Universal Speech Interfaces

Roni Rosenfeld, Carnegie Mellon University

Dan Olsen, Brigham Young University

Alex Rudnicky, Carnegie Mellon University

Introduction

In recent years speech recognition has reached the point of commercial viability realizable on any off-the-shelf computer. This is a goal that has long been sought by both the research community and by prospective users. Anyone who has used these technologies understands that the recognition has many flaws and there is much still to be done. The recognition algorithms are not the whole story. There is still the question of how speech can and should actually be used. Related to this is the issue of tools for development of speech-based applications. Achieving reliable, accurate speech recognition is similar to building an inexpensive mouse and keyboard. The underlying input technology is available but the question of how to build the application interface still remains. We have been considering these problems for some time [Rosenfeld et. al., 2000a]. In this paper we present some of our thoughts about the future of speech-based interaction. This paper is not a report of results we have obtained, but rather a vision of a future to be explored.

From the inception of speech recognition research, the goal of “natural” interaction with computers has been cited as the primary benefit. The concept of talking with a machine as fluently and comfortably as with another human being has attracted funding and interest. We wanted to create HAL from “2001: A Space Odyssey” with a gentler personality. Naturalness of communication, however, is not the only goal of speech recognition, nor is naturalness reserved for speech. Based on the idea that “A picture is worth a thousand words,” graphical user interfaces have been developed purposely for their naturalness. If naturalness is not the key driver for speech then what is? We think there are at least three fundamental advantages for speech.

1. Speech is an ambient medium rather than an attentional one. Visual activity requires our focused attention while speech allows us to interact while using our other faculties (e.g. visual, sensory-motor) to do something else.
2. Speech is descriptive rather than referential. When we speak we describe objects in terms of their roles and attributes. In visual situations we point to or grasp the objects of interest. For this reason, speech and pointing are to a large extent complementary, and can often be combined to great effect.
3. Speech requires more modest physical resources. Speech-based interaction can be scaled down to much smaller and much cheaper form-factors than visual or manual modalities.

We see a future for speech, not so much for its naturalness but for its ubiquity. Natural language interfaces are still an interesting and important topic, but speech interaction is both more and less than that. For this reason we choose to focus on speech as an input/output modality rather than as a medium for natural language.

We see interactive systems as mechanisms for humans to express needs to and obtain services from machines. By machines we mean not only computers in the lay sense of the word, but also any gadget, appliance or automated service which, in order to be fully utilized, must be reconfigured, controlled, queried or otherwise communicated with. We are surrounded by dozens of such machines today. The exponential drop in the fundamental costs of computing will cause hundreds more to be developed in the near future. Examples of such interactivity include:

- Configuring and using home appliances (VCRs, microwave and convection ovens, radios, alarms…)
- Configuring and using office machines (fax machines, copiers, telephones…)
- Retrieving public information (e.g. weather, news, flight schedule, stock quotes…).
- Retrieving and manipulating private information (e.g. bank or other accounts, personal scheduler, contact manager, other private databases).
- Handling asynchronous communication (voice, email, fax).
- Controlling miscellaneous user and consumer applications (map following, form filling, web navigation).

To clarify our ideas, we distinguish between intelligent and simple machines. Suppose that one possessed a household robot of the kind envisioned by Isaac Asimov. While seated at the table one might hand a dirty dish to the robot and say “Take care of this.” This is a natural statement that most children would understand. The desire is to have the dish taken to the kitchen, rinsed and placed in the dishwasher. The naturalness of this communication comes from the large amount of world knowledge that the robot must possess to infer the correct response. This is not so much an interaction problem it is a natural language and inferential reasoning problem. This problem is not necessarily related to speech. A more natural form of communication would be to hand the dirty dish to the robot and then expect the right thing to happen without any speech at all. We would consider this an intelligent machine.

Our focus for speech interfaces is on simple machines. In a simple machine the user can, at least in principle, possess a mental model of the machine’s capabilities and of the machine’s rough state. Further, the user is assumed to know ahead of time what is desired, although they do not have to know *how* to get it done. Under this paradigm, high-level intelligent problem solving is done by the human; the machine is only a tool for getting needed information, modifying it, and/or issuing instructions to the desired service. We find speech interaction with simple machines to be interesting because it completely sidesteps the artificial intelligence problems of natural language in favor of the ubiquitous nature of speech interaction. We believe that in the future human beings will be surrounded by hundreds if not thousands of simple machines with which they will want to interact.

In particular, the approach we are proposing is not aimed at applications requiring truly intelligent communication. For example, an air travel reservation system will fall within our focus only if the machine is used to consult flight schedules and fares and to book flights, while actual planning and decision making is done by the user. The machine plays the role of a passive travel agent, who does not do much thinking on their own, but mostly carries out the explicit requests of the user. The more intelligent travel agent, however desirable, is outside the scope of this discussion.

Why Speech?

In dealing with simple machines one might ask why one should go to the effort of recognizing speech, when a visual technique would suffice. When considering technologies that might meet the needs of ubiquitous interactivity, the issues of situation, cost, breadth of application and the physical capabilities of human beings must be considered. When we talk of embedding interactivity into a wide variety of devices and situations, the keyboard and mouse are not acceptable interactive devices. In any situation where the user is not seated at a flat surface, a keyboard or mouse will not work. In any application with any richness of information, a bank of buttons is unacceptably restrictive. In a large number of cases only speech provides information rich interaction while meeting the form factor needs of the situation.

If we consider future exponential growth, it is clear that any interactive solution relying primarily on processing and memory capacity will over time become very small and very very cheap. Speech interaction requires only audio I/O devices (i.e., microphone and speaker), which are already quite small and cheap, coupled with significant processing power which is expected to become cheap. No such projections hold for keyboards, buttons and screens. Visual displays have made only modest gains in pixels per dollar over the last 20 years and no order of magnitude breakthroughs are expected. Visual displays are also hampered by the size requirements for discernable images and by the power required to generate sufficient light energy. Buttons are cheap but are also restricted as to size and range of expression. Any richness of interaction through the fingers quickly becomes too large and too expensive for ubiquitous use. Only speech will scale along with the progress in digital technology. A Palm Pilot can get more powerful, but it cannot get physically smaller. Its size and form factor are dominated by its screen, touch pad and the limitations of human beings [Olsen 99]. Speech interfaces have much smaller minimal sizes and power requirements.

Spoken language dominates the set of human faculties for information-rich expression. Normal human beings, without a great deal of training, can express themselves in a wide variety of domains. As a technology for expression, speech works for a much wider range of people than typing, drawing or gesture because it is a natural part of human existence. This breadth of application is very important to ubiquitous interactivity.

State of the art in developing speech interfaces

Speech interfaces are only beginning to make an impact on computer use and information access. The impact thus far has been limited to those places where current technology, even with its limitations, provides an advantage over existing interfaces. One such example is telephone-based information-access systems, because they provide a high input-bandwidth in an environment where the alternative input mode, DTMF (Touch-Tone telephone buttons), is inadequate. We believe that speech would achieve much higher penetration as an interface technology if certain fundamental limitations were addressed. In particular:

- Recognition performance
- Accessible language (for the users)
- Ease of development (for the implementers)

That is, it should be possible for users to verbally address any novel application or artifact they encounter and expect to shortly be engaged in a constructive interaction. At the same time it should be possible to dramatically reduce the cost of implementing a speech interface to a new artifact such that the choice to add speech is not constrained by the cost of development. There is a strong interplay between the interface needs of an application and the means to economically develop that interface.

We consider three current approaches to the creation of usable speech systems: natural language, dialog trees, and commands. First, we can address the problem of accessible language by allowing the user to use unconstrained natural language in interacting with an application. That is, given that the user understands the capabilities of the application and understands the properties of the domain in which it operates, he or she is able to address the system in spontaneously formulated utterances. While this removes the onus on the user to learn the language supported by the application, it places it instead on the developer. The developer needs both to collect a large corpus of user expressions and to create an interpreting component that maps user inputs into application actions. Examples in the literature of this approach include ATIS [Price 90] and Jupiter [Zue 97].

This approach not only places a huge usability burden on the developer to understand any reasonable utterance by the user, but also carries the additional burden of supporting a discovery dialog. All computing systems and human beings have limitations on their knowledge and the services that they provide. For a user to effectively use a system they must have a means for discovering those abilities and limitations. A structured language implicitly communicates the functional limits of the application: what is supported is exactly what can be expressed. An unconstrained language, on the other hand, must do so explicitly: in the natural language approach the developer must not only develop the application language but also a sufficiently broad meta-language to support user discovery.

The natural language approach attempts to exploit transfer of prior human experience with other people to simplify the learning of a new application. This handcrafted language development strategy, however, does not produce any transfer of learning among human-machine interfaces. Even if there is learning in the context of a particular application (say through the modeling of spoken utterance structure [Zoltan-Ford 91]) there is no expectation that this learning will transfer across applications, since development efforts are not systematically related.

A second problem with natural language interfaces is that they do not systematically address the issue of constraining language with a view to improving recognition performance; the only improvement in recognition performance is through the accumulation of a domain-specific corpus. For the foreseeable future constraining possible alternatives is the most important method for producing high recognition accuracy. Even human beings in high stress or high noise situations use restricted formalized vocabularies for accurate communication. The freely spoken natural language model works against this need.

One of the primary contentions in favor of unconstrained natural language interaction is that there is zero learning time because all human experience transfers to an application that understands natural language. In practice there is no natural language experience that has zero learning time. All human-human situations require time to acquire terminology, shared history, shared goals and relationships of trust. Human beings regularly find themselves in situations where they do not understand and cannot communicate effectively even though all parties are speaking the same language. More importantly, unconstrained language does not inherently eliminate communication difficulties. Because the knowledge and capacities of human beings are so rich, we require rich language structures to build up the shared understandings that support natural language. When communicating with “dumb machines”, such rich mutual understandings with their shades and nuances of meaning are not only not required but in many cases are an impediment to getting service. Simplifying the language and unifying its structure will reduce these problems. Simple machines as defined earlier are fundamentally limited in the services that they can perform. We want a system that will readily and efficiently communicate those limits, so that users can make their own judgments about effective usage. We expect predictable, efficient subservience from our simple machines, not independent behavior and thought.

As an alternative to allowing the user to speak freely (and compensating for this in the parsing and understanding component of the system) we can constrain what the user can say and exploit this constraint to enhance system performance. We consider two such approaches, *dialog-tree* systems and *command and control* systems.

Dialog-tree systems reduce the complexity of recognition by breaking down activity in a domain into a sequence of choice points at which a user either selects from a set of alternatives or speaks a response to a specific prompt (such as for a name or a quantity). The drawbacks of such systems, from the user’s perspective, center on the inability to directly access those parts of the domain that are of immediate interest, and to otherwise short-circuit the sequence of interactions designed by the developer. A space with many alternatives necessarily requires the traversal of a many-layered dialog tree, as the number of choices at any one node will necessarily be restricted. From the designer's perspective such systems are difficult to build, as they require being able to break an activity down into the form of a dialog graph; maintenance is difficult as it may require re-balancing the entire tree as new functionality is incorporated. While dialog-tree systems may be frustrating to use and difficult to maintain, they do simplify the interaction as well as minimize the need for user training. What this means is that the user's contribution to the dialog is effectively channeled by the combination of directed prompts and the restricted range of responses that can be given at any one point. This is the approach commonly used in commercial development of public speech-based services.

Command and control interfaces reduce complexity by defining a rigid syntax that constrains possible inputs. The language consists of a set of fixed syntactic frames with variable substitution (for example, "TURN VALVE <valve-id> TO POSITION <position-value>"). In a restricted domain, such a language provides the user with sufficient power to express all needed inputs. At the same time the recognition problem is simplified since the language is predictable and can be designed to avoid confusion. The utility of such interfaces depends on the willingness of users to spend the time to learn the language for such a system. The drawback of command and control systems stems from the investment that the user makes in learning the language. Being able to communicate with additional applications requires that this effort be duplicated. This may be feasible for a few applications but it is unlikely to scale to tens or hundreds of applications, which is what would be required of an interface that permits ubiquitous access to machines. It also suffers in comparison with dialog-graph based systems in that this interaction style is not inherently self-explanatory; the user is not guided into the correct input style by the structure of the interface.

Table 1 summarizes the properties of the three approaches to speech interfaces discussed above. As can be seen, no approach is satisfactory along all dimensions.

Speech interface approach:	Unconstrained Natural Language	Dialog trees	Command and Control
User's effort	low	moderate	high (moderate with use)
Developer's effort	very high	moderate	moderate
User training	none	none	required for each application
Supports discovery?	no	yes, inefficiently	possibly
Scales to complex tasks?	unknown	no	no
Scales to hundreds of applications?	yes, but effort not amortized	yes	no
Stress on speech recognizer & parser	high	low	moderate

Lessons from the GUI revolution

In seeking a solution for effective development of spoken language interfaces we can learn from the history of graphical user interfaces. The current state of research into speech interaction and its supporting tools is very similar to the state of research into graphical user interfaces in the early 1980s. At that time it was assumed that tools for constructing graphical user interfaces must remain very general in their capabilities. A primary goal was to provide maximum flexibility for designers to create and implement any interactive dialog. It was clearly stated that tools should not introduce bias as to the types of interactive dialogs that should be developed [Newman 68, Jacob 82].

The commercial advent of the Apple Macintosh changed all of that thinking. The Macintosh clearly showed that wide variability in interactive dialog was harmful. A fundamental maxim of the Macintosh was its style guide and the toolkit that supported it. The shift was away from all possible interactive styles to uniformity of interactive style. When menus, scrollbars, buttons, dragging and double-clicking all work in the same way, global ease of use is much greater than when these items are uniquely crafted for each situation. Some major keys to the Macintosh success are: 1) once a user learns the basic alphabet of interactive behaviors, those behaviors can be transferred to almost any Macintosh application, and 2) when faced with a new Macintosh application the set of things to try is very clear. This transference of experience and the presentation of a clear space of exploration as a way to master new applications has been wildly successful in graphical user interfaces, and has come to dominate virtually all interactive computing.

The uniformity of the Macintosh style also has impact on the way in which interactive applications are developed. Instead of creating tools that supported general dialogs of any sort, a widget-based strategy dominated. In the widget strategy the paradigm is prebuilt, pretested pieces that are assembled to address the needs of a given application. These pieces are parameterized so designers can fill in the application-specific information. The widget strategy for development has been extended to the notion of interactive frameworks. In such frameworks, developers are provided with a complete skeleton application with all standard behaviors predefined. The particular needs of the application are then fitted into this framework. This tool and implementation strategy not only simplified development but also reinforced the uniformity of applications. The cost of GUI development and the range of programmers with such skills both improved dramatically

The key insight of the Xerox Star/Macintosh revolution is that usability is a global rather than a local phenomenon. Users do not use one or two applications in the course of their work. They use many different applications and are repeatedly faced with learning new ones. Furthermore those applications must continually work together to create a productive whole. Uniformity in the basic interactive alphabet increases the transference of learning from one application to another and enhances the overall usability of the entire interactive environment. Such transference becomes enormously important in a world of ubiquitous simple machines. A home owner does not care nearly as much about saving a few seconds in effectively controlling the oven as they do about readily controlling every appliance in the house.

Speech interaction can learn further from experience with GUI standardization. When originally introduced, a standardized "look and feel" was presented as the key to learning transference between applications. Over time, however, we have learned that a uniform look (visual appearance) is not as important as a uniform feel (input behavior). Users readily adapt to scroll bars or buttons that have different stylistic appearances. They do not adapt as readily, however, to differences in interactive inputs. Uniform input behavior is important because the input behavior is invisible and must be remembered. This insight is critical in spoken language interfaces because all interaction is invisible. Uniform structure, terminology and input behavior are critical for an invisible interactive environment to be successful.

Standardizing the "sound and say"

We have focused on simple machines rather than intelligent ones and thus escaped the difficulties of artificial intelligence. We have identified lessons from the history of graphical interaction to guide the movement forward. The next issue is to identify how spoken language interfaces to simple machines should be developed. Our solution is a universal speech interface (USI) style. By this we mean a standardized "look and feel" (or, rather, "say and sound") across varied applications and domains. There are several components to such a style:

- **A universal metaphor**. A metaphor allows the user to map concepts from a familiar conceptual landscape into a new, less familiar domain. The success of the GUI revolution depended to a great extent on the desktop metaphor. A metaphor for human machine speech communication is even more important because the limited bandwidth of speech severely restricts the cues that can be provided at runtime --- it is important that the user have a good mental model of the application and of how to accomplish various goals within it. Ideally, a single, universal metaphor should be used to cover a wide range of application classes. Any deviation from the metaphor will require substantial dialog between the human and the machine to identify the differences.
- **Universal user primitives**. Many aspects of dialog interaction are universal, i.e. they recur in many applications. The list of such building blocks is quite long: recognition error detection and recovery, barge-in, taking or relinquishing the floor, answering multiple-choice questions, asking for help, navigating etc. There must be standardized ways for the user to express each of these. For a given application, idiosyncratic ways of expressing these primitives may be better. However, the uniformity and the usability that uniformity brings dominate other issues. When faced with a new spoken language interface to a simple machine the user should already know how to find out about that machine's capabilities. It is very important that the interactive scaffolding that supports users be uniform. Spoken language studies by Arons [Arons 91] strongly indicate that language for orientation and navigation must be consistent.
- **Universal machine primitives.** Machine primitives are similar to user primitives in that a small set of machine prompt and response types recur in many applications, and in fact constitute the great majority of turns in these applications. Standardizing these

machine-side universals will allow the machine turns to be shorter and clearer. This could be accomplished by using appropriate common phrases, by making use of prosody, and by packing the audio channel with non-verbal information (beeps, melodies and other earcons). For work in this direction, see [Kamm et. al., 1997, Shriver et. al., 2000]. If responses are uniform in style, user understanding is increased. Communicating with people with speech is much more error prone than visually because of the transient nature of audio. The user has no opportunity to study audio output nor to control its order of presentation. People expect to visually scan a display several times to understand its content. They are very much less patient when an audio help message is played for the third or fifth time. Uniformity of the output simplifies the user's extraction of what they need to know.

The use of a universal interface will greatly reduce the development time of spoken language interfaces. Much of the interaction can be standardized and included in an application framework. This limits the difficulty of producing new speech applications. To the extent that we reduce costs and shorten the time required to develop a new application, we empower the creation of thousands of speech applications rather than tens.

The use of the universal interface also simplifies the speech recognition problem. The uniform treatment of the universal primitives will result in a substantially lower expected branch-out factor for the user's language, with a direct impact on recognition accuracy.

We believe that such a USI is possible, and will be effective precisely because of the restricted nature of computing technology. Devices and applications function in terms of their information state. Communicating that state to the user and allowing that user to modify the state is fundamental to such interactive situations. Any interaction style that does this clearly, effectively and uniformly will to some level be successful as a USI. Virtually every programming language supports numbers, strings, groups of things (objects/records/structs), recursive composition of information etc. These same fundamental structures appear over and over because they reflect the way human being organize information. These same structures appear in every GUI toolkit for the same reasons. Though the dialog structure and supporting information required for spoken language information will be quite different, those same fundamental information structures will appear. A USI built around those structures should be effective in achieving the same transference of skills and generality of application.

A research agenda

It should be noted that the Universal Speech Interface approach to spoken language interfaces is a departure from the natural language-based approaches taken by others. The USI approach assumes that, at least for simple machines, the advantage of speech lies not so much in its "naturalness" but in the interactive contexts where screens, buttons and other more traditional interaction forms do not apply. There is a contrary hypothesis to ours, which states that natural language-based speech will be more effective than a uniform style. There is an additional contrary hypothesis that some other interactive modality such as buttons, lights, projectors, cameras etc. will be more effective in achieving ubiquity of interaction. The testing of these hypotheses is the heart of an interesting research agenda.

In comparing natural language with a constrained language the key measures are learnability, transfer of learning and robustness and expense of the technology. The key contention for natural language is that transfer of speech expertise from the human-human to the human-machine world will mitigate most learning and transfer problems. The contention for a USI is that the inherent limitations of each simple machine will produce uncertainty and confusion among users of a new natural language interface. There is some early evidence to support the assertion that users prefer a less efficient but more predictable and trusted interface[Walker 98]. Standardizing the interface will hopefully give users a more efficient path for discovering and utilizing simple machines by exploiting transfer among human-machine interfaces rather than from the human-human domain.

Initial learnability is also a key issue. Natural language presumably requires no initial learning. Any deviation from natural language will require initial training. First time GUI users do not understand double-click, right-click or drag, but once they do, they are empowered. A USI must have a shallow learning curve to be successful. A good example of this approach can be found in the Graffiti™ character writing system currently deployed with the PalmPilot PDA. Graffiti is an invented "language", which was engineered to optimize automatic handwriting recognition. It bears strong resemblance to the way people naturally write letters, digits and punctuation marks. The original, natural style of writing was modified minimally, only when it was prudent to do so for ease of recognition. As a result, it takes only a few minutes to become a productive Graffiti writer. Similar approaches may work in speech.

The only way to resolve the natural vs. constrained spoken language debate is by constructing many prototypes and a great deal of user experimentation. Researchers in User Interface Management Systems (UIMS) worked for years on a wide variety of approaches to inexpensive development. Most of them were discarded before widget libraries and Interface Development Environments (IDEs) such as Visual Basic became wildly successful. There is no reason to believe that the speech tools research process will be different. A key measure for the technological success of a USI will be the breadth of applications that its restricted language can effectively service. Many graphical user interfaces are very easy in Visual Basic and yet there are common types of interfaces that are very difficult within its framework. A test of the USI tools will be whether they can develop a large enough class of applications, cheaply enough to diminish the impact of those limitations.

A good USI system should have

- A shallow learning curve to get started with readily accessible mechanisms to build expertise both in the application and in the USI concepts themselves. GUI users do not learn command key shortcuts at first, but when they do learn them, their effectiveness increases on all applications.

- Standardized ways to accomplish standardized tasks. This is the key to the transference.
- Reliable and predictable handling of the vagaries of speech recognizers.

There are also serious experimental questions to be answered when comparing speech to other means for interacting with simple machines. The ambient nature of speech gives it both advantages and disadvantages over visual and manual techniques. The nature of when and how these tradeoffs occur is not well understood. The work of Oviatt sheds some light on these questions [Oviatt 99]. Speech has a much wider breadth of expression, but buttons are more reliable and responsive. Most of these comparison questions remain unanswered due to the lack of research in speech application tools. The advent of cheap recognition technologies opens this fruitful area of research.

Summary

Speech by its very nature opens up interactive possibilities that are beyond graphical user interfaces. We believe that understanding how those possibilities are different from the natural language problem will produce many more spoken language interfaces than ever before. In many situations interaction by text, pointing and/or using a display is not feasible or desirable. Whenever this is the case, speech is indispensable. When graphic interaction is possible, speech can still take its place alongside it. Research into the nature of speech interfaces and their tools is entering an exciting new phase. We have articulated here one vision of what speech interaction between humans and semi-intelligent machines could be like in the foreseeable future. Our ongoing efforts in this vein can be found in http://www.cs.cmu.edu/~usi and http://www.cs.byu.edu/ice. But these cannot but represent only a handful of points in this vast design space. It is our hope to see many more attempts in this direction.

References

[Arons 91] Arons, B., "Hyperspeech: navigating in speech-only hypermedia" Proceedings of the third annual ACM conference on Hypertext, 1991, 133 - 146

[Jacob 82] Jacob, R. *Using Formal Specifications in the Design of Human-Computer Interface*. **Proceedings of Human Factors in Computer Systems**, March, 1982, 315-22.

[Kamm 97] Kamm, C., Walker, M., Rabiner, L. "The role of speech processing in human-computer intelligent communication." **Speech Communication**. 23 (1997), 263-278.

[Newman 68] Newman, W. *A System for Interactive Graphical Programming*. **SJCC,** Washington, D.C.: Thompson Books, 1968, 47-54.

[Olsen 99] Olsen, D. R. *Interacting in Chaos*. **Interactions**, Sept 1999, 42-54.

[Oviatt 99] Oviatt, S. *Ten Myths of Multimodal Interaction*. Communications of the ACM. November 1999, 42(11), 74-81.

[Price 90] Price, P., *Evaluation of Spoken Language Systems: the ATIS Domain*, **Proceedings of the Third DARPA Speech and Natural Language Workshop**, Richard Stern (editor), Morgan Kaufmann, June 1990.

[Rosenfeld et. al., 2000a] Roni Rosenfeld, Dan Olsen and Alexander Rudnicky, "A Universal Human-Machine Speech Interface," *Technical Report CMU-CS-00-114,* *School of Computer Science, Carnegie Mellon University*, Pittsburgh, PA, March 2000.

[Rosenfeld et. al., 2000b] Roni Rosenfeld, Xiaojin Zhu, Stefanie Shriver, Arthur Toth, Kevin Lenzo, Alan W Black, "Towards a Universal Speech Interface", in *Proc. ICSLP 2000*.

[Shriver et al., 2000] Stefanie Shriver, Alan W Black and Roni Rosenfeld, "Audio Signals in Speech Interfaces", in *Proc. ICSLP 2000.*

[Walker 98] Walker, M. A., Fromer, J., Di Fabbrizio, G., Mestel, C, and Hindle, D. *What Can I Say?: Evaluating a Spoken Language Interface to Email*. Human Factors in Computing Systems (CHI 98), April 1998, 582-589.

[Zoltan-Ford 91] Zoltan-Ford, E. "How to get people to say and type what computers can understand" **International Journal of Man-Machine Studies** (34), 1991, 527-547.

[Zue 97] Zue, V, *From interface to content: Translingual access and delivery of on-line information,* **Proceedings of Eurospeech'97** (Rhodes, Greece), 2047-2050, 1997.