Employing Voice Back Channels to Facilitate Audio Document Retrieval

Chris Schmandt
Speech Research Group, Media Laboratory
Massachusetts Institute of Technology, 20 Ames St., Cambridge, MA 02139
geek@media-lab.media.mit.edu

Abstract

Human listeners use voice *back channels* to indicate their comprehension of a talker's remarks. This paper describes an attempt to build a user interface capable of employing these back channel responses for flow control purposes while presenting a variety of audio information to a listener. Acoustic evidence based on *duration* and *prosody* (rhythm and melody) of listeners' utterances is employed as a means of discriminating responses by discourse function without using word recognition. Such an interface has been applied to three tasks: speech synthesis of driving directions, speech synthesis of electronic mail, and retrieval of recorded voice messages.

1 Audio document access

This paper describes research in progress to develop a user interface to facilitate voice retrieval of on-line information over a telephone connection. Information may be *synthesized* from text such as human authored electronic mail or a response to a database query, or it may be *recorded*, for example a telephone message or a dictated document. We need to control the rate and order of presentation of such audio information for an efficient interaction. We desire to exploit those aspects of human dialog behavior whereby the listener gives cues to the information provider indicating comprehension and ability to keep up. We are attempting to build an intuitive and robust user interface based on the *duration* and *prosody* (rhythm and melody) of the listener's voice responses independent of any word recognition. Such an interface must take advantage of the *structure* of a normal information providing dialog.

This work focuses on what might be called *packet-oriented information transfer*, in which outgoing data is normally broken down into logical "chunks" presented sequentially. Between each packet the listener may respond to indicate status of the transfer, such as an acknowledgment or request for repetition or clarification. Our initial effort studied discourse behavior among humans while exchanging driving directions and led to a computer system synthesizing directions while listening for responses. The somewhat unexpected success of this first simple dialog interface motivated us to extend it to speech synthesis of electronic mail and playback of recorded telephone messages.

A variety of audio information systems are amenable to such packetization. Text-to-speech synthesis can provide telephone access to electronic mail [11,9]; although content is unknown since the text was human authored, punctuation and paragraph structure are indicative of logical units. Step-by-step instructions such as tutorials [8], driving directions [2], or assembly instructions naturally divide into paragraphs reflecting task structure. Recorded speech may be segmented by conversational answering machines [9,6,4] or by pause detection [1].

2 User feedback options

Because speech is slow and necessarily serial, efficient user interfaces to audio information systems are rare. For the packet-oriented voice applications mentioned above, the computer needs some method of flow control to be sure the user has received and understood the message. Without flow control control the machine must wait between packets long enough for even the slowest plausible user.

If access is *local* any number of user interfaces employing menus, mice, touch screens, keyboards, etc. may be employed and are likely to be more efficient than one relying on error-prone voice recognition. For example, a section

of an audio recording could be played, and the user requested to click a mouse button to continue to the next segment or click a different button to replay the current segment. Part of the attractiveness of audio information, however, is the possibility of greatly enhanced *remote* access over ordinary voice telephone circuits; but such systems must rely on acoustic feedback (including touch tones). This is the basis of our interest in voice interfaces to facilitate retrieval using audio media.

Speech recognition and touch tones are the obvious feedback channels but both are limited. Speech recognition does not work well over telephones because of variable noise levels and line characteristics. Recognizers understand limited vocabularies and speech must be spoken carefully to achieve good results, making them difficult to employ for naive users. Most recognizers are speaker dependent, limiting access to previously enrolled callers. Touch tones are reliable but not available from all telephones, and some pay phones disable tone generation once a connection is made.

Our desire to build user interfaces which need no prior explanation argues against both these channels. Both word recognition (because of limited vocabularies) and touch tone signalling employ *artificial* discourse constructs which must be explained by the system and learned by the listener. We believe a successful acoustic discriminator can be built given adequate knowledge of discourse structure and human conversational behavior.

If we observe human conversations, we find them rich in *back channels* [13] whereby the listener indicates by words, paraverbals ("uh-huh"), gestures, and facial expressions her degree of comprehension and interest in what is being said. These back channels facilitate the conversation by providing important cues to the talker as to the listener's degree of understanding and willingness to proceed to the next topic [7] even in an environment in which the only cues are acoustic. Our approach has been to classify the types of messages used as back channels in terms of discourse function, develop corresponding flow control abstractions for packet transmission, and attempt to discover robust acoustical cues based on pitch and energy to differentiate utterances.

3 Human use of back channels

To explore discourse behavior, several tasks were set up in which directions were relayed between subjects. Talker and listener could not see each other and were recorded on separate audio tracks. The listener was required to take adequate notes which would allow real navigation to the destination. The talker was allowed as much time as required to pre-plan the route, but was not constrained in any way in the presentation.

Our observations suggest that talkers break the route down into logical segments, or paragraphs, each containing a relatively simple set of instructions between landmarks. The talker pauses between each segment, which allows (or perhaps solicits) a response from the listener. The nature of these responses was our main interest, and we group them into four classes according to their function in information exchange.

- *implicit acknowledgment.* The listener says nothing; after a suitable timeout period (a function of the complexity of the most recent outgoing utterance) the talker assumes understanding and continues with the next paragraph. Especially over a voice-only telephone connection, the talker may engage in *channel checking* [5] if no response is forthcoming after several paragraphs. Channel checking ("Are you there?") demands a response and verifies the connection.

- *affirmative.* The listener indicates understanding by an *explicit acknowledgment.* Examples: "O.K", "uh-huh", "yup", "Yes, I know where that intersection is..." or "...right after the third light. O.K." This is a cue to the talker that the conversation can proceed without further delay, encourages the next portion of information, and fulfills a social role of reflecting the listener's interest in what the talker is discussing.

- *negative.* The listener indicates lack of understanding. Examples: "Take a right *where?*", "But I thought I was going to Cambridge.","*Ames* street?","How will I recognize Kendall Square?". Confusion may be *local*, i.e. a word misunderstood, especially a proper noun such as a street name, or it may be *global*, as in a loss of continuity in portions of the route or inability to relate the directions to some known landmarks.

- *synchronization*. The listener indicates a timing problem, usually needing more time to write down the directions. Examples: "Could you hold on a moment?", "Just a second", "Repeat that last part please." *Echoes* may also be used to hold the floor without contributing any new information or really taking a dialog turn; the listener repeats what she is writing down, perhaps with pauses or a slow speech rate. These utterances are somewhat problematic to classify, as the echo also serves as a verification mechanism.

Note that we do not consider responses which introduce a new topic. Without word recognition it is impossible to determine what this topic is, so the program could not discuss it. Instead, we simply act obstinate and refuse to recognize a topic change.

4 Simple acoustic classifiers

Given this taxonomy we want a classifier to determine the nature of a back channel response in real time based on acoustic evidence. In this section we describe our first pass at acoustic discrimination, an algorithm employing only utterance *duration* to classify speech input by discourse function. Although very simple, it is surprising how often this discriminator makes the correct decision. Even when it fails or produces ambiguous results, we may be able to take advantage of discourse structure or an explicit "repair mechanism" to get the dialog back on a known track.

The program pauses after each paragraph of speech output to listen for a response. The beginning of an utterance is detected when the average magnitude of the incoming signal exceeds background level, which is adjusted continuously to account for variable phone line noise levels. When speech is detected, the program waits for silence and then classifies the utterance by length. Short utterances (less than 800 ms) are treated as affirmative. Anything longer may be a request for more information or retransmission, or it may be a synchronizing response. Since this simple classifier cannot discriminate these two, a long response is classified as negative, as we would prefer to transmit redundant rather than insufficient information. Silence, of course, is an "implicit acknowledgment".

Although we cannot discriminate between synchronization and a negative response, these may be differentiated by succeeding behavior. After a long utterance response, the direction giving task waits a few seconds before repeating the previous paragraph. If the listener had in fact requested more information, she will wait patiently or repeat the request, which at this point is very likely to be long. On the other hand, "Just a moment..." is likely to be followed by "O.K.", so a short following utterance indicates that a synchronization request has just been satisfied and we have executed the correct dialog behavior.

5 Applications of the back channel interface

Our goal in studying the way people converse and building an acoustic classifier of back channel utterances was to use these as an interface to provide flow control for computer presentation of audio data. We began with a task which uses speech synthesis to give driving directions. The direction domain was chosen for several reasons: we have prior experience with it [2], it can be well structured and hence tractable, and most important it is an interaction with which everyone is familiar, allowing observation of human behavior under natural conditions, which is key to our approach. Later, we extended the interface to several other applications.

5.1 Direction giving

The original task involves the computer reciting driving directions between points in Boston using speech synthesis [2]. The outgoing text is divided into paragraphs and a pause after each allows a response from the listener. An affirmative (i.e., short) reply causes the program to proceed to the next paragraph. Silence (i.e., implicit acknowledgment) causes the program to continue to the next paragraph after a timeout; the lack of explicit confirmation is noted for in possible repair behavior later.

A long response can indicate either request for clarification or synchronization. We assume the former and repeat the same paragraph more slowly to compensate for difficulties understanding synthesized speech. Since we are not certain that a repetition was requested, we must state explicitly "I'll repeat that" to avoid the listener believing that we are reciting new information and adding a spurious link to the route. A second long response for a particular paragraph causes an alternate, more detailed explanation to be spoken.

Some special behavior is required to convince users that the program listens to them, so at the beginning of an interaction several questions are asked just for that purpose. If the listener never says anything during the direction recitation, the program will engage in channel checking, and then remind the user that it is listening as well as talking. At the other extreme, if the listener always responds with seemingly negative acknowledgments, the program invokes repair behavior, explains its capability, and offers "If you say nothing now, I will continue. If you say anything at all, I will repeat the route up till now."

We observed surprisingly successful interactions between naive listeners and the direction giving program; subjects were told only that the computer was going to give them directions to a local bakery. This led to interest in adapting the interface to other forms of audio information retrieval, for which there was not a simple "natural" parallel task in ordinary conversations.

5.2 Synthesis of electronic mail

A previous project had used speech synthesis to access human-authored electronic mail over the telephone [11]. Despite the difficulty understanding synthetic speech, especially given the mistakes and informal language structure of much electronic mail, about a dozen users had read their mail using this interface over a period of six months until we switched to new computers and discontinued the project.

The original mail reader had a backup strategy but no pausing between sentences or paragraphs. One of the touch tone keys acted as the "repeat" command and would cause the most recent sentence to be replayed at a slower speed (120 words per minute instead of the normal 170). A second "repeat" would cause the entire sentence to be spelled letter by letter. It was simple to extend this approach to the back channel interface, as we were already using sentence structure to segment the text message for flow control. A pause was added between sentences and the same listener response protocol employed. One difference we note is that the pauses between sentences need to be much smaller with mail than directions, as the listener is hearing a message for overall content, and rarely writing down as much information as when taking notes on directions.

With directions, however, we had alternate instructions as a secondary repetition strategy, an option not available with mail. We wanted to improve the "spell mode" fall back strategy as it had proven slow and awkward. Instead, we ran our output through a spelling checker and spelled out words not in the dictionary. This is in fact quite successful, as it catches most typos, which are impossible to pronounce intelligibly, as well as proper nouns and acronyms, which tend to be the most difficult words for successful text to sound rules [12].

5.3 Playback of telephone messages

A logical continuation was to extend the text mail message interface to voice mail. Some years ago we had built a conversational answering machine, the *Phone Slave* [9], which asked callers a series of questions and digitally recorded each answer into a separate sound file. Breaking the caller's responses into smaller chunks facilitates the owner's message retrieval and also tends to obtain more complete messages [4].

To apply the back channel interface, we treated each recorded response (typically 1 to 5 seconds of speech) as a "packet", playing them in turn with a pause to allow the listener's response. A single phone message consists of up to 5 such segments. In addition to the recorded audio portions, each message was preceded by a synthesized utterance, also treated as a packet, identifying the caller and message time.

The repetition strategy was again simplified, as the only option is to repeat the message. Even if we could play the

speech back at a slower rate, that does not really aid intelligibility if the original problem was a poor talker or noisy phone line. As opposed to previous cases, there is no need to comment "I'll repeat" or the like, as this is obvious from the recorded audio playback of human speech.

6 Implementation environment

All of the applications described above were programmed in C under Unix on Suns. For text-to-speech synthesis commercially available devices from Speech Plus were used. The back channel duration detection was done using an IBM PC with a Dialogic speech card running as a server to the Suns [10]. The server was also used for gathering and playback of telephone messages. Another version of the server employing a Texas Instruments speech card and Linear Predictive Coding speech analysis algorithm is used for pitch and energy extraction mentioned below.

7 Future work

Affirmative replies tend to be short, so we have assumed that all short replies are acknowledgments. But a number of very common questions are also short, such as: "Where?", "What?","Left?". We hope to detect these *short* questions on the basis of their interrogative pitch contours, and are currently conducting experiments in which pitch contours of recorded speech are modified to determine how listeners discriminate simple questions from statements.

A likely form of synchronization is *echoing*, wherein the listener repeats some portion of the directions to pace the interaction (as well as to confirm the directions explicitly). We need to better differentiate these echoes from requests for clarification, as length is inadequate. We believe listeners use *final lowering* [3] to indicate the end of their turn, which is if anything exaggerated by the synchronizing context, suggesting a drop in pitch as an echo cue. Another form of echoing is the repetition of each word as it is written. Since we speak faster than we write, this results in a very staccato reply, which appears to be unique to this function in the human discourse analyzed so far.

Interruption over telephone connections can be problematic. Because telephone systems use a single pair of wires to carry both sides of a conversation it is difficult to separate the received audio from that which is being transmitted. Without sophisticated echo-cancellation hardware, it is nearly impossible to use speech recognition to understand the interruption. We hope, however, to reliably detect the increase in energy of the received signal that indicates an interruption, and incorporate an appropriate response into our discourse strategy.

8 Conclusions

Understanding discourse structure and the natural use of back channel comments by cooperating humans has important power for improving speech interfaces to audio information systems. This domain is particularly relevant to timely or remote information which must be accessed by telephone. We do not see such an interface as a replacement for speech recognition or word spotting, but rather a parallel channel which can contribute useful information in a well designed context.

Although there are many possible refinements of the acoustical classifier of back channel discourse events, it is surprising how effective even a simple classifier can be. In part this shows us the adaptability of human speech behavior, of course, but it is quite encouraging. Most of the observations of the viability of this interface come from watching naive listeners use the driving direction task. The electronic mail synthesis and audio message playback systems have yet to be so tested, in part because it is difficult to design an experiment without explaining the interface in advance, which is exactly what we wish to be able to avoid!

9 Acknowledgments

Thanks to Lorne Berman for writing the main logic of the direction giving software, to Kevin Landel for work with observing human direction givers, to Mike McKenna for writing pitch and energy analysis tools, and to Jim Davis for discovering the direction giving domain in the first place and much work with the linguistics literature, as well as a review of this paper. This work was supported by DARPA, Space and Naval Warfare Systems Command, under contract number N00039-89-C-0406 and by NTT, the Nippon Telegraph and Telephone Public Corporation. Hardware support was provided by Sun Microsystems and Speech Plus.

References

[1] S. Ades and D. Swinehart. Voice annotation and editing in a workstation environment. In *Proceedings of 1986 Conference*, pages 13–28, American Voice I/O Society, Sept 1986.

[2] James R. Davis and Thomas F. Trobaugh. *Direction Assistance.* Technical Report, MIT Media Laboratory, Dec 1987.

[3] Starkey Duncan, Jr. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[4] Stephen Furner. *Rapid prototyping as a Design Tool for Dialogues employing voice recognition.* Technical Report TE R19/7/87, British Telecom, 1987. presented at poster session of European Conference on Speech Technology in Edinburg 2 Sept 87.

[5] Phillip J. Hayes and Raj Reddy. Steps towards graceful interaction in spoken and written man-machine communication. *Int J. Man-Machine Studies*, 19:231–284, 1983.

[6] H. Kojima, J. Nishi, and L. Gomi. A voice man-machine communication system based on statistical information received from telephone conversations. In *Proceedings of 1987 Conference*, pages 101–110, American Voice I/O Society, Oct 1987.

[7] Robert E. Kraut, Steven H. Lewis, and Lawrence W. Swezey. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 41(4):718–731, 1982.

[8] Nakatani L. H., et. al. TNT: a talking tutor 'n' trainer for teaching the use of interactive computer systems. In *Human Factors in Computer Systems, CHI 86 Proceedings*, ACM SIGCHI, 1986.

[9] C. Schmandt and B. Arons. A conversational telephone messaging system. *IEEE Trans. on Consumer Electr.*, CE-30(3):xxi–xxiv, 1984.

[10] C. Schmandt and M.A. McKenna. An audio and telephone server for multi-media workstations. In *Proceedings of the 2nd Workstations Conference*, IEEE, 1987.

[11] Christopher Schmandt. Speech synthesis gives voiced access to an electronic mail system. *Speech Technology*, 2(3):66–69, 1984.

[12] Murray F. Spiegel. Pronouncing surnames automatically. In *Proceedings of 1985 Conference*, American Voice I/O Society, Sept 1985.

[13] Victor H. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting*, pages 567–578, Chicago Linguistics Society, 1970.