen, T. J. (2002) ASR dependent techniques for speaker identification, Proc. ICSLP.

Rabiner, L. R. & Juang, B.H. (1993) *Fundamentals of Speech Recognition*, Prentice Hall.

Reynolds, D. (1995) Speaker identification and verification using Gaussian mixture speaker models, *Speech Communications*, 17(1-2), 91-108.

Reynolds, D. A., Quatieri, T. F., & Dunn, R.B. (2000) speaker verification using adapted Gaussian mixture models, *Digital Signal Processing Review Journal*.

Sadek, D. (1999) Design considerations on dialogue systems: from theory to technology - the case of Artimis, Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems, 173-188.

Seneff, S. (1988) A joint synchrony/mean-rate model of auditory speech processing", J. Phonetics, Vol. 16, pp. 55-76.

Seneff, S. (1992a) TINA: A natural language system for spoken language applications, *Computational Linguistics*, 18(1), 61-86.

Seneff, S. (1992b) Robust parsing for spoken language systems, *Proc. ICASSP 92*, 189-192.

Souvignier, V., Kellner, A., Rueber, B., Schramm, H., & Seide, F. (2000) The thoughtful elephant: strategies for spoken dialogue systems, *IEEE Trans. SAP*, 8(1), 51-62.

Stallard, D. & Bobrow, R. (1992) Fragment processing in the DELPHI system, *Proc. DARPA Speech and Natural Language Workshop*, 305-310.

Starkie, B., Findlow, G., Ho, K., Hui, A., Law, L. Lightwood, L. Michnowicz, S., & Walder, C. (2002) Lyrebird: Developing Spoken Dialog Systems using Examples, *Proc. ICGI*, 309-211.

Stevens, K. N. (1995) Applying phonetic knowledge to lexical access, *Proc. Eurospeech,* 3-11

Stevens, K. N. (1998) *Acoustic Phonetics*, Cambridge, MA: MIT Press.

Sutton, S., et al., (1998) Universal Speech Tools: The CSLU Toolkit*, Proc. ICSLP*, 3221-3224.

Tang, M., Seneff, S. & Zue, V. (2003) Modeling Linguistic Features in Speech Recognition, *Proc. Eurospeech*, 2585-2588.

Ward, W. (1990) The CMU air travel information service: understanding spontaneous speech, *Proc. ARPA Workshop on Speech and Natural Language*, 127-129.

Ward, W. (1994) Integrating Semantic Constraints into the SPHINX-II Recognition Search, *Proc. ICASSP*, II-17-20, 1994.

Thomson, D. & Wisowaty, J. (1999) User confusion in natural language services, *Proc. ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, 189-196.

Wang A. & Acero, A. (2001) Grammar Learning for Spoken Language Understanding, *Proc. ASRU Workshop*.

Weinstein, U., Steele, K., Agarwal, A., & Glass, J. R. (2004) personal communication.

Zue, V. W. (1983) Proposal for an Isolated-Word Recognition System Based on Phonetic Knowledge and Structural Constraints, *Proc. of the Tenth International Congress of Phonetic Science*, 299-305.

Zue, V. W., Glass, J. R., Phillips, M., Polifroni, J. & Seneff, S. (1990) The SUMMIT speech recognition system: phonological modeling and lexical Access," *Proc. ICASSP 92*, 49-52.

Zue, V., & Glass, J., (2000) Conversational interfaces: advances and challenges, *Proc. IEEE*, *Special Issue on Spoken Language Processing*, Vol. 88, No. 11, 1166-1180.