On Organic Interfaces

Victor Zue

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139 USA
zue@csail.mit.edu

Abstract

For over four decades, our research community has taken remarkable strides in advancing human language technologies. This has resulted in the emergence of spoken dialogue interfaces that can communicate with humans on their own terms. For the most part, however, we have assumed that these interfaces are static; it knows what it knows and doesn't know what it doesn't. In my opinion, we are not likely to succeed until we can build interfaces that behave more like organisms that can learn, grow, reconfigure, and repair themselves, much like humans. In this paper, I will argue my case and outline some new research challenges.

Index Terms: speech-based interfaces, dialogue systems

1. Introduction

Speech is the most natural means for humans to communicate; nearly all of us can talk and listen to one another without special training. It is flexible; it can free our eyes and hands to attend to other tasks. Speech is also very efficient; one can typically speak several times faster than one can type or write. Nowadays, with the pervasiveness of landline, cellular, and internet phones, speech is also one of the most inexpensive ways for us to communicate. It is therefore not surprising that speech-based interfaces are in the minds of every techno-visionary and science fiction writer or movie producer.

Over the past four decades, we have witnessed remarkable progress in the development of speech input/output technologies. The introduction in the early seventies and the subsequent wide-spread use of the stochastic techniques known as Hidden Markov modeling [39, 25, 24] have resulted in a continuous reduction in word error rate while the task complexity continues to grow [37]. Similarly, the intelligibility and naturalness of synthetic speech has also improved with time, thanks to corpus-based techniques and the availability of large corpora [42]. We are beginning to see the emergence of spoken dialogue systems that understand and respond to spoken queries [13, 58, 50]. In some cases, compelling systems are being demonstrated in research laboratories around the world, combining speech with pen, gesture, and other modalities [38, 20]. Today, our lives are touched almost daily by systems that can allow us to dial phone numbers, issue verbal commands, perform transactions, or perhaps dictate a letter, all using the devices we are born with. As proud as we should be about the progress that this community continues to make, however, we are far from reaching human capabilities of recognizing and understanding nearly perfectly the speech spoken by many speakers, under varying acoustic environments, with essentially unrestricted vocabulary. Synthetic speech still sounds stilted at times, and lacking in real personality and emotion.

In this paper, I intend to offer a perspective drawn from evolution in computer science, one that views complex systems as living organisms that can learn, grow, reconfigure, and repair themselves. I would argue that such a perspective can lead us to the type of interface that is truly anthropomorphic. I will focus my discussions on spoken dialogue systems, namely, systems that integrate several human language technologies – speech recognition/synthesis, language understanding/generation, and discourse/dialogue modeling – that can help users solve problems incrementally and interactively. In the next section, I will outline what I mean by *Organic Interfaces* and describe some of their properties. This will be followed by a discussion of a few challenges, along with examples to illustrate my points. Some of the ideas are admittedly half-baked. In many ways, more questions are raised than are answered. It is my hope that this paper will trigger some discussion among us that will lead to further refinements of the ideas and perhaps engender new directions in our collective research agenda.

2. Organic Interfaces

It is perhaps informative to first examine the evolution of computing and computer science as a discipline [28]. Early computers were used primarily for computing of static functions under static conditions, whose specifications were well understood (e.g., billing, inventory management, medical records, etc.) The foundation of computer science is based on logic, and there is a clear notion of correctness. However, computers today are fast becoming pervasive, and are often required to operate under dynamic environments that are constantly changing. Computation is often augmented with functions such as communication, sensing, and control. Data of all kinds (audio, video, sensor readings, natural language text) from noisy, distributed, and unreliable sources have to be processed and fused. Increasingly, interaction with humans is a key aspect of the computers' functionality. Since we cannot know the details of the environments in which they will be deployed, nor the behavior of the human operators, these systems must be able to execute based on incomplete information and be able to adapt to varying environments. The optimum solutions may not be definitive, and can be arrived at only through consideration of a variety of trade-offs. To make these complex and interconnected systems more robust, we need to build into them the ability to adapt to dynamically changing environments, and to deal with uncertainty. Recent efforts have gone by the names of autonomic computing [29], cognitive computing [48], and organic computing [22]. The idea is to incorporate properties of living organisms that can learn, grow, reconfigure, and recover from errors.

Some of the properties of organic computing systems are particularly important to the design of interfaces. Organic systems are robust to changes in the environment and operating

conditions. Some of the characteristics that we may want to emulate in organic interfaces include redundancy and degeneracy [46], often two sides of the same coin. Many biological systems are redundant (e.g., kidneys); they possess vastly more capacity than is necessary for the tasks they are designed to accomplish. They are also highly degenerate: there are usually many ways to satisfy a given requirement. For example, we can metabolize carbohydrates, fats, and proteins, even though the mechanisms for digestion and for extraction of energy from each of these sources is quite distinct.

An organic system can evolve over time by learning from experiences. For humans, learning can often be accomplished with just a few examples, rather than with voluminous amounts of data [7]. Through learning, the system can increase its knowledge base and expand its capabilities. For learning to take place, it must be self aware and context aware; it must be able to observe itself in varying operating conditions and modify its behavior based on this observation. It must also detect what it doesn't know, and find ways to incorporate this knowledge into the system for future use. Context-awareness, learning, and adaptation are three inter-related properties.

3. Research Challenges

Following this line of thinking, we can reexamine how today's speech-based interfaces can be made more organic. In this section, I will discuss some of the desirable properties of organic interfaces, and provide some illustrations of what has been done in these areas and some of the as-yet-unmet challenges. Since the space spanned by spoken dialogue systems is very large and space for this article is limited, I will primarily focus my attention on two aspects: speech understanding and dialogue management. As such, I will only address a few of the large number of challenges.[1] I am likely to draw heavily from our own experience in developing such systems at MIT over the past two decades. This is a consequence more of familiarity than of enthnocentricity.

3.1. Robustness

Human communication is inherently robust. We can understand spoken input in extremely adverse acoustic conditions, from talkers with varying amounts of accents, and do so with different kinds of input – text, speech, gesture, and facial expressions. Future systems' robustness in performance can conceivably be improved in many dimensions, some of which will be described below.

3.1.1. 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. High recognition accuracy has also been achieved over the telephone for systems with a working vocabulary of several thousand words [18]. 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. 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.

[1] See [4] for an excellent discussion of speech recognition and understanding challenges.

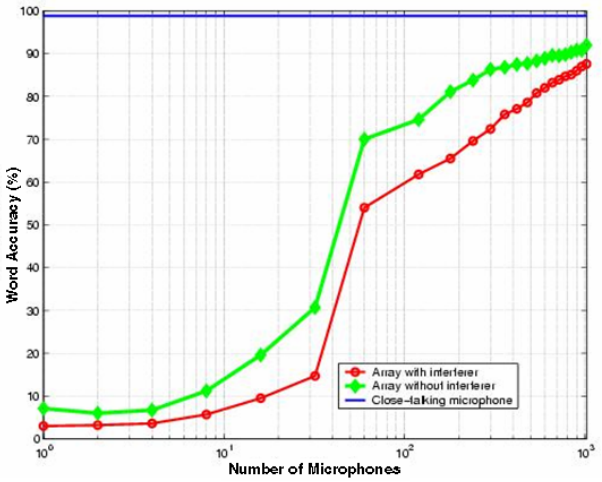

Figure 1: *Recognition accuracy as a function of the number of microphones in a microphone array. There is a modest degradation when multiple competing speakers are present.*

Researchers have demonstrated that the use of microphone arrays using beam-forming techniques to capture the desired signal can significantly improve performance [16]. As an example, Figure 1 shows the performance improvement as the number of microphones in the array increases from 1 to 1,000 [52].

While the use of microphone arrays is a promising direction for achieving robust data capture, one can not help noticing that, in this instance, 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 [32]. Some researchers have explored the use of auditory models as a recognition front-end [44, 19, 3]. These auditory front-ends typically yield similar performance to Fourier-based representations for clean speech (e.g., [34]). 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 or a Perceptual Linear Prediction representation that attempt to mimic some of the known properties of the human auditory system [24].

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 and understanding.

3.1.2. Lexical Access

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

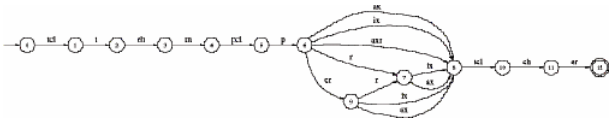

Figure 2: *A pronunciation graph for the word "temperature," after phonological expansion has been performed on the lexical baseform.*

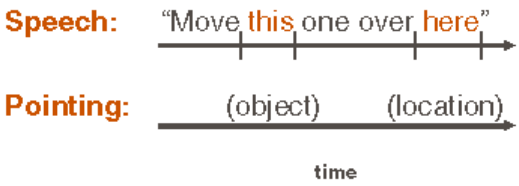

Figure 3: *A schematic plot, as a function of time, of the sequence of words determined by a speech recognizer, and the interpretation of the object and its target location determined from the visual signal.*

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 [21], 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 [45]. 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 [56]. Recently, this line of investigation has been revived and extended to continuous speech with promising results [47].

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 is 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 [24]. How such a search strategy can be formulated formally and implemented efficiently should be a topic of further research.

3.1.3. Multimodal Interactions

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 and add redundancy, 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 3, 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 also 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. 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 dialogues, the system should support user interruption, back-channel, and other cues that are common in human dialogues.

Our own research on a multimodal restaurant domain application has allowed us to explore a number of design issues related to both multimodal input (e.g., drawing a line along a street on a map while asking, "What restaurants are on this street?"), and multimedia output, interfacing with Google maps to provide richly informative visual feedback, which in turn reduces the importance of verbal summarization [20]. To provide an integrated search in multimodal understanding, Johnston and Bangalore [26] have developed a novel strategy for tight coupling between speech and mouse clicks via a joint interpretation within a weighted finite state transducer (FST) framework. Another example of a rather unique multimodal dialogue system is WITAS [30] which allows the user to interact with a simulated robotic helicopter, via speech and mouse clicks on a map. The user can instruct the virtual helicopter to fly to different locations, follow vehicles, and deliver goods.

For the multi-modal interaction to be effective, we must develop a unifying linguistic formalism that can describe multimodal 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 deci-

sion 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).

3.2. Establishing Context

Context setting is an important aspect of spoken language communication. Knowing that we are speaking in a noisy environment, for example, enables us to adapt and disregarding the interference, whether it be traffic noise, music, or competing talkers. Knowledge about linguistic constructs enables us to favor one set of words over another (e.g., "euthanasia" *vs.* "youth in Asia"). Discourse knowledge is crucial for us to interpret the meaning of sentences based on previous parts of the conversation. For example, there are many ways to interpret the user query, "What about Japanese?," including it's culture, geography, food, weather, etc. But the question would be unambiguous if it is known that the previous sentence is "Where is the nearest Chinese restaurant?"

Much work has been done by the research community on context setting. In speech recognition, for example, the use of context-dependent phone models, *n*-gram language models, and trigger-based language models are all attempts at establishing the context [24]. In recent research transcribing Broadcast News, some researchers pre-segmented the input signal to mark changes in environment and talker in order to improve speech recognition performance [8]. A discourse component is typically included in a spoken dialogue systems to help establish the context [58]. However, there is much more to be done; I will briefly describe a couple of ideas in this section.

A common practice in spoken language interface design is the assumption that speech is whatever a microphone picks up. This is clearly not the case in our everyday life, where our ears are bombarded with a large variety of sounds, some bearing linguistic information and others not. One simple solution would be to use a "push to talk" mechanism to ensure that the system only *listens* to the signals that it should process and interpret. However, this can be cumbersome for the user, and difficult to implement for applications such as processing meeting recordings. Besides, the non-speech intervals may contain useful information that would be helpful for speech decoding.

A promising approach would be to process the entire audio signal, segmenting it into acoustically homogeneous chucks, and classifying each segment into different categories – speech, music, speech with background music, two people speaking simultaneously, etc. This kind of auditory scene analysis [14] can potentially provide a rich description of the acoustic signal such that the appropriate processing steps can then be taken.

3.3. Adaptation

Much has been done in the area of adapting an interface to the environment in which it operates. For example, both on-line and off-line adaptation techniques have been applied to good effect to minimize the mismatch between the training data and the test data, whether it be the result of differences in the environment, speaker, or language model [24].

The development of conversational systems shares many of the research challenges being addressed by the speech recognition community for other applications such as speech dictation and spoken document retrieval, although the recognizer is often exercised in different ways. For example, in contrast to desktop dictation systems, the speech recognition component in a conversational system is often required to handle a wide range of channel variations. Increasingly, landline and cellular phones are the transducer of choice, thus requiring the system to deal with narrow channel bandwidths, low signal-to-noise ratios, diversity in handset characteristics, drop-out, and other artifacts.

Another problem that is particularly acute for conversational systems is the recognition of speech from a diverse speaker population. In the data we collected for JUPITER – a telephone-based spoken dialogue system [57], for example, we observed a significant number of children, as well as users with strong dialects and non-native accents. The challenge posed by these data to speaker-independent recognition technology must be met [33], since conversational interfaces are intended to serve people from all walks of life.

A solution to these channel and speaker variability problems may be adaptation. For applications in which the entire interaction consists of only a few queries, short-term adaptation using only a small amount of data would be necessary. For applications where the user identity is known, the system can make use of user profiles to adapt not only acoustic-phonetic characteristics, but also pronunciation, vocabulary, language, and possibly domain preferences (e.g., user lives in Boston, prefers aisle seat when flying).

However, there is much more that could be done. For example, the interface should be able to adapt to the choice of words and grammar of a user, in order to improve its understanding capability. To make the interaction more productive and enjoyable, it could also learn the likes and dislikes of the user and make suggestions when appropriate.

3.4. Learning

Over the past several decades, we have steadily seen stochastically-motivated learning techniques being applied to human language technologies. Hidden Markov modeling, for example, illustrates how powerful stochastic techniques can be used to perform speech recognition [25], language understanding [40, 36, 55], and machine translation [6]. By formulating a statistical model of sounds and words and using training data to estimate model parameters, the community has been able to dramatically increase both the performance and the complexity of the systems.

The taxonomy of learning as it applies to speech-based interfaces can be quite complex. At one extreme, the system could learn by observing user behaviors over time, and then making use of a statistical model of observed patterns to bias future decisions. The user might not even be aware that the system is altering its model of the world. Alternatively, the system may need to actively engage the user in dialogue in order to learn their preferences explicitly or to acquire new knowledge about the world. Finally, the system may want to learn user behavior explicitly through imitation. Some of the learning can be done off line, whereas other forms must be done during actual usage.

Note that passive learning is closely related to the issue of adaptation. Take speaker adaptation, for example, where the system typically updates the acoustic model parameters incrementally during usage to tune them to the speaker's voice characteristics. Any system which involves enrollment can in theory also benefit from repeat usage, learning not only low-level voice characteristics, but also higher level features such as their language usage patterns or even preferences such as a favorite cuisine or a bias towards cheap restaurants [51]. Such knowledge can be used then to alter information presented in summative responses when a large set of database items match a specified constraint. An interesting example of tailoring response gener-

ation content to a user model in the flight domain is described in [1].

3.4.1. Statistical Dialogue Management

Statistical methods, and more generally machine learning techniques, have been unusually slow to penetrate dialogue management in spoken dialogue systems. The main reason is that such techniques depend critically on large corpora of manually annotated data. For dialogue systems, this translates into the need for detailed, annotated log files for thousands of user dialogues with a pre-existing system. Another roadblock has been uncertainty in how to formulate a tractable machine learning paradigm for the highly heuristic task of dialogue management. A powerful method for side-stepping the data collection issue is to collect synthetic data from user simulation runs [9, 43], although one ultimately has to confirm that the results carry over to real users. The beauty of simulation is that the user's intended actions are known, so that no manual annotation is required. And the developer can simulate compliant or non-compliant behavior, known or unknown vocabulary choices, etc., in controlled experiments, generating enormous amounts of training data effortlessly.

Levin et al. [31] were pioneers in introducing machine learning techniques combined with user simulation, in experiments where they showed that a reasonable policy could be learned in the flight domain through trial and error. Bayesian reinforcement learning, Markov Decision Processes (MDP's), and Bayesian Belief Networks (DBN's) [35] have evolved into the more complex "POMDP" (Partially Observable MDP) model [53], which is beginning to catch on as a method to replace heuristic rules governing dialogue management decisions. However, it is as yet unclear whether these techniques can scale to complex domains.

Some researchers have utilized machine learning techniques to sovle a restricted part of the dialogue management problem, rather than attempting to completely replace an existing dialogue manager. For example, user simulation data can be used to train RIPPER rules [12], to produce a rule-based system for deciding whether to invoke implicit or explicit confirmation, or to seek a spoken spelling of a potentially unknown word [15]. Bohus and Rudnicky [5] have developed a data-driven approach which integrates information across multiple turns to aid in the decision process for implicit vs explicit confirmation. Wu and Seneff [54] showed improved speech understanding in a spoken dialogue system by using a genetic algorithm to bias N-best selection towards utterances whose speech act is primed due to dialogue context. Straightforward machine learning techniques have also been successfully applied to the task of deciding whether to reject a user's utterance due to suspicions of gross recognition error [17]. Johnston and Bangalore [27] borrowed techniques from the statistical machine translation community to train a set of "edit rules," leading to more robust handling of multimodal inputs. Through an automated intent-mapping algorithm, Tur [49] has shown how annotated data from one domain can be used to build an initial speech understanding model for a new but related domain.

3.4.2. Interactive Learning

The type of learning discussed above is largely achieved off-line in that the data are typically used in a training phase whereby the parameters of the systems are adjusted prior to actual use. By relying on a large corpus, the system development must be preceded by a data collection phase, and as such the resulting systems tend to be highly task dependent. This raises the question about portability, i.e., can we generalize from one task to another without having to be continually on the tread-mill of data collection?

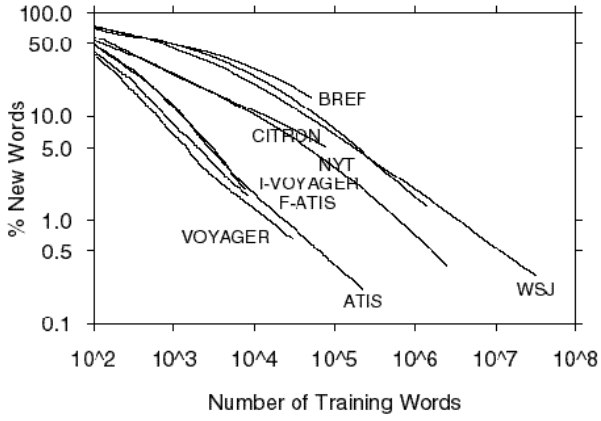

Figure 4: *The percentage of unknown words in previously unseen data as a function of the size of several training corpora used to determine the vocabulary empirically [23].*

In many other cases, on-line, interactive learning is critical. Consider the problem of unknown words, for example. 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 [23]. This point is illustrated in Figure 4. For the Air Travel Information System, or ATIS, task, for example, the probability of the system encountering an unknown word, is about 0.002, even after encountering a 100,000-word training corpus. In real applications, 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 nearly enough attention to the unknown word problem. If the system's working vocabulary is open, then 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. If unknown words are always going to be present, then a spoken dialogue system must be able to cope with unknown words, because 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 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.

What is needed, then, is a generic capability to handle unknown 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 ex-

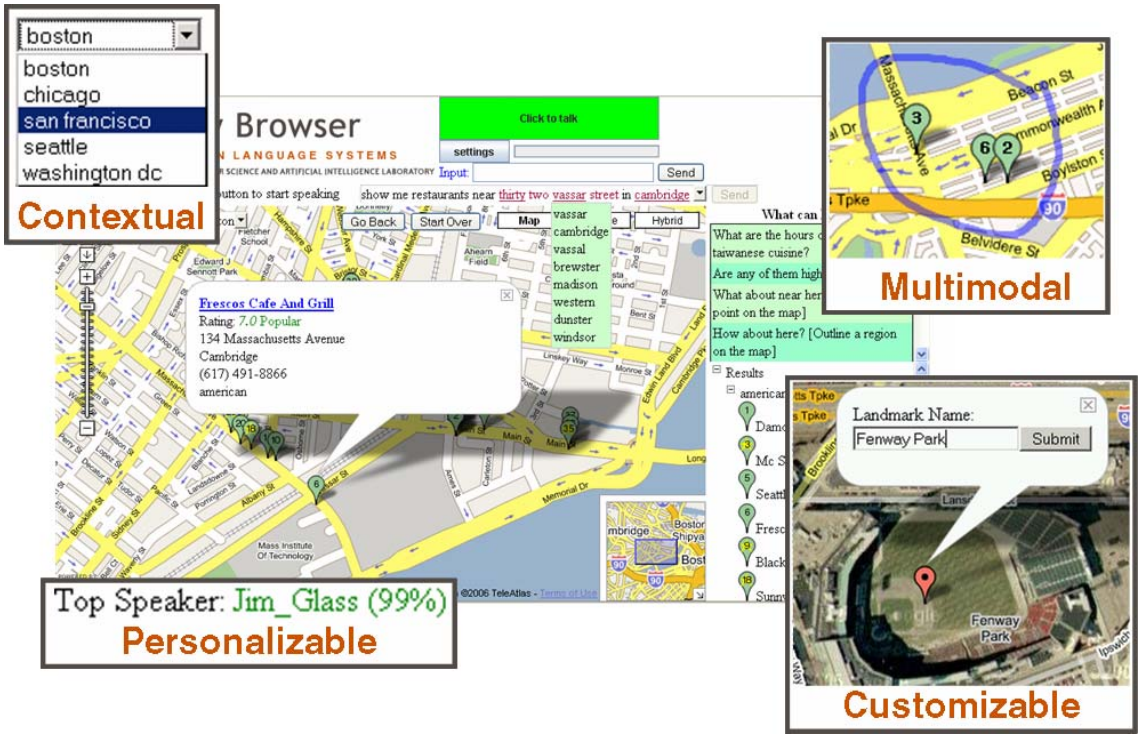

Figure 5: *An illustration of the MIT City Browser, which can provide information about restaurants, landmarks, and transportation for several major cities in the U.S.*

plicitly and understands its usage. For example, Chung et al. developed a mixed-initiative spoken dialogue system that can flexibly incorporate new words from users and from dynamic information sources retrieved from the Web [11]. Specifically, the system enlists a software agent to seek the data entries from the Web, given the current dialogue context. Subsequently, the system updates its vocabulary and language models with the newly retrieved data subset. A desirable feature of the system is that changes in the database content via updates, such as new restaurants, do not require re-compilation of the main finite-state transducers (FSTs) in the recognition or the natural language parser. As an illustration, a user may ask, "What is the number of *Flora* in Arlington?". The system initially understood the query as "What is the number of ¡unknown word¿ in Arlington." Based on the context, the system will retrieve all the restaurants in Arlington from a restaurant database, and select *Flora* based on its acoustic similarity to the input. It will then respond to the user with the requested information, and updates the speech recognition and language understanding components so that the previously unknown word can be used subsequently. This can all be done on one sentence.

Thus far, the new word acquisition problem has only been attempted in cases where the new word is assumed to be a member of a previously defined class, such as a user or an establishment name [10, 20]. In the future, we can anticipate systems whose knowledge base can slowly grow through direct interaction with end users. The system would logically guide the user through a subdialogue asking for information to fill each slot in a new database entry. Such knowledge could also be acquired multimodally, by allowing the user to speak and type the new word, or by taking advantage of existing handwriting recognition systems for pen-based script input of the new word. Advancing to the level of an entirely new class of objects is far more challenging. Pioneering work by Roy [41] has begun to address the problem of acquisition of knowledge through multimodal *input* – by having a robot learn to associate spoken words with attributes associated with objects pressented to its visual field.

3.4.3. Learning by Imitation

In *learning by imitation*, the user can simply *show* the systems how to performance certain tasks, and in most cases, provide a spoken commentary to be associated with that task for future use. For example, a complex sequence of clicks on a menu hierarchy in a smart phone can be directly linked to the verbal command "enable Bluetooth."

There has been, to my knowledge, very little prior research within the speech community in the area of learning by imitation. Most notable however is the recent research by James Allen and his team [2] involving walking through a sequence of steps at a Web page in order to teach the system how to search and summarize from complex linguistically understood queries. Task models are constructed by fusing together information from language understanding with the observed demonstration. The active learning process allows the system to be able to master the process from a single example, due to the linguistic scaffolding provided in accompanying spoken dialogue interaction. This area seems ripe for new ideas emerging out of the space of mobile devices, now that it is feasible to run speech recognition systems on these devices. This is especially appealing because such devices are becoming so capable that menu-driven approaches to accessing services are beginning to become too unwieldy to be appealing.

4. Concluding Remarks

In this paper, I try to argue that future interfaces should behave more like a living organism that can provide robust performance in a wide range of operating conditions, learn from their experiences and adapt to the environment, user, and task. Some of the challenges of developing such an interface are being pursued by the research community with good results, while others will need increased attention. It is my hope that some of the ideas may find their way onto the research agenda of others in the community, in addition to my own.

Figure 5 shows an example of a spoken dialogue interface under development at MIT that possesses some of the properties that I have discussed in this paper. The interface is integrated to a number of online information sources. The system is context-aware in that it will customize its domain knowledge based on a user's specification of a particular city of interest. Pen gesture is integrated with speech to respond to inquiries such as, "Are there any Italian restaurants in this area?" A user can customize the system by adding personal landmarks. The system can also monitor the characteristics of the speaker, and can bring in speaker-specific models to improve performance. This is admittedly only a small step towards the realization of an organic interface. Nonetheless, it provides a platform for us to explore some of the challenges described in this paper.

5. Acknowledgements

The ideas described in this paper have been influenced by my collaboration with many past and current students and staff of the Spoken Language Systems Group at the MIT Computer Science and Artificial Intelligence Laboratory, especially Stephanie Seneff and Jim Glass. Their contributions are gratefully acknowledged. This research is sponsored by the T-Party Project, a joint research program between MIT and Quanta Computer Inc.

6. References

[1] M. Absdil and O. Lemon,"Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems," *Proc. ACL,* 2004.

[2] Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Swift, M., and Taysom, W., "PLOW: A Collaborative Task Learning Agent," *Proc. Twenty-Second Conference on Artificial Intelligence*, 2007.

[3] J. Allen, "How do humans process and recognize speech", in Ramachandran and Mammone, (Eds.), *Modern Methods of Speech Proceessing*, 251–275, Kluwer, 1995.

[4] J. Baker, L. Deng, J. Glass, S, Khudanpur, C. Lee, and N. Morgan, "Historical Development and Future Directions in SPeech Recognition and Understanding," *MINDS Workshop: Report of the Speech Understanding Working Group,*, 2007.

[5] D. Bohus and A. Rudnicky, "Constructing Accurate Beliefs in Spoken Dialog Systems," *Proc. ASRU*, 272–277, 2005.

[6] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Laffert , R. Mercer, P. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, 16(2), 79–85, 1990.

[7] S. Carey and E. Bartlett, "Acquiring a single new word," *Papers and Reports on Child Language Development,* 15, 17–29, 1978.

[8] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion," *Proc. Broadcast News Transscription and Understanding Workshop,* 127–132, 1998.

[9] G. Chung, "Developing a Flexible Spoken Dialog System Using Simulation," *Proc. ACL,* 63–70, 2004.

[10] G. Chung, S. Seneff, and C. Wang, "Automatic Acquisition of Names Using Speak and Spell Mode in Spoken Dialogue Systems," *Proc. HLT-NAACL*, 197–200, 2003.

[11] G. Chung, S. Seneff, C. Wang, and L. Hetherington, "A Dynamic Vocabulary Spoken Dialogue interface,"*Proc. ICSLP*, 2004.

[12] W. Cohen, "Fast Effective Rule Induction," *Proc. 12th Intl. Conference on Machine Learning*, 115–123, 1995.

[13] M. Denecke, and A. Waibel, "Dialogue Strategies Guiding Users to Their Communicative Goals," *Proc. Eurospeech,* 2227–230, 1997.

[14] D. Ellis, "Model-Based Scene Analysis," Chapter 4 in D. Wang and G. Brown, (Eds.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications,* Wiley/IEEE Press, 115–146, 2006.

[15] E. Filisko and S. Seneff, "Learning Decision Models in Spoken Dialogue Systems via User Simulation," *Proc. AAAI Workshop: Statistical and Empirical Approaches for Spoken Dialog Systems*, 2006.

[16] J. Flanagan and E. Jan, "Sound capture form spatial volumes: matched-filter processing of microphone arrays having randomly-distributed sensors", *Proc, ICASSP*, 1996.

[17] M. Gabsdil and O. Lemon, "Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems," *Proc. ACL,* 2004.

[18] J. Glass,T. Hazen, and I. Hetherington, "Real-time telephone-based speech recognition in the JUPITER domain," *Proc. ICASSP*, 1999.

[19] O. Ghitza, O. "Auditory models and human performance in tasks related to speech coding and speech recognition," in Ramachandran and Mammone, eds., *Modern Methods of Speech Proceessing*, Kluwer, 1995.

[20] A. Gruenstein, S. Seneff, and C. Wang, "Scalable and Portable Web-based Multimodal Dialogue Interaction with Geographical Databases," *Proc. INTERSPEECH*, Pittsburgh, PA, 2006.

[21] T. Hazen, , I. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Proc. PMLA Workshop,* 99–104, 2002.

[22] H. Herzel, C. von der Malsburg, W. von Seelen, and R. Wurtz, (Eds.) *Organic Computing: Towards Structured Design of Processes*, Interdisciplinary Symposium, 2001.

[23] L. Hetherington and V. Zue, "New words: Implications for Continuous Speech Recognition," *Proc. EUROSPEECH*, 475–931, 1991.

[24] X. D. Huang, A. Acero, and H. Hon, *Spoken Language Processing*, Prentice Hall, New York, 2001.

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

[26] M. Johnston and S. Bangalore, "Finite-state Multimodal Parsing and Understanding," *Proc. 18th Intrenational Conference on Computational Linguistics*, 369–375, 2000.

[27] M. Johnston and S. Bangalore, "Learning Edit Machines for Robust Multimodal Understanding," *Proc. ICASSP '06*, pp. I617–I620, 2006.

[28] L. P. Kaelbling, "A New Computer Science," Unpublished Manuscript.

[29] J. Kephart and D. Chess, "The Vision of Autonomic Computing," *J. IEEE Computer Society*, 41–50, 2003.

[30] O. Lemon, A. Gruenstein, and S. Peters, "Collaborative Activities and Multitasking in Dialogue Systems, *Traitement Automatique des Langues*, 43(2), 131–154, 2002.

[31] E. Levin, R. Pieraccini, and W. Eckert, "Using Markov Decision Process for Learning Dialogue Strategies," *Proc. ICASSP*, 201–204, 1998.

[32] R. J. Lippmann, "Speech Recognition by Humans and Machines," *Speech Communication,* 22(1), 1–15, 1997.

[33] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," *Proc. ICASSP,* 2000.

[34] H. Meng and V. Zue, "A Comparative Study of Acoustic Representations of Speech for Vowel Classification using Multi-Layer Perceptrons", *Proc. ICSLP,*, 1990.

[35] H. Meng, C. Wai, and R. Pieraccini, "The Use of Belief Networks for Mixed-Initiative Dialog Modeling," *IEEE Transactions on Speech and Audio Processing*, 11(6), 757–773, 2003.

[36] S. Miller, R. Schwartz, R. Bobrow, and R. Ingria, "Statistical Language Processing Using Hidden Understanding Models," *Proc. ARPA Speech and Natural Language Workshop*, 278–282, 1994.

[37] National Institute of Standards and Technology (NIST), "RT-03F evaluation," http://www.nist.gov/speech/tests/rt/rt2003/fall/rt03f-evaldiscd oc/index.htm, 2003.

[38] S. Oviatt, "Multimodal Interfaces for Dynamic Interactive Maps," *Proc. Conference on Human Factors in Computing Systems: CHI '96*, 95–102, 1996.

[39] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, N.J., 1993.

[40] R. Pieraccini, E. Levin, and W. Eckert, "AMICA: The AT&T mixed initiative conversational architecture," *Proc. Eurospeech,* 1875–1879, 1997.

[41] D. Roy and A. Pentland, "Learning Words from Sights and Sounds: A computational Model," *Cognitive Science*, 26(1), 113–146, 2002.

[42] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR ν-Talk Speech Synthesis System," *Proc. ICSLP*, 483–486, 1992.

[43] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. (2006), "A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies." *Knowledge Engineering Review,* 21(2): 97–126, 2006.

[44] S. Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics,* (16), 55–76, 1988.

[45] S. Seneff and C. Wang, "Statistical Modeling of Phonological Rules through Linguistic Hierarchies," *Speech Communication*, 2004.

[46] G. Sussman, "Building Robust Systems," Unpublished manuscript.

[47] M. Tang, M., S. Seneff, and V. Zue, "Modeling Linguistic Features in Speech Recognition," *Proc. Eurospeech,* 2585–2588, 2003.

[48] T. Tether, *Statement to the U.S. Senate Committee on Armed Services*, April 2002.

[49] G. Tur, "Multitask Learnign for Spoken Language Understanding," *Proc. ICASSP*, I585–I588, 2006.

[50] M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Owen Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, D. Stallard, "DARPA Communicator: Cross-system Results for the 2001 Evaluation," *Proc. ICSLP*, 273–276, 2002.

[51] M. Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy (2004), "Generation and Evaluation of User Tailored Responses in Dialogue," *Cognitive Science*, 28, 811-840, 2004.

[52] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "LOUD: A 1020-Node Microphone Array and Acoustic Beamformer," to appear at the International Congress on Sound and Vibration (ICSV), 2007.

[53] J. Williams and S. Young, "Scaling POMDPs for Dialog Management with Composite Summary Point-Based Value Iteration (CSPBVI)," *Proc. AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*, pp. 7–12, 2006.

[54] H. C. Wu and S. Seneff, "Reducing Recognition Error Rate based on Context Relationships among Dialogue Turns," *Proc. Interspeech*, 2007.

[55] S. Young, J. Schatzmann, K. Weilhammer and H. Ye. (2007), "The Hidden Information State Approach to Dialog Management." *Proc. ICASSP*, 2007.

[56] V. Zue, "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, 1983.

[57] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A Telephone-Based Conversational Interface for Weather Information," *IEEE Trans. SAP*, 8(1), 85–96, 2000.

[58] V. Zue and J. Glass, "Conversational Interfaces: Advances and Challenges," *Proc. IEEE, Special Issue on Spoken Language Processing*, 88, 2000.