AUDIO SIGNALS IN SPEECH INTERFACES

Stefanie Shriver, Alan W Black and Ronald Rosenfeld

Language Technologies Institute
Carnegie Mellon University
{sshriver, awb, roni}@cs.cmu.edu

ABSTRACT

This paper discusses a variety of types of non-lexical signals such as beeps, prosodic variation and speaker style changes, and we consider four cases in which such signals might be used to good effect. We discuss the results of user tests to determine if specific types of non-lexical signals are better in some situations than in others, and we discuss the advantages and disadvantages of using such signals.

1. INTRODUCTION

This paper discusses options for including non-lexical cues in audio output in order to convey pertinent information to users of speech interface systems. By non-lexical cues we mean any noises or supra-lexical features such as prosody or pitch which can be inserted or altered in an otherwise lexical string. These non-lexical cues can be arbitrary (e.g. a beep to suggest that an item is optional) or non-arbitrary (e.g. a ticking clock sound to indicate that the item in question is a time value).

We believe that non-lexical cues can be particularly useful for applications that have limited or no visual displays, or for speech output coordinating with visual displays that use icons or punctuation to convey important information. Such signals can also be used for situations in which quick interactions are important.

Non-lexical cues can also be valuable for helping to standardize applications. We discuss possibilities for non-lexical signals within the framework of the Universal Speech Interface (USI) [5], which is being developed at Carnegie Mellon University to provide a standard interface for communication between humans and simple machines such as information servers, cell phones, and DVD players. We believe that using consistent non-lexical cues in such an environment may help users of one USI application learn new USI applications quicker, since they will already be familiar with the meanings of the various sounds. Standard non-lexical signals also could help make development of new USI applications easier since output signals can be transferred from one application to another.

We focused our assessment of non-lexical cues on their appropriateness for situations that occur in the USI. The current USI prototype application is a movie line which allows users to access information about movies and theaters in the Pittsburgh area. The cases we have considered for non-lexical signals are:

- system confusion
- list continuation/finality
- feature specification
- lexical entrainment

While the information presented and the syntax with which it is delivered is specific to the USI and the movie line application, we feel that the cases examined are sufficiently general for the results to be used in many other applications.

We analyzed the following types of non-lexical cues for each of the cases above:

- prosodic variation
- inserting beeps or other noises
- adding beeps or other noises as background
- changing speaker style (e.g. changing speaker gender or synthesis style).

2. TESTING

For each type of phenomenon we wanted to investigate, we created utterances using the non-lexical cues listed above and consisting of phrases the USI movie line application would output when the phenomena occurred. The USI uses limited-domain synthesis created with the Festival speech synthesis system [2][3]. Most of the utterances were created by recording the same female, American English speaker used in the USI limited-domain system and inserting the appropriate non-lexical cues into the recordings. About a third of the utterances were generated using diphone synthesis of the same female voice; we then used the Sable markup language [6] to modify pitch, power or speed as the non-lexical cue on these utterances. The complete matrix of 33 utterances can be found at http://www.speech.cs.cmu.edu/usi/audiomarking.htm

We then tested the appropriateness of the utterances for each case by asking users to listen to and assess the recordings. In the first part of the test, users were asked to listen to one of three sets of 11 of the created utterances and then answer the question "What information did you get from it? (Be as specific as possible)" for each one. The subjects were given no context for the utterances, as we wanted to see whether the cues had any intuitive, non-contextual meaning for users.

In the second part of the test, users were asked "Which of these utterances best conveys <case>" for each situation and were asked to listen to and select the best option from all of the utterances created for that case.

We surveyed 14 people for this experiment; most of whom were unfamiliar with the USI project. The survey was administered via the internet and can be found at http://www.speech.cs.cmu.edu/usi/audiomarking.htm.

3. RESULTS

In general, we found that users did not ascribe any meaning to the non-lexical cues when asked to interpret the utterances without context. Most users simply transcribed the recordings: of 118 listener analyses of individual utterances, only 27 included references to noises or to the way part of an utterance sounded different. The class of non-lexical cues that was most reliably commented on was beeps and other artificial noises; 11 out of 24 analyses of recordings including these contained some reference to the sound.

The cue that was least frequently commented on was prosodic variation, particularly in the diphone synthesis utterances. Many subjects noted the unnaturalness and unintelligibility of the diphone synthesized voice, and these comments point to possible reasons why the imposed prosodic variation tended not to be noticed in these cases. First, if the synthesized speech is unintelligible, then listeners must devote more of their cognitive resources to decoding it, leaving less processing power available for perceiving variation. Second, if the speech is unnatural, any change in it may be perceived as part of the unnaturalness and not as a deliberate, significant variation.

The following sections will discuss the different cases and non-lexical signals in more detail and present the results of the second part of the user tests, in which participants were asked to choose the utterance that best conveyed the given information.

3.1. System Confusion

In the USI system, confusion can occur when a user has input invalid data ("February thirty-first") or when a user's command has failed to parse (which could be because of either a recognition error or an ill-formed command). The system could also work in confusion mode if the speech recognition engine has given the decoded input a low confidence score. The dual goal of the confusion cue is to make the user aware that a problem exists in their input and to point out specifically where the problem is.

We created eight utterances for system confusion, each of which said "theater is the Manor, date is February thirty-first," and a confusion signal was placed to indicate that the system did not understand "thirty-first." Each utterance included one of the following non-lexical confusion signals:

- Natural prosody, for which the speaker was recorded reading the sentence in a natural voice, but with rising, questioning prosody over the number "thirty-first."
- Increasing pitch via signal processing for the word "thirty-first;" the entire utterance was created with a diphone synthesizer.
- Increasing power via signal processing for the word "thirty-first;" the entire utterance was created with a diphone synthesizer.
- Reducing the speed of the word "thirty-first;" again, the entire utterance was created with a diphone synthesizer and the variation was created using signal processing.
- Insertion of a rising-tone, "question" beep between "February" and "thirty-first" in a natural recording of the utterance.
- Playing a steady tone in the background during "thirty-first" in a natural recording of the utterance.
- Switching to a male speaker for the word "thirty-first."
- Switching to the diphone synthesized female voice for "thirty-first."

A significant number of the participants in the survey selected natural prosody as the best utterance for indicating confusion. This coincided with our expectations, as it was the only one of the utterances whose correlation to the confusion phenomena is reasonably non-arbitrary, at least for native speakers of American English. This was also our first choice method of conveying errors in the USI system.

The disadvantage of using natural prosody lies in its very naturalness. Since it is not yet possible to automatically alter diphone-synthesized speech to reliably recreate the rising, stressed prosody of uncertainty, natural prosody is best used in limited-domain synthesis. However, because errors are by nature unpredictable, in order to cover the space of possible error utterances in a system, the recorded data in the limited-domain system needs to include "confused" recordings of virtually all the words in the domain in addition to the set of possible "normal" utterances. For the USI movie line application this proved to create too large a set to record reasonably, as the vocabulary contains about 800 words (and, as most are movie names, needs frequent updating). We suggest that natural prosody is best for indicating errors when limited-domain synthesis is used with a fairly small vocabulary.

The second most popular choice (though not significantly better than any of the remaining options) was the background tone played concurrently with the erroneous part of the sentence. This has the advantage of being fairly easy to implement in any kind of synthesis system. One participant pointed out that the tone we used in the survey sounded exactly like the call-waiting beep used in American telephones, which would certainly cause confusion if it were used in a phone-based system, so care must be taken to ensure that any sounds used do not already have meanings that might confuse the user.

3.2. List Continuation/Finality

Responses in the USI system often take the form of a list. Since recitation of a very long list of items to a user does not generally allow the user adequate time to process and retain each item, and because we want to encourage turns to be as brief as possible, USI lists are output in groups of three or four. Rather than explicitly saying something like "say 'more' to hear the rest of the list" each time a longer list has been split up, we would like to convey this information non-lexically.

For list continuation, we created eight utterances. The utterance used to test this case was a list of movie genres: “action, comedy, drama.” Three examples used the variations on pitch, power, and speed as described for system confusion, here with the change made to the last item in the list. Three utterances incorporated the background tone and speaker switching as described above, again with the change on the word “drama.” One utterance was created with a quick three-beep signal added after the last word in the list; we thought of this sound as an audio equivalent to an ellipsis (…). Finally, we created a natural prosody representation of the list in which the speaker did not use a falling tone at the end of the list.

Once again, participants selected the utterance using natural prosody as the best option for conveying the information in question. This surprised us a bit, as we felt that this particular natural prosody recording did not have an especially strong instance of a non-falling tone at the end, and also that the three-beep signal was not completely arbitrary (this signal was chosen as the next most popular option, however). That a significant number of subjects selected this option indicates that humans definitely have a good ear for prosody, and that it can be used fairly reliably even if the prosodic indications are not all that strong.

As with natural prosody for confusion alerts, using prosody to indicate list continuation requires a limited-domain synthesis system to achieve the best effect, and can require a substantial amount of extra recording if the contents of lists in an application exhibit a lot of variability. For smaller systems, or for those in which the information does not change frequently, we suggest that using natural prosody is an effective way to convey list continuation.

3.3. Feature Specification

USI commands are generally composed of slot + value phrases. At any point in an interaction, a user can ask "now what?" and get a list or description of what slots or values they can say next (in the form of “movie is dadada, theater is dadada…,” where dadada represents some user-specified value like a movie name). Non-lexical clues can signal features of the slots and values. Within the USI framework, we wish to indicate whether a slot takes a standardized USI value type (such as time, date, or amount), and if so, which one, and whether a slot is required or optional.

3.3.1. Value Types. We only created two recordings for value type specification, using the base utterance “show time is dadada, ticket price is dadada.” We used the sound of a clock ticking to indicate that show time takes a time value, and the sound of a cash register to indicate that ticket price takes an amount value. In one recording, the sounds were played as background noise during the dadadas; in the other they replaced the dadadas altogether. We did not feel that any of the rest of the techniques used for the other cases lent themselves well enough to value type specification.

Nearly all subjects preferred the recording in which the sounds replaced the dadadas, but since listeners only had two similar options to choose from, this is not a significant finding. An important point to note here is that non-arbitrary noises such as these are quite noticeable and identifiable; in part one, five out of seven analyses of these recordings made specific mention of the sounds.

The biggest disadvantage of using non-arbitrary sounds such as these is that it is not always possible to find appropriate sounds for all situations. Besides amount and time, the USI also uses date and number as standard value types, and we have not yet come up with suitable sounds for these cases.

An incidental finding related to value type is that users were not completely confused about the function of “dadada.” Subjects expressed the idea that “dadada” means “something” in about half of the part one analyses of utterances containing that phrase.

3.3.2. Required vs. Optional. For required vs. optional we created seven recordings, again based on “show time is dadada, ticket price is dadada;” the idea was to convey that the ticket price information was optional. As in the other cases, three of the recordings used pitch, power and speed variations on the ticket price phrase, although in this case power was reduced and speed was increased. One recording changed to diphone synthesis for the final phrase, one changed to a male speaker, and one inserted a short beep before the optional phrase. We also included a natural prosody example in which the speaker attempted to express prosodically that the ticket price information was optional.

Participants expressed no clear preference for any of the options in this category. This was expressed both by the overall variety of answers chosen and by explicit post-survey comments like “I really didn’t get the idea from any of the utterances…that one of the items was optional, [although] I thought I could make a distinction with all the other questions.” This was not very surprising to us, as all of the options seemed arbitrary, even the natural prosody example, as the speaker herself was not confident of the best way to express optionality through prosody. Optionality seems to be a fairly common feature however, and a non-lexical cue for it would be a valuable item, because it is not even clear that there is a simple, unambiguous way to express it lexically. As we continue to develop the USI, one thing we will assess is how easily learnable arbitrary non-lexical cues are, such as those for optionality.

3.4. Lexical Entrainment

USI prompts are phrased so as to "lexically entrain" the user [4]. That is, instead of asking "what movie do you want to see?" the system says "movie is dadada." This format encourages the user to learn the proper form of the command and in doing so should make the overall interaction faster and more user-driven. It also simplifies application development, since the information is always presented in a standard form.

Often the entire string output by the USI is an indication of what the user is expected to say, as in “movie is dadada.” If the user is expected to input something that is part of a very large class however, it is probably impossible to have complete lexical entrainment, but it may be appropriate for part of the string. An example of this case would be a string like “state a neighborhood or city, or say ‘is what?’” (‘is what?’ is the mechanism for querying in the USI).

We assessed eight non-lexical cues for their lexical entrainment effectiveness. Using the "state a neighborhood…" string noted above, we created recordings using the same eight techniques described previously for system confusion to highlight the phrase "is what," with the following differences: the inserted beep was a brief, steady tone, and the natural prosody was achieved by pausing briefly before "is what" and then pronouncing it in a slightly more distinct manner.

The participants in the survey preferred the change of speaker by a significant margin (the inserted beep and the natural prosody split the rest of the votes). This finding verifies the claim of Balentine and Morgan [1] that it can be useful to incorporate a "system operator/model user" duality to enhance lexical entrainment.

Speaker switching for lexical entrainment merely requires dividing the utterances recorded for limited domain synthesis between two speakers, and in a system like the USI in which the set of things the user is supposed to say is fairly small (large classes can be explained, as in "state a neighborhood…," rather than enumerated), this is a manageable task.

4. CONCLUSION

Our research has shown that natural prosody in limited domain synthesis can be an effective method for indicating confusion and list continuation. Since it can require a lot of extra recording to cover all possible errors and list configurations, it is probably not reasonable to use for large domains, but for small domains it can be used to good effect. Inserted noises are more easily implemented, and we plan to assess in future work how well inserted, arbitrary noises can be learned.

We have also found that switching speakers is a good method of effecting lexical entrainment, and that it seems difficult to clearly convey the idea of optionality. Our work has also reinforced the notion that diphone synthesis is hard to understand, and people are not likely to perceive variations in it. If it is necessary to use diphone synthesis, learned, arbitrary noises may be the best method for conveying information such as errors as users may not necessarily be able to understand fully lexical prompts.

5. REFERENCES

1. Balentine, B. and Morgan, D. *How to Build a Speech Recognition Application*, Enterprise Integration Group, San Ramon CA, 1999

2. Black, A. and Lenzo, K. "Limited domain synthesis," *ICSLP 2000*, Beijing, China.

3. Black, A., Taylor, P. and Caley, R. The Festival Speech Synthesis System, http://www.cstr.ed.ac.uk/projects/festival.html. 1998.

4. Boyce, S., Karis, D., Mané, A. and Yankelovich, N. "Speech User Interface Design Challenges," *SIGCHI Bulletin* Vol. 30 (2) p. 30-34. 1998.

5. Rosenfeld, R., Zhu, X., Toth, A., Shriver, S., Lenzo, K., and Black, A. "Towards a Universal Speech Interface," *ICSLP 2000*, Beijing, China.

6. Sproat, R., Hunt, A., Ostendorf, M., Taylor, P., Black, A., Lenzo, K. and Edginton, M. "SABLE: A Standard for TTS Markup," *ICSLP 1998*, Sydney, Australia.