A Conversational Interface to News Retrieval

by

James C. Clemens

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

August 9, 1996

Author

Department of Electrical Engineering and Computer Science

August 9, 1996

Certified by

Chris Schmandt

Thesis Supervisor

Accepted by

F. R. Morgenthaler

Chairman, Department Committee on Graduate Theses

A Conversational Interface to News Retrieval

by

James C. Clemens

Submitted to the

Department of Electrical Engineering and Computer Science

August 9, 1996

In Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

As people find themselves travelling more and more, remote access to information becomes imperative. And as they find themselves trying to juggle more and more activities at once, the ability to multitask becomes imperative as well. NewsPhone, a conversational interface to news, attempts to provide a ubiquitous, speech only solution to the stated problems. As a telephone based service, NewsPhone uses speech recognition and speech synthesis to provide a user news information. Furthermore, as a project developed in the Speech Interface Group - as part of the News in the Future Consortium - at the MIT Media Lab, NewsPhone addresses the issue of a conversational interface and the underlying architecture necessary to support it.

Thesis Supervisor: Chris Schmandt

Title: Principal Research Scientist, MIT Media Lab

This work was supported by the News in the Future Consortium

Acknowledgments

Chris Schmandt, my advisor, gave me a chance. He took me in as a senior with hardly any experience in speech systems, and allowed me to learn the field by working on meaningful projects. His patience and understanding have made my final year at M.I.T. more than I could have asked. For this, I am indebted to him.

Lisa Stifelman, for always having time. Her experience was indispensable for this work. I thank her for making the past year more than the daily grind.

Matt Marx. If he hadn't been working in the area, this work may never have been. Taking time whenever necessary, helping me understand the tools used in this project, Matt was a great help.

Jordan Slott, the man with the answers. I cannot thank him enough for all of his help. His work with the Sun sound servers made this project possible. I wish him much luck in the future.

Minoru Kobayashi, Brenden Maher, Jeff Herman and Deb Roy, my fellow speech group members who were a constant source of ideas, anecdotes and fun.

Jeff Wong, Ross Yu, Jordan Slott, Eugene Lin, Matt Antone, and Charles Chen-Cheng my MEng buddies. Their friendships have made my time at MIT memorable. My respect each of them is unbelievable. I wish them all luck with whatever endeavors they pursue.

Walter Bender, Jack Driscoll and the News in the Future Consortium. Their constant support and experience with news and media was invaluable. I thank them for teaching me so much during my stay at the Media Lab.

Paul Martin and the members of the speech team at Sun Microsystems Laboratories. They have provided wonderful tools and technical help, and I thank them.

My friends from M.I.T. and elsewhere, Joe, Matt, Eric, Enzo, Mike, Glenn, Erich, Dave, Eric, Norm, Kareem, Martin, Ji and Chris. They have provided me with constant support and friendship. I cannot thank them enough.

Ultimately, my parents, James and Dianna, and my sister, Amanda, have given me love, hope, and shown me what dedication and perserverence can do.

Table of Contents

1. Introduction

1.1 What is a "Conversational Interface"?

In the speech community, the term "conversational interface" has become something of a buzzword. However, it's meaning is extremely ambiguous. As experienced speakers of the English language, we know that a conversation can take on many forms. A conversation can be long or short. The language we use can be free and loose when speaking with a friend, or restricted and carefully selected while interacting with a superior. A conversation can be extremely fluid or utterly frustrating.

The NewsPhone system is a "conversational interface" to news retrieval. When a user telephones the system, she communicates, through speech, with a computer. It is then the computer's job to determine the user's news interests, find relevant text articles, and read them back to the user. NewsPhone uses a speech recognizer to understand the user's spoken words, and a speech synthesizer to read the articles back to the user.

So then, what exactly does "conversational" mean in terms of the NewsPhone system? Does it simply mean that the user is asked questions and can respond? If the responses are just "yes" and "no," then this can hardly be considered conversational. In the context of NewsPhone, the interface is made "conversational" not just because the mode of communication is speech, but because the system is able to understand - to some extent - the context of the news articles in its database. This understanding allows the user to make queries about content specific information, "Is there anything about Westinghouse?" and allows the system to suggest related articles, "No, but there are two articles about Honeywell."

These two techniques are what make NewsPhone truly "conversational." In addition, other features such as speech recognition error handling, help capabilities and well-formed phrases also contribute to NewsPhone's conversationality. For example, by providing feedback that uses its predefined syntax and vocabulary, NewsPhone coaxes the user into speaking sentences it is prepared for. However, it is the ability of both system and user to speak content sensitive sentences, that makes the NewsPhone dialogue more than just a game of '20 questions'.

1.2 Motivations and Research Challenges

1.2.1 Why a Conversational Interface?

More and more, people are travelling. With business expanding domestically and globally,

people find themselves away from their workplaces and homes much too frequently. When people don't have a familiar newspaper on their doorstep in the morning, or new electronic mail popping up on a 21-inch display in front of them, remote access to news, mail and other services becomes vital. Cellular networks are growing, and services such as RadioMail [RadioMail] are allowing people to receive their email on portable personal digital assistants. Yet, the most ubiquitous and practical device for remote access applications remains the telephone. Built to support speech quality audio quickly and robustly the telephone lends itself ideally to a conversational interface. And although such an interface is certainly limited in some regards, it is extremely advantageous in others.

1.2.2 What are the Advantageous of a Conversational Interface?

A conversational interface is powerful because it allows a system to capture the expressiveness of human speech. For example, speech is natural - most people have been speaking for most of their lives. Speech is fast - people's experience with human conversation allows them to speak their thoughts quickly and precisely. Speech is also visually independent - it allows users to perform tasks with their eyes and hands while listening or speaking. In the context of NewsPhone, a news retrieval system, these reasons unite to give the user an interface that accepts efficient queries in a variety of situations.

1.2.3 How Does NewsPhone Organize its News Articles?

News articles form an information space of immense proportions. Although NewsPhone deals strictly with technology news, the diversity and amount of information contained in technology news is still tremendous. To deal with such a vast amount of information, NewsPhone attacks news articles from two directions, statically and contextually. First, NewsPhone creates static news categories, such as computers, electronics, media, etc.. This representation of news allows a user to easily find a category of interest and search the news articles in that section. The categories don't change and the user can form a mental model of the organization of the news. However, lists of articles are slow to scan. Perhaps a list would be sufficient for browsing several voice mails, but it would not be nearly sufficient enough for a long list of news articles.

In order to make news browsing more efficient, and to take advantage of a conversational interface, NewsPhone also has a content-based representation of the news. Each day, when new articles are retrieved, they are parsed and classified. This allows users to take advantage of the conversational interface mentioned early. Queries for specific content in articles can be made, and related articles of specific content can be offered.

1.3 The NewsPhone System Model

The conversational interface and information space just described work in harmony to

allow NewsPhone users efficient querying. A conversational interface provides for expressive input from users, but would be useless if there was no underlying structure to support the various methods of querying. Similarly, classification of news articles becomes pointless if there is no way to interact with a user such that the architecture can be exploited.

The NewsPhone system is depicted in Figure 1. The user interacts directly with the speech interface, which is responsible for pulling the query from the conversation and passing it to the search engine. The search engine then uses the query to navigate the pool of news articles, and find those the user may be interested in. These articles are then passed back to the speech interface which lets the user know that the query was satisfied.

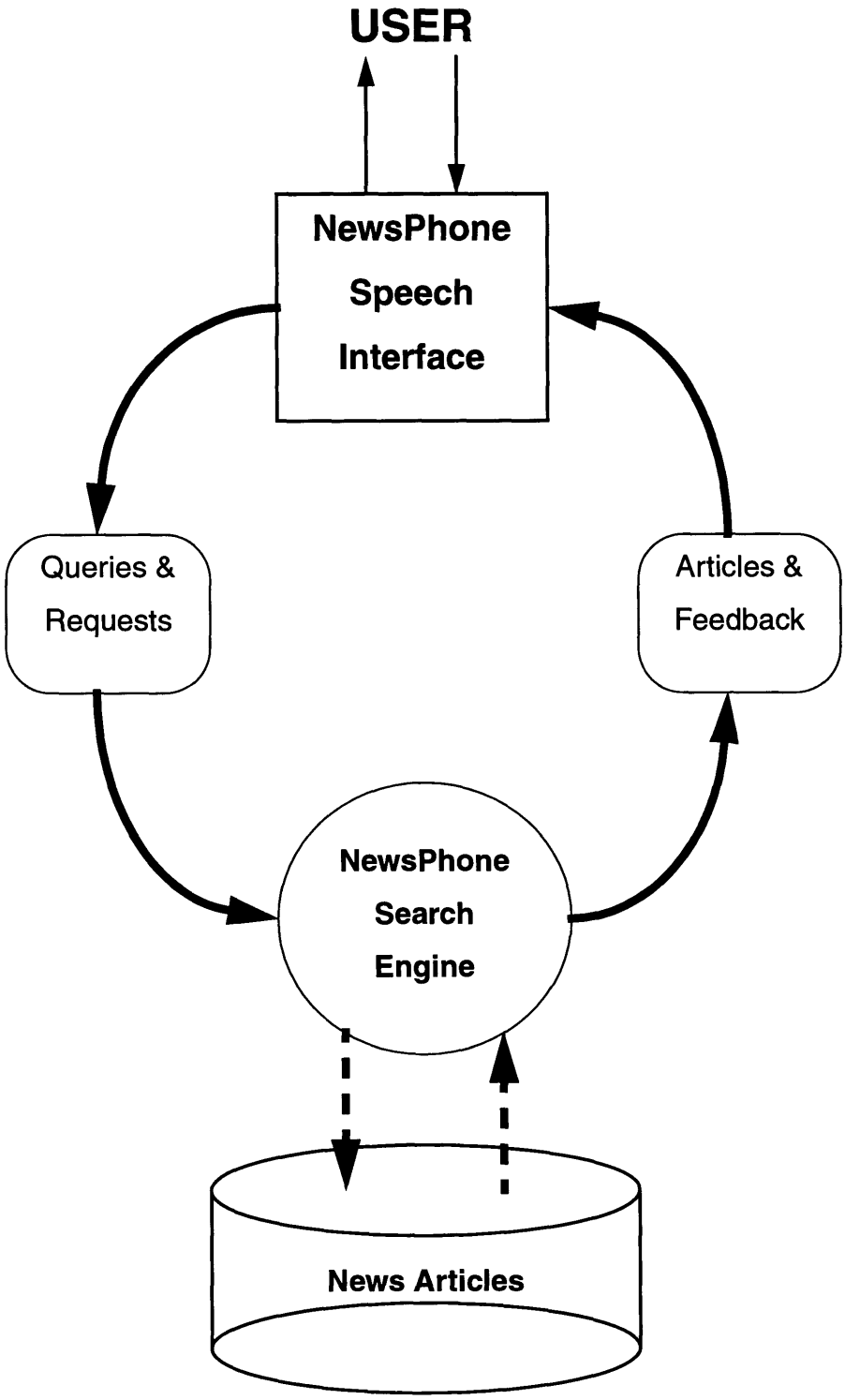

Figure 1. NewsPhone Engine

1.4 Overview of Document

Chapter 2, *Related Work*, describes the contributions of relevant work in the fields of speech recognition interfaces, conversational navigation, and news retrieval. Specifically, the NewsTalk [Herman 1995] system will be investigated in depth to show how, as the foundation for this work, it has influenced many design decisions.

Chapter 3, *The User Interface*, presents the navigation system as viewed by the user. Transcripts of actual NewsPhone sessions are thoroughly analyzed to show the rationale behind the various types of behavior the system exhibits. Implicit and explicit feedback are discussed, as well as error handling.

Chapter 4, *The Information Datastructures*, describes in detail the datastructures used to categorize the news articles. Presented in this chapter are the sacrifices and trade-offs that must be made to support various methods of navigation.

Chapter 5, *The NewsPhone System Architecture*, discusses speech interface design tools and their contributions to the NewsPhone system.

Chapter 6, *Thoughts and Future Work,* reflects upon the success and failures of the NewsPhone system. It also looks at the direction in which this research is headed.

2. Related Work

The research problems of conversational systems and news retrieval pursued in NewsPhone have a long history. The following chapter introduces some of those projects whose impacts upon NewsPhone have been fundamental.

2.1 Phoneshell

The Phoneshell system [Schmandt 1993] aims to present "the telephone as computer terminal." A user of Phoneshell is provided mobile access to desktop applications, such as email, voice mail, calendar information, weather, news, and rolodex over the telephone. Not only does Phoneshell allow a user to hear email messages and calendar entries, it allows her to respond to those messages and augment her calendar by either voice or text. The interface to these services was provided as touchtones. Fast and error-free, touchtones provide an extremely robust and intuitive interface over a telephone. By exploiting the telephone and it's keypad as an interface to audio and speech data, Phoneshell has become extremely successful, seeing prolific use by members of the Speech Interface Group.

2.2 VoiceNotes

VoiceNotes [Stifelman 1992; Stifelman 1993] brought a speech interface to a hand-held device to allow the recording, categorization and lookup of voice notes. These notes can be any spoken utterances. For example, thoughts, ideas and general remarks can all be stored digitally in the device and managed so that later lookup is quick. The system addressed three substantial research problems: capture and lookup of spontaneous speech, speech as a datatype, and speech as an interface to a hand-held device. The user had the ability to create new categories by saying "record" outside of a category and then speaking the new categories name. Recording a note was done similarly, by saying "record" while in a category and then speaking the note. Moving a note between categories was accomplished by saying "move" just after hearing the note read in the source category, and then speaking the name of the destination category.

Feedback in the VoiceNotes system is provided by both recorded speech and non-speech audio. For example, if the user requests a category by saying "things to do," the system responds by saying "moving into things to do" to explicitly confirm that the user's choice of action. Furthermore, when a user requests the next note in a category, VoiceNotes responds with the sound of a page flipping before it reads the note. This auditory icon serves to quickly tell the user that she was understood. A good mix of speech and non-speech audio helps keep the feedback provided by VoiceNotes fresh, while at the same

time reliably conveying important information.

2.3 SpeechActs

SpeechActs [Yankelovich 1994] provides a speech-recognition interface to a variety of desktop applications very similar to Phoneshell. It integrates third-party recognition and synthesis tools as well as an in-house natural language parser, SWIFTUS [Martin 1994], to provide users access to email, calendar information, weather, stock quotes, and voice and text messages. The SpeechActs project looked at the problems associated with translating typically graphically interfaced user applications to purely speech interfaced ones. In doing so, extensive user testing was done to determine what users were likely to say to the system, how they perceived information without visual cues and how to handle speech recognition errors.

2.4 MailCall

MailCall [Marx 1995] is a conversational messaging system, similar to SpeechActs, but dealing solely with electronic mail, voice mail, and voice dialing. Speech recognition is used to engage the user in a conversation, while a combination of mail filtering and random access allow the user to search his space of mail messages quickly and thoroughly.

Furthermore, the system uses a number of sources, such as a user's rolodex, calendar and email, to dynamically update the system's vocabulary. Therefore, if "Matt Howell" is listed in my rolodex, MailCall will recognize this and allow me to say "Call Matt Howell at work." MailCall makes significant strides in dealing with the problems of information navigation by means of these dynamic vocabularies.

2.5 NewsTalk

NewsTalk is the immediate predecessor of NewsPhone. NewsTalk is a telephone based news retrieval system, built upon a newspaper metaphor. When someone sits down to read a newspaper, she has the option of opening to one of a number of sections, e.g. business, sports, living arts, etc. NewsTalk emulates this process. Once engaged with the system, a user can say "go to business," to effectively "flip" to the business category. Upon entering a category, the user is presented with a series of news article headlines, to which she can respond "read it," or "skip it." The articles contained in these categories are determined by the NewsTalk personalized information agent. This agent has an understanding of each user's listening history, and uses this information to create two category groupings, one with the most popular stories, and one with the most relevant stories for each user.

As a conversational system, NewsTalk has many successes, as well as a few drawbacks. The input vocabulary is small, and therefore very robust. Misrecognitions are not very

common, and the system can be used in situations where telephone audio quality is not ideal or there is much ambient noise - for example over cellular telephones. Moreover, the interface is extremely intuitive. Once the user learns the categories she can toggle between, navigation is very fluid. At any point, a user can move out of one category an into another by stating "go to <category>." This also prevents users from becoming lost in a sea of news data. However, the fact that articles are arranged as lists under each category heading presents a problem. Speech as a communication medium is fast to transmit, but slow to receive. Thus, listening to a series of twenty headlines in the technology category can take a good deal of time.

Moreover, NewsTalk could have been implemented using only touch-tones. For example, "Read it" and "skip it" could be replaced by pressing the 1 key or the 3 key. NewsPhone attempts to go beyond sequential - touch-tone type - access and towards associative searching. Although NewsPhone makes use of static categories for quick reference, it also accepts content sensitive queries from users, and offers content related articles. Thus, no longer are articles just related to others in the same category, but can be related to any other article in any other category if NewsPhone finds they to have similar content.

2.6 Comparison of NewsPhone and Related Work

NewsPhone, like every speech interface system, has inherent problems. Speech is slow to

listen to, speech recognition errors are common, and synthesized speech is simply not very pleasant on human ears. The solutions to these general speech interface problems, as well as those introduced by both the NewsPhone interface, find their roots in the works just mentioned.

For example, the success of Phoneshell more than justifies the exploitation of the telephone as the preferred interface for remote access applications. SpeechActs and MailCall both prove that a speech recognition system can work, if misrecognitions are handled properly, and a system exists that allows users to efficiently perform the tasks at hand. Moreover, the importance of implementing explicit and implicit feedback, was shown by VoiceNotes to be essential in success of a recognition system. With these projects as predecessors, the NewsPhone system strives to take another step forward in the areas of conversational interfaces, and spoken news retrieval.

3. The User Interface

3.1 Overview

During a NewsPhone session, the user's speech and the computer's synthesized voice combine to form a dialogue. In typical a human conversation, the dialogue can be very free. Conversants can choose their words freely, form their sentences freely and move from subject to subject freely. However, the dialogue created between NewsPhone and its user lacks this freedom. Because the system has been programmed to understand only a fixed size grammar, and speak only a predefined list of sentences, the dialogue contains well defined boundaries.

The next section reveals pieces the NewsPhone grammar by transcribing a typical user session. Following sections will then discuss in detail the rationale behind the NewsPhone grammar design.

3.2 A NewsPhone Session

1. NewsPhone: "Welcome to NewsPhone. Your daily technology news source. Who's this?"

2. User: "Hi, this is Jim."

3. NewsPhone: "Hi Jim. Would you like me to read the top stories or would you like to go to particular section?"

4 User: "What sections are there?"

5. NewsPhone: "There are articles in the following sections Computers, Electronics, Media and Telecom. Which section would you like to go to?"

6. User: "Go to computers."

7. NewsPhone: "Going to computers. There are articles about the following topics, Apple, Microsoft, Comaq, and Dell."

8. User: "What are the articles about Apple?"

9. NewsPhone: "The 2 articles about Apple are, *Apple Gets Tough New COO, Apple Introduces NetFinder For Mac-Based Servers On Web.*"

10. User: "Read me the second article."

11. NewsPhone: "*Apple Introduces NetFinder for Mac-Based Servers On Web*. As part of its program to

make Internet easier to use for 'the rest of us,' Apple Computer released NetFinder to facilitate...

...It works with the Apple Internet Solution or any other Mac OS server running Quarter deck's WebSTAR."

12. User: "Read the previous one."

13. NewsPhone: "*Apple Gets Tough New COO*.

Last week, Apple Computer's newly appointed chief executive officer, Dr. Gilbert Amelio, outlined four new product"

14. User: < interrupts reading with a keypress >

15. NewsPhone: "yes?"

16. User: "Is there anything else about Apple?"

17. NewsPhone: "There are no more articles about Apple. But, there is a related article about Sun."

18. User: "Read it."

19. NewsPhone: "*Sun Licenses* Java Software Platform...

<reads to completion.>"

20. User: "Hang up."

21. NewsPhone: "Did you say you are hanging up?"

22. User: "Yes."

23. NewsPhone: "Ok. Good bye Jim."

3.3 Designing a Spoken Language Model

In human-human conversation, there are few circumstances in which one participant says something the other is completely unprepared to hear. However, in human-computer interaction, this is more than likely. Speech recognizers must be programmed to listen for specific utterances. If a user of NewsPhone or any speech recognition system does not phrase his words in one of the defined forms, the recognizer will simply not understand it. Therefore, it is extremely important that a speech recognition system coaxes its users into speaking phrases it can understand. The NewsPhone system does this by using directive prompts [Kamm 1994]: phrases that exhibit the same syntax and vocabulary the user should use. For example, the following question can be answered in several ways, all of which conform to the syntax and vocabulary used in the question itself.

NewsPhone:	"Hi Jim. There is new technology news. Would you like me to *read the top stories*, or would you like to **go to** a particular section?"
User Response 1:	"*Read the top stories.*"
User Response 2:	"**Go to** electronics."

User Response 3: "What sections are there?"

The first response directly answers the first part of the question, using exactly the same vocabulary and syntax. The second response is that of an experienced user. Knowledge of an existing category allows the user to jump to it, using the same vocabulary presented in the question. Finally, the third response does not exactly match the syntax of the question, but uses similar vocabulary. This response is necessary for inexperienced users, not familiar with the section headings.

3.4 Handling Errors

Speech recognizers are not perfect. If a user mumbles, slurs his words, speaks with a heavy accent, or is in a noisy environment, a recognizer will have a hard time understanding his speech. Even if the user speaks nearly perfectly, the recognizer may misinterpret him, since the models recognizers use to understand speech are probabilistic [Schmandt 1994]. These problems are only compounded by the fact that speech recognizers must be programmed to understand a fixed size grammar. Not only does a user need to pronounce his words clearly, he also has to phrase his words in one of the predefined forms. This suggests that recognition errors will be abundant.

Recognition errors can be categorized into three classes [Schmandt 1994]:

1. **Rejection**: the user speaks a word in the recognizer's vocabulary, but it is simply not recognized.
2. **Substitution**: the user speaks a word from the vocabulary, but it is understood as something else.
3. **Insertion**: the recognizer thinks that ambient noise - breathing, background noise, sniffling - is actually input.

It is important to note that the speech recognizer used in the NewsPhone system (DAGGER [Hemphill 1993]) lumps categories 1 and 2 together. This occurs because all input utterances are matched to a phrase in the recognition grammar.

The NewsPhone system deals with the errors just described by using a combination of explicit and implicit confirmations.

3.4.1 Explicit Confirmation

Explicit confirmations to user utterances are formed in two situations. The first is after NewsPhone hears a request to end a session. Since this utterance implies a terminating action, it is imperative that the system explicitly confirm this is the user's intention.

User: "Hang up."

NewsPhone: "Did you say to hang up?"

User: "Yes."

NewsPhone: "Ok. Good-bye."

The second instance in which NewsPhone would explicitly confirm a user's utterance is when the speech recognizer returns a low confidence score. For every user utterance, the DAGGER speech recognizer returns the predefined part of the grammar the utterance my have matched, and a confidence score, indicating the strength of the match. If the NewsPhone finds that DAGGER's confidence was below a certain level, then it explicitly confirms the utterance.

User: "Is there anything else about Mitsubishi?"

NewsPhone: "I thought I heard you say 'Is there anything else about Mitsubishi.' is this correct?"

User: "Yes."

NewsPhone: "There is a related article about Mitsubishi."

3.4.2 Implicit Confirmation

Implicit confirmations to user utterances are plentiful in the NewsPhone system. Formed by speaking with the same words in the user utterance, implicit confirmations serve a dual purpose. Not only do they tell the user whether her utterance was understood, but they also help reinforce the vocabulary and grammar of the system.

User: "What are the articles about Apple?"

NewsPhone: "The 2 articles about Apple are, *Apple Gets Tough New COO, Apple Introduces NetFinder For Mac-Based Servers On Web.*"

User: "Read me the second article."

NewsPhone: "*Apple Introduces NetFinder For Mac-Based Servers On Web.* As part of its program to make Internet easier to user for..."

The previous example points out two different styles of implicit confirmation. The first two lines demonstrate a very direct means of implicit confirmation. NewsPhone forms its response using the vocabulary from the user's input. If NewsPhone had said, "The 3 articles about Apex are..." the user would have immediately known there was a misrecognition. The second half of the exchange demonstrates a subtler type of implicit confirmation. By repeating the headline of the second article before actually reading it, NewsPhone tells the user right away which article it is reading. If there was a recognition error, and NewsPhone began reading the wrong article, the user would immediately know.

3.5 Navigation Techniques

NewsPhone is a conversational system. All communication between the system and the user is auditory. So, when using the NewsPhone interface to navigate through a collection of information, it is extremely easy for a user to become disoriented. If NewsPhone is to be successful then it must understand that users will get lost. It must also understand that being lost is frustrating. And a frustrated user will not be a user much longer. Employing a navigation technique that allows users to immediately find a reference point relieves the frustrations of becoming lost. At the same time, employing a navigation technique that allows users follow their fancies and get lost is fun. NewsPhone wants to be "user-friendly," and attempts to do both.

3.5.1 Static Categories

The NewsPhone system defines five static categories into which news articles are grouped. Static categories are necessary in NewsPhone for several reasons. First, as their name implies, they are static. A user can be confident they will not change from day to day, or during a session. Second, because the interface is purely speech, there can be no visual representation of the news article's categorization. Therefore, as was shown in [Herman 1995], it becomes necessary to create a physical space metaphor, through which users create a mental model of how the news is organized. Static categories become the foundation for this mental model, allowing users to jump to any category at any time simply by saying, "go to <category>."

Direct access to static categories becomes extremely important for navigation purposes in NewsPhone. Since users are allowed to move between categories by making content based queries, it is very likely that they will eventually get lost. Being able to jump back to a known category allows lost users to relocate themselves. The following example demonstrates this idea.

User: "Is there anything else about Sun?"

NewsPhone: "No, but there is a related article about Digital."

User: "Go back to media."

In the previous example, it is possible that the user began her news search in the media section. She may have heard an article about Apple, then been referred to an article about Sun, and then queried about Sun. The response to the query about Sun told her that there was a related article about Digital available. It did not give her any location information, only confirmation that there is related information about Digital. Not knowing exactly where she was, nor being very interested in news about Digital, the user requests to go back to the media section so she can begin her queries again.

3.5.2 Content Based Queries

While static categories are important for the reasons just mentioned, they are not terribly exciting. If a user relied solely on sequentially searching lists, news browsing for items of interest would become painfully tedious. To make news retrieval more dynamic, News-Phone allows its users to make content based queries. These searches present themselves in two situations, as demonstrated by the following example.

NewsPhone: "In the computers section, there are articles about the following topics, Apple, Microsoft, Compaq and Dell."

User: "Read me one about Microsoft."

NewsPhone: "Microsoft Wants No Part Of Internet Box. Bill Gates announced yesterday that..."

User: "Is there anything else about Gates?"

NewsPhone: "There is a related article about Gates."

The first situation in which the user was allowed to query about specific content was immediately after the system gave a summary of the articles in the computers section. By explicitly telling the user there are articles about specific companies, NewsPhone suggests that it can then accept queries about those specific companies. The second situation in which content was used for querying was after the reading of the article. At this point in the dialogue, the user was able to query the system about any proper noun encountered in the article. The process by which NewsPhone finds keywords in both the headlines and the

articles to use for content sensitive searching is explained in chapter 4.

3.6 Article Suggestion

As a conversational system, NewsPhone attempts to be active rather than passive. Instead of simply listening to user's requests and responding accordingly, NewsPhone initiates dialogue. For example, after the user has listened to an article, NewsPhone may suggest another related article.

NewsPhone: "Sales of ship-related products boost Nikon profits.
Major Japanese camera and computer product make Nikon...
... Nikon plans to introduce the next-generation F4 body, the spokesman said, but he declined to comment further."

User: "Is there anything else about Nikon?"

NewsPhone: "No, but there is a related article about Canon."

NewsPhone not only tries to determine the salient keyword associated with each article,

but it also tries to associate those keywords with each other. For example, in the above dialogue, NewsPhone suggested to the user an article about Canon. It was able to do this because Nikon and Canon are marked in the database as companies that have similar product lines. The process by which articles and their content is correlated is explained in detail in the next chapter.

4. The Information Datastructures

NewsPhone has two separate strategies by which it arranges news articles. The first is by grouping articles into five technology based static categories: computers, electronics, media, telecom and top stories. The second grouping is an internal representation that links articles of similar content together.

4.1 Static Categories

Every day, NewsPhone fills the static categories with news articles it finds in the ClariNet News database [ClariNet]. The static categories are implemented straightforwardly. Under any subject heading there is a list of articles, arranged in no particular order.

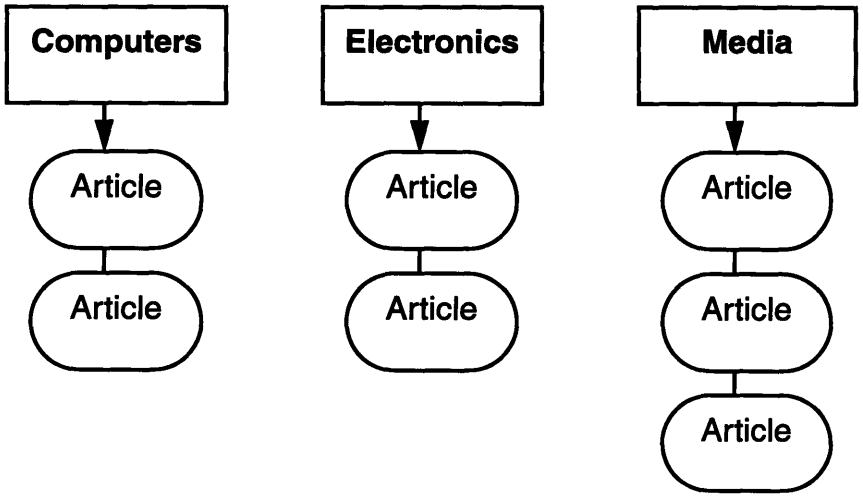

Figure 2. NewsPhone Static Categories

4.2 Content Relations

Understanding the limitations of static categories, NewsPhone seeks a second structural classification. In doing so, NewsPhone takes advantage of the fact that the news articles are text documents. By parsing them for content, the system then has the ability to link articles of different categories together if they share keywords. For example, an article containing the word 'Sun' in the computers section would be linked to an article in media, if it too contained the word 'Sun.' There are four separate ways NewsPhone relates articles in its database. They are explained in the following sections.

4.2.1 Salient Keywords

In the previous chapter, the idea of category summarization was presented. For example, when a user chooses to go to the computers section, NewsPhone may respond as follows: "`There are articles about the following topics, Microsoft, General, Sun, and Compaq.`" The system is able to do so, because it has classified each article by a *salient keyword.* Every time the NewsPhone system gathers an article, it looks at its headline. It then searches the headline for a word contained in a predefined list of technology specific proper nouns. If it finds a match, then it marks this word as the article's salient keyword. Hence, in the example just stated, articles had *Microsoft*, *Sun* and *Compaq* as salient keywords, while those unable to be classified were marked *General.* These keywords not only give the user some idea of the content associ-

ated with the articles in the category but also allow the user to request articles about a specific topic.

4.2.2 Salient Keyword Associations

As NewsPhone classifies articles by salient keywords, it also categorizes the articles. Each keyword in the predefined list has been grouped into a category. For example, Nynex and Bell South belong to TELCO, Apple and IBM belong to PC, Java and Netscape belong to INTERNET. These internal categorizations allow NewsPhone to perform article suggestion. If a user asks, "Is there anything else about Apple?" it first tries to find other articles that mention Apple. If there are no others, then it uses the fact that Apple falls into the PC category and suggests articles that have salient keywords of the same category. For example, NewsPhone may respond, "`No, but there is a related article about IBM.`"

4.2.3 Proper Noun Lists

NewsPhone always wants to allow a user to query by content. However, not all articles can be categorized by a salient keyword. Thus, another relation strategy must exist. Every article processed by NewsPhone is parsed to find all proper nouns. Thus, associated with each article is a list of its proper nouns. This allows the user to make very general content based

queries, by asking for other articles that also contain any proper noun found in the current one. So, if a user has just finished hearing an article about 'Sony', but 'Sony' was not found in its headline - and thus not marked as a salient keyword - the query "Is there anything else about Sony?" can still be successfully made.

4.2.4 A Fourth Relation System

Finally, the news articles are classified in yet another fashion. Using the SMART [Salton 1983] text correlation system, NewsPhone attempts to link any relevant articles in the entire corpus together. It may be the case that two articles of very similar content were not linked because one or both did not have a salient keyword. In this situation, the SMART system provides fault-tolerance. SMART takes as input, one profile article and a collection of relation articles. A correlation score is then calculated for the correspondence of each relation article to the profile article. NewsPhone performs a SMART correlation between every article in it's information pool. Those articles that have a correlation coefficient above a certain threshold are then linked.

4.2.5 Time Factors

All the article relations discussed in this section are done off-line. That is, when NewsPhone collects articles from ClariNet News it also processes them. Thus, the speed of the

system is not hindered at all by parsing articles for proper nouns or performing SMART correlations at run time.

4.3 Why Multiple Representations?

Static categories, as mentioned, have certain benefits. However, the fact is that they are slow, nonscalable and boring. The dynamic navigation techniques just described are what make the NewsPhone system fresh. At some point, even static categories may vanish from such a news retrieval system. Once the number of articles becomes large enough, categories are pointless. For example, if NewsPhone were dealing with thousands of articles, there would need to be either hundreds of categories with several articles under each, or several categories with hundreds of articles under each. Both choices are equally bad. However, a good dynamic navigation approach would make a database of thousands of articles efficiently searchable It is clear then that this is the selling point of NewsPhone, and that which helps it stand apart from more conventional news retrieval systems.

4.4 A NewsPhone Article

The picture below is the representation of an article that NewsPhone uses. Each article has a defined salient keyword, associated category, a linked list of SMART related articles that

point to other articles, and a linked list of proper nouns contained in the article itself.

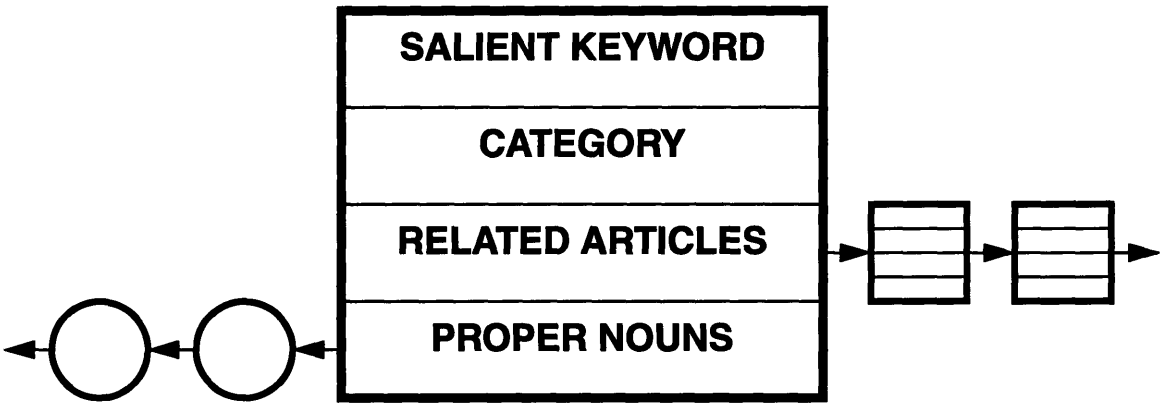

Figure 3. A NewsPhone Article

5. The NewsPhone System Architecture

The NewsPhone system consists of several key elements. Loosely, they can be grouped into two halves. The first half is the user interface and input grammar, detailed in chapter 3. The second is the organization of news articles, explained in chapter 4. However, both of these parts could not work in harmony without a number of tools. For example, the input grammar used in the NewsPhone system exists in two parts: one for the natural language parser, SWIFTUS, and one for the speech recognizer, DAGGER. This chapter looks at the many vital tools that were incorporated to make NewsPhone a working speech application.

5.1 Grammar Tools

5.1.1 SWIFTUS

The SWIFTUS system is a natural language parser developed at Sun Laboratories. The goal of the SWIFTUS parser is to provide a system designer an environment in which to

develop a speech recognizer independent "unified grammar." A unified grammar consists of two separate parts. The first is a language specification. Written by the designer of a speech application, a language specification defines the English language phrases that NewsPhone is able to understand. The second is a lexicon, or vocabulary. A lexicon incorporates the words that are used in composing the phrases found in the language specification.

Once a system designer has written a unified grammar, the SWIFTUS system compiles it into two separate grammars. One is a representation of the grammar to be used by SWIFTUS, and the other is a representation to be used with the speech recognizer defined by the system. It is in this sense that a unified grammar is speech recognizer independent. In the case of NewsPhone, the unified grammar file is `newsphone.ug`. This file is compiled by the SWIFTUS system to create two new files, `newsphone.gram` - for the SWIFTUS parser - and `newsphone.cfg` - for the DAGGER speech recognizer.

5.1.2 DAGGER

The DAGGER system is a real-time, software-based, speaker independent speech recognizer. This means that it runs in software concurrently with the NewsPhone system and can be used by anyone without training. DAGGER gets the `newsphone.cfg` file from SWIFTUS and compiles it for its own use. DAGGER receives all spoken utterances from the user, matches these against the grammar in `newsphone.cfg`. It then returns to

NewsPhone application the English text of what it believes the utterance to be, and a confidence score indicating how strongly it believes that it matched the utterance correctly.

5.1.3 Dynamic Updating of the Newsphone Grammar

As previously mentioned, every time an article is read to a user, NewsPhone updates its grammar so that it can listen for the proper nouns that appear in that article. Both SWIFTUS and DAGGER have mechanisms for updating the input grammar dynamically or "on the fly." Specifically, NewsPhone needs to update that part of the grammar that corresponds to proper noun queries. These are the phrases, "Is there anything else about <proper-noun>?" and "What else is there about <proper-noun>?" In the SWIFTUS unified grammar, `newsphone.ug`, these queries are coded as such:

```
{properNoun := "Is there anything" ["else"]
            ("about" | "on") ["the"]
            [sem=proper-noun] sem=proper-noun;
      head proper-noun;
      discourse-segment ^= 'query-ProperNoun;
}
```

In the DAGGER context-free grammar, `newsphone.cfg`, the same query is represented this way:

```
PROPERNOUN --->  is there anything[ else] (about | on)[
the]
```

```
[ SEM-PROPER-NOUN] SEM-PROPER-NOUN .
```

In the SWIFTUS example, the grammar will match any sentence "Is there anything else about <proper-noun>?" where <proper-noun> is any word defined in the lexicon and whose semantic part of speech is proper-noun. Therefore, to update SWIFTUS with the proper-noun, Amanda, for example, the following construct will be added to the `news-phone.lex` file:

```
(Amanda
  (
    (sem . proper-noun)
    (name . Amanda)
   )
  )
```

Similarly, the DAGGER context free grammar will match any query of the form "Is there anything else about <proper-noun>?" where <proper-noun> is any word in the grammar, and defined under the SEM-PROPER-NOUN nonterminal. Updating DAGGER is then parallel to updating SWIFTUS. However, DAGGER incorporates vocabulary in the same file as the grammar, while SWIFTUS does not. The following line is added to `news-phone.cfg` to add the proper-noun Dianna to the grammar:

```
SEM-PROPER-NOUN --->  Dianna .
```

Once both the SWIFTUS and DAGGER files have been updated with the new proper-nouns, they are recompiled by a simple function call. For SWIFTUS, NewsPhone calls:

```
swiftus_update("newsphone.lex");
```

For DAGGER, NewsPhone calls:

```
r_update_grammar(rfd,"newsphone.cfg",TELEPHONE);
```

The DAGGER function is part of the recognition server described later in the chapter.

5.2 Speech Group Servers

Over the years, the Speech Group at the MIT Media Lab has put together a variety of application programming interfaces (APIs) to allow developers to interface low level devices cleanly and easily. The following section describes those servers that play an integral role in the NewsPhone system.

5.2.1 The Audio Server

The audio server [Schmandt 1995] controls all audio data through audio devices on a given workstation. For example, all user utterances that are spoken into the telephone. However, it is the audio server that captures this speech and delivers it to DAGGER for recognition. The audio server is crucial in the communication between NewsPhone and its users.

5.2.2 The Phone Server

The phone server [Schmandt 1989] establishes a channel of communication between the

computer and the telephone. Any activity by the phone's hook is understood by the phone server. Thus, it knows when a user has placed a call to NewsPhone, and similarly, when a user has hung up.

5.2.3 The Recognition Server

Although DAGGER comes with its own API, NewsPhone does not make use of it. Instead, NewsPhone uses the recognition server [Ly 1993] developed by the Speech Group. All calls to DAGGER - for example starting and stopping recognition, and updating the grammar - are made through this API.

5.3 The NewsPhone Architecture

The following diagram demonstrates how the specific tools just describe interact to form the NewsPhone system.

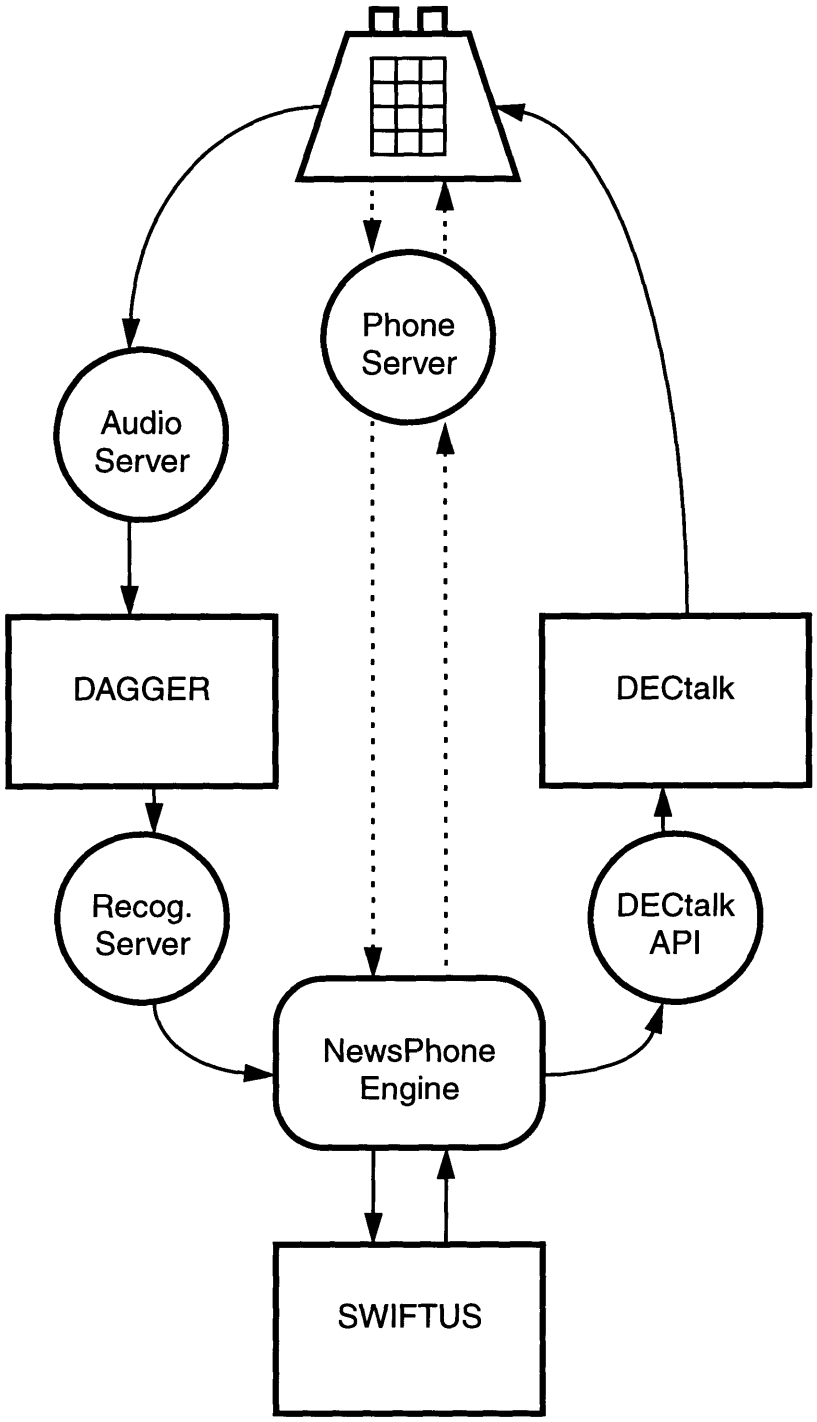

Figure 4. Overall NewsPhone Architecture

Using the previous diagram, the flow of information during a NewsPhone session can be explained. When a user dials in, the Phone Server picks up and tells the NewsPhone engine that a call has been placed. From this point on, all utterances by a user are captured by the Audio Server. The Audio Server passes the speech input to DAGGER to be processed. DAGGER matches the utterance against a predefined phrase in its grammar. The match, and a confidence score, are then sent to the NewsPhone engine as text. If the confidence is above a certain threshold, the text is sent to SWIFTUS. At this point the text utterance is parsed and sent back to the NewsPhone engine. NewsPhone determines the appropriate action or feedback and speaks through the DECtalk, which has a direct connection to the phone line. The user hears the synthesized speech and the process continues.

6. Thoughts and Future Work

6.1 The Contributions of NewsPhone

NewsPhone brings content based news retrieval to speech interfaces. As way found in [Yankelovich 1994], content based querying is extremely natural, and thus should be used in a news browsing environment. Using some simple parsing techniques, NewsPhone is able to find salient keywords in news article's headlines, and proper nouns in an article's body. Because NewsPhone uses the DAGGER speech recognizer and SWIFTUS natural language parser, it can update its grammar on the fly. This means that when a user requests an article to be read, the grammar is automatically augmented with the salient keyword and proper nouns corresponding to the article. The user can then query the system using content specific vocabulary.

6.2 Problems with NewsPhone

Some of the very same reasons that make NewsPhone successful also make it problematic. First, proper nouns present problems for the speech recognizer. An article typically has 10 to 15 distinct proper nouns in its body, but can have more than 25. Many proper nouns do not follow the con-

ventional rules of pronunciation that a speech recognizer uses. For example, the author's last name, 'Clemens', is pronounced as 'Kleemins' by DECtalk. If DAGGER's rules are similar, then it seems that upon introduction, I should say, "Jim Kleemins." This is hardly natural. Another example is the 'Hitachi' company, pronounced by DECtalk as 'HIT-a-chee' rather than 'hi-TA-chee.' Therefore, when an article is finished reading, and a user asks, "Is there anything else about Hitachi?" it is very likely that DAGGER believes he said, "Is there anything else about ITT?" It is clear then that propernouns present a formidable task for recognizers, system designers and system users.

Making the problem even worse, in the previous example everything but the last word in the question was matched with much certainty. Thus, DAGGER will return a high confidence score. This makes the process of determining when an misrecognition occurs in a content based query extremely difficult. The following example shows what could happen to a user in this circumstance.

User: "Is there anything else about Motorola?"

NewsPhone: "There are no more articles about Alcoa."

User: "I said, is there anything about Motorola."

NewsPhone: "There are no more article about Alcoa."

...

In such a situation a user can become extremely frustrated. If DAGGER were able to do partial recognition, telling the system that it understood everything but the last word with confidence,

NewsPhone would be able to handle there errors more elegantly. Moreover, if DAGGER were able to return multiple matches for the last word - for the above example, Motorola with confidence .7 and Alcoa with confidence .8 - NewsPhone again would be able to better satisfy the user. Until these abilities are a reality, NewsPhone must do it's best to avoid the "brickwall" effect demonstrated above and use a more progressive type of assistance [Marx 1995].

6.3 Future Work

The content based questioning and article suggestion used in NewsPhone are a first step, but are in no way complete. There is much work to be done as far as content understanding, more specific user queries, and the implications upon a speech only interface. With more correlation between companies and their officers, products, competitors and so forth, more advanced queries would be possible. For example, upon hearing an article about the release of the PowerPC processor, a user might be able to ask: "Did Bill Gates have any reaction to this?", "How does Intel plan to react?", "What did Apple's stock do yesterday?", "How soon will a PowerPC machine be available, and how much will it cost?". All these are legitimate questions when trying to find more information about a particular topic. Users should be able to ask intelligent questions concerning related material, rather than just finding out if there is more available on a certain subject.

Furthermore, in the future, NewsPhone may look to incorporate a personalized information agent as in [Herman 1995]. This would allow the system to have an idea of the user's news interests and thus be more responsive to queries about those areas. Instead of suggesting any article from a related area, the system could suggest a more specific article from that same area. For example, if

the system knew a user was interested in cellular phones, it may respond to a query about PacTel, with "`There are no more articles about PacTel, but there are several about McCaw.`"

As speech recognizers become more robust, and the general public becomes more familiar with speech interfaces, applications such as NewsPhone will demand more attention. This suggests that not only will technological advancement be necessary for popularizing speech interfaces, but more importantly time and exposure.

References

Chapanis 1975 — A. Chapanis. "Interactive Human Communication." *Scientific American*, 232, 1975. pp. 36-42.

ClariNet — http://www.clari.net

Davis 1989 — J. Davis and C. Schmandt. "The Back Seat Driver: Real Time Spoken Driving Instructions." *Proceedings of the IEEE Vehicle Navigation and Information Systems Conference*, Toronto, Canada, September, 1989.

Hemphill 1993 — C. Hemphill. "DAGGER: a parser for Directed Acyclic Graphs of Grammars Enhancing Recognition." Texas Intruments Inc., 1993.

Herman 1995 — J. Herman. "NewsTalk, A Speech Interface to a Personalized Information Agent." SM thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, May 1995.

Ly 1993 — E. Ly, C. Schmandt, B. Arons. "Speech Recognition Architectures for Multimedia Environments." In *Proceedings of the American Voice Input/Output Society*, San Jose, CA, 1993.

Martin 1994 — P. Martin and A. Kehler. "SpeechActs: A Testbed for Continuous Speech Applications." *AAAI-94 Workshop on the Integration of*

Natural Language and Speech Processing, 12th National Conference on AI, Seattle, WA, July 31-August 1, 1994.

Marx 1995 — M. Marx. "Toward Effective Conversational Messaging." SM thesis, Department of Media Arts and Sciences, Massachusetts Institute of Technology, May 1995.

Marx 1996 — M. Marx and C. Schmandt. "MailCall: Message Presentation and Navigation in a Nonvisual Environment." In *Proceedings of the ACM SIGCHI '96*, Vancouver, BC Canada. 1996.

RadioMail — http://www.radiomail.com

Schmandt 1989 — C. Schmandt and S. Casner. "Phonetool: Integrating Telephones and Workstations." IEEE *Communications Society*, November 27-30, 1989.

Schmandt 1993 — C. Schmandt. "Phoneshell: the Telephone as Computer Terminal." In *Proceedings of the ACM Multimedia Conference*, Anaheim, CA, August 1993.

Schmandt 1994 — C. Schmandt. Voice Communication with Computers. Van Nostrond Reinhold: New York. 1994.

Schmandt 1995 — C. Schmandt and J. Slott. "An Asynchronous Audio Server." MIT Media Lab Technical Report, 1995.

Stifelman 1992 — L.J. Stifelman. "VoiceNotes: An Application for a Voice-Controlled Hand-Held Computer." SM thesis, Department of Media Arts and

Sciences, Massachusetts Institute of Technology, May 1992.

Stifelman 1993 — L.J. Stifelman, B. Arons, C. Schmandt, E.A. Hulteen. "VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker." In *Proceedings of INTERCHI '93*, ACM, New York, 1993.

Stifelman 1995 — L.J. Stifelman. "A Tool to Support Speech and Non-Speech Audio Feedback Generation in Audio Interfaces." In *Proceedings of the ACM Symposium on User Interface Software and Technology*, Pittsburg, PA, November 14-17, 1995. pp. 171-179.

Yankelovich 1994 — N. Yankelovich and E. Baatz. "SpeechActs: A Framework for Building Speech Applications." In *Proceedings of the American Voice Input/Output Society*, San Jose, 1994.

Yankelovich 1995 — N. Yankelovic, G. Levow, M. Marx, "Designing SpeechActs: Issues in Speech User Interfaces." In *Proceedings of ACM SIGCHI '95*, Denver, CO, May8-11, 1995.