Generating Time-Constrained Audio Presentations of Structured Information

Brian Langner[1], Rohit Kumar[1], Arthur Chan[2], Lingyun Gu[1], Alan W Black[1]

[1]Language Technologies Institute, [2]Computer Science Department
Carnegie Mellon University, Pittsburgh, USA
{blangner,rohitk,archan,lgu,awb}@cs.cmu.edu

Abstract

Presenting complex information in an understandable manner using speech is a challenging task to do well. Significant limitations, both in the generation process and from the human listeners' capabilities, typically make for poorly understood speech. This work examines possible strategies for producing understandable spoken complex information working within those limitations, as well as identifying ways to improve systems to reduce the limitations' impact. We discuss a simple user study that explores these strategies with complex structured information, and describe a spoken dialog system that will make use of this work to provide a speech interface to structured information in a more understandable manner.

Index Terms: speech synthesis, natural language generation, information presentation.

1. Introduction

What to do with complex information when it is necessary to use spoken presentation is a problem that has, and will continue to, occur in many speech applications. *Complex information*, here, refers to information that is most naturally represented in a more elaborate fashion than simple text. Very often such information will have an inherent structure to it, such as lists and tables. Problems with understandability exist even when using natural speech to communicate complex information, and these difficulties are exacerbated when using synthetic speech due to its decreased understandability. While there has been some work investigating the use of stylistic changes to improve the understandability of synthetic speech [1], our primary focus at this time is looking at lexical changes and strategies that make it easier to remember spoken complex information.

Frequently, we have seen automatic spoken dialog systems employ a "read out everything" approach, despite this naïve strategy often failing miserably at actually conveying information to a person, particularly when there is a large amount of information. Our goal with this work is to explore presentation methods that result in improved human understanding.

For this work, we will exclusively use complex information with a well-defined structure, primarily because structured information provides a slightly easier, more limited situation for audio presentation than general, open-domain, unconstrained information. It is our intent, however, to identify and employ strategies that will be effective for a wide range of information, including unconstrained situations.

Cognitive psychology provides some direction here. The primacy and recency effects [2, 3] are well-studied phenomena, suggesting important items should be presented either first or last. Furthermore, there is the standard "seven, plus or minus two" rule [4] that is commonly referred to regarding human memory. However, more recent work in the field suggests the number of items being presented is not the limiting factor in remembering them, but instead the length of the sound being listened to; humans have approximately two seconds of auditory memory [5] to work with. Exceeding that length of time significantly reduces the likelihood of successfully conveying information. Finally, there are several good reasons to avoid approaching these general limits when designing a human-computer interface [6]; they represent the upper limit of human perfomance, and interfaces which require a person to continuously function at or even near their limits will quickly be regarded as frustrating and stressful, making them less effective at their tasks.

2. Presentation Strategies

2.1. Requirements for Humans

Given that the limitations of human capabilities are unlikely to change, it is clear that we will need to tailor language generation systems to account for those limits. Specifically, the default behavior should be to limit the length of time of spoken utterances, since longer utterances in general will be harder to understand fully. However, simple tactics to achieve this, such as a "fast-talking" synthesizer or producing only severely abbreviated, disfluent language are not sufficient; at best these methods will be perceived as strange and unnatural, and at worst they will actively hamper human understanding. Fluent or mostly fluent utterances, despite being longer, are often more understandable because they are more like what the user is expecting to hear, and because conversational "filler" phrases can draw attention to the important information and then be ignored, thus not taking up the human's relatively short auditory memory.

Additionally, people often will want varying levels of information; a user asking for a restuarant's menu probably does not want to hear the names and descriptions of all 80 items on the menu but rather a higher-level description of choices like "seafood, steak, and pastas". In contrast, the answer to a query about available toppings at a pizzeria should not be "various meats and vegetables". Combining these requirements leads us to agree with Grice's *cooperative principle* of conversation [7]: answers should be infor-

mative, but brief and relevant; do not be more informative than required.

2.2. Proposed Solutions

For the structured information we are investigating, it can be reasonable to represent the information in a tree-like manner, such that the leaves of the tree represent the most detailed information available, and higher-level nodes are increasingly general summaries of their children. An example of this can be seen in Figure 1. Our proposed language generation system would thus follow a relatively simple algorithm to produce utterances. We begin by determining the maximum time for the utterance. In most cases, this should be five seconds or less, as a default, since this length takes human capabilities into account. Other possible lengths – based on user knowledge, user preferences, and conversational context, among other things – include 10 seconds, 15 seconds, and unlimited. The last should be used extremely rarely, only when absolutely required or explicitly preferred, as there will be a significant understandability decrease as utterances lengthen.

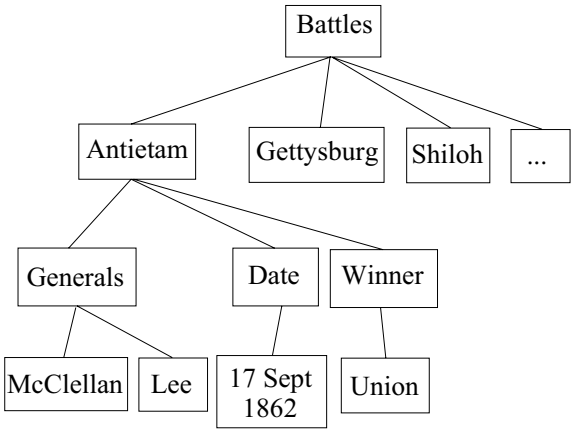

Figure 1: Example of structured information, represented as a tree.

We next establish the maximum pieces of information that can be presented in the allowable time frame. Fortunately, since we are producing synthetic speech, we will have easy access to the length of utterances that we are generating. However, determining this will require taking into account the information we are presenting (the information may be capable of being grouped logically to allow for chunking [4]) as well as any knowledge we have of the user (users familiar with the information being presented will more readily understand abbreviations, and thus we can make use of them). The default, when we have no knowledge of the user's capabilities, is to assume the user is unfamiliar with the subject being discussed. Often, establishing the optimal amount of information that can be presented will require domain-specific knowledge. While this means that the language generation system is also domain dependent, this is already often the case, and we also feel that a domain-independent system can show sufficient understandability improvements even without producing optimal utterances.

In cases where we have multiple pieces of information which are similar or related, either in a single response or in response to subsequent queries, we can use *tapered presentation* as in [8] to produce shorter utterances. The preferred behavior should be to start with the most detailed information that answers the user query, then generate the shortest possible utterance. If that utterance exceeds the allowable length, proceed up the information tree to the next most general node that can satisfy the query. This process should continue until either an utterance is produced that is within the time requirements or no utterance that answers the query can be produced in the allowable time. In the latter case, we should generate the shortest answer that can provide an answer.

Another strategy that we will examine is prosodic modification of the synthetic output, primarily to emphasize important information. Additionally, we would also like to use altered prosody to call attention to any difficult or unusual information, much as humans will do. However, this is significantly more challenging than other potential tactics, and thus more likely to be attempted after we have implemented the simpler strategies.

3. An Initial User Study

3.1. Description and Setup

In order to determine which strategies are effective at improving understandability, we designed and performed a simple user study. Subjects were asked to listen to 16 synthetically-produced utterances, with varying time lengths, and then describe in their own words the information from the utterance. The utterances were generated using a modern, high-quality commercial synthesis engine. All of the utterances contained information about papers or sessions from Interspeech 2005; several examples can be seen in Figure 2. The utterances took the form of fluent sentences; that is, they did not simply say "25 recognition sessions", but "There are 25 speech recognition sessions". As we did not want to explicitly test memory, but understandability, subjects were allowed to hear utterances twice. Utterances had one of four lengths – under 5 seconds, under 15 seconds, under 30 seconds, and unlimited – and fell into three distinct stylistic categories: naïve "read everything", summaries produced by a freely available text summarization tool, and utterances that could be produced by the algorithm described above. Eight subjects participated in this study, all of whom were familiar with speech technology.

> There are 12 poster and 12 oral sessions on Wednesday.
> There are 9 speech recognition, 3 signal processing, dialog, and prosody, and 4 other sessions on Tuesday.
> "Influence of syntax on prosodic boundary prediction" is an oral presentation on Wednesday at 10:40 am.

Figure 2: Example utterances from the user study. These examples are from the two shortest time categories.

Subjects were evaluated in two ways; first, on how many of the concepts present in each utterance they understood, and second, on how many concepts associated with the information in the utterance they understood. These can differ because shorter utterances may not attempt to convey as much of the information as the longer ones, and thus could look artificially better despite providing less information to the user. For example, an utterance

could describe to the user 5 out of the 10 total concepts related to a query, of which the user understood 4. The user would thus have 80% understanding of what was presented to him, but only 40% understanding of the full answer to the query. This contrasts with an utterance presenting 9 out of 10 concepts for which the user correctly understands 5; this would result in 56% and 50% understanding rates, respectively.

3.2. Results

Overall, subjects did quite well at understanding the content of the shorter utterances. Not unexpectedly, their performance dropped steadily as the utterances became longer, with exceedingly long utterances showing very poor understanding. These results can be seen in Table 1.

Utterance Time	Correct (Presented)	Correct (Total)
Under 5 seconds	95%	42%
Under 15 seconds	56%	33%
Under 30 seconds	44%	34%
Unlimited	13%	13%

Table 1: Results from this user study, grouped by utterance length

It is interesting to confirm that presenting more information does not necessarily result in more information being understood; in fact it is clear that presenting too much information has a detrimental effect on overall understanding, as can be seen with the Unlimited category. The reverse also seems to be true; the best overall understanding occured when only small percentages of the information were given (the under-5 category), though the result is not statistically significant. It could also be that the shorter utterances had fewer concepts to convey, in which case they would appear superficially better. More results exploring this are needed to explain the cause.

It should be noted that most subjects, when encountering longer utterances, commented on the increased difficulty of remembering and understanding what they were being told. For the time-unlimited utterances (some of which could exceed two minutes in length), *every* subject indicated they felt the task was too difficult. This would seem to be supported by the relatively poor performance for those utterances, and strongly suggests system utterances should not exceed 30 seconds for any reason.

There was no noticable improvement by using automatically generated summaries as opposed to reading the full abstract, and in some cases, the full abstract had a higher percentage of provided concepts understood. We feel that this is most likely due to the overall low quality of the summaries produced by the freely available tool; to fully test this strategy, we intend to use state-of-the-art automatic summarization tools in the future, as well as human-generated summaries, in order to gauge the effectiveness of summarization for understandability.

4. Building a Real System

Related to this work on language generation and presentation, we are developing a spoken dialog system that will be able to provide conference information. The intent is for such a system to be used by conference attendees to get information about the schedule, sessions, and papers, among other things, specifically during Interspeech 2006 in Pittsburgh. We are developing this system using the Ravenclaw Dialog Management Framework [9]. Users will be able to able to ask unstructured queries to obtain relevant conference information. We are using Sphinx3 [10] for speaker independent speech recognition. The language model is trained on valid utterances generated from the task grammar, with the recognized input parsed using a robust parser. An example interaction with the proposed system is shown in Figure 3.

System: Welcome to the Interspeech 2006 Information Line. What information do you need?
User: Are there any synthesis sessions on Tuesday?
System: Yes, text-to-speech 2 is at 10:30 am on Tuesday.
User: Tell me about that session.
System: Text-to-speech 2 is an oral session with 5 presentations, in room 116. Do you want to know about the papers?
User: Who will be presenting in the session?
System: The presenters are not known, but Yannis Pantazis, Gabriel Webster, Tomoki Toda, Ulrich Reubold, and Makoto Tachibana are the first authors for these papers. Can I help you with anything else?
User: What is the title of Toda's paper?
System: Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. Do you want to know the abstract?
User: No. Thank you.

Figure 3: Example interaction with our proposed system.

Once the desired information has been retrieved, we will employ the presentation strategies outlined above as required. Since most of the conference information will be too long or detailed to present quickly, it is likely that most queries will need to be answered by using some or all of the discussed strategies.

By deploying this system in a real-world environment, we hope to collect a significant amount of data from users who are genuinely interested in the results. We feel that such users are far more suitable to evaluate the effectiveness of the presentation strategies; users who care about the results will not allow system errors, misunderstandings, or imprecision that causes incorrect or incomplete answers to succeed.

5. Conclusions and Future Directions

There are several interesting results to take from the study above. Notably, that utterances longer than 30 seconds or so are so poorly understood that producing them is effectively worthless. Human listeners seem overwhelmed by the amount of information in long utterances and simply "give up" at understanding them – in fact, the only concepts the study participants understood were names that were previously familiar to them. Any novel information was simply lost. For information-giving systems, that scenario represents complete failure, and needs to be avoided. Despite being high quality speech synthesis, the prosody of the synthetic speech is still clearly unnatural, and while it is likely that understandabil-

ity was degraded at least in part due to that, it is still the case that long utterances are simply not understood well. Working within these limitations is undoubtably required for successful information presentation.

However, when there is a large amount of information to be presented, it is still currently unclear what the best approach should be. Obviously, presenting a significant amount at once will not be understood, but presenting information in shorter utterances requires "leaving out" possibly important information. This can be dealt with, at least partially, by having an interactive application, but there are some speech applications, such as a news reader, that are not or should not be interactive. These applications have an even higher requirement to balance the amount of information in utterances with the length of the utterance; achieving this balance is most likely the key to improving the understandability of speech systems. Interactive applications, such as those involving spoken dialog systems, have the advantage of being able to provide some information to the user, and then reacting based on whether the provided information was what the user wanted, accurate but not sufficiently detailed, or not the right information.

Since interactive applications are, in some senses, easier to implement and evaluate, our focus with this work will be on those systems in the near future. Certainly, with the conference information system we are building, we have the expectation of gathering data in a real setting with the intention of evaluating these strategies in a live interactive system. There are other applications we are planning to investigate, such as large document summarization. Determining an understandable method of interactively presenting a long, complex document, especially without frustrating or annoying the human user, is a major goal of this work.

6. Acknowledgements

This work is supported in part by the National Science Foundation IGERT Fellowship for assistive technology.

7. References

[1] B. Langner and A. Black, "An examination of speech in noise and its effect on understandability for natural and synthetic speech," Tech. Rep. CMU-LTI-04-187, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2004.

[2] B. B. Murdock, "The serial position effect of free recall," *Journal of Experimental Psychology*, vol. 64, pp. 482–488, 1962.

[3] W. A. Bousfield, G. A. Whitmarsh, and J. Esterton, "Serial position effects and the "Marbe effect" in the free recall of meaningful words," *Journal of General Psychology*, vol. 59, pp. 255–262, 1958.

[4] G. A. Miller, "The magical number seven plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 63, pp. 81–97, 1956.

[5] A. D. Baddeley, N. Thomson, and M. Buchanan, "Word length and the structure of short-term memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 14, pp. 575–589, 1975.

[6] D. C. LeCompte, "3.14159, 42, and 7±2: Three Numbers That (Should) Have Nothing To Do With User Interface Design," http://www.internettg.org/newsletter/aug00/article_miller.html, 2000.

[7] H. P. Grice, "Logic and conversation," in *Syntax and Semantics: Speech Acts*, Cole and Morgan, Eds., vol. 3. Academic Press, 1975.

[8] N. Yankelovich, G. Levow, and M. Marx, "Designing SpeechActs: Issues in speech user interfaces," in *Conference on Human Factors in Computing Systems*, Denver, CO, 1995.

[9] D. Bohus and A. Rudnicky, "RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda," in *Eurospeech03*, Geneva, Switzerland, 2003.

[10] Carnegie Mellon University, "The Sphinx 3 recognition system," http://www.cmusphinx.org/.