SPOKEN LANGUAGE GENERATION IN A MULTIMEDIA SYSTEM

Shimei Pan, Kathleen McKeown

Columbia University
Department of Computer Science
New York, NY 10027
{pan, mckeown}@cs.columbia.edu

ABSTRACT

In this paper we address two important issues in generating spoken language within a multimedia system: the design of a speech generator to facilitate coordination between media, and extensions to the functionality of a written language generation system to produce natural speech output. We demonstrate how a speech generator can produce information that allows for temporal coordination between multiple media. We describe how our speech generator takes advantage of rich and accurate syntactic and semantic information during text planning and speech realization. This enables the system to accurately predict, generate, and utilize prosodic features to facilitate coordination of speech with graphical actions such as highlighting.

1. INTRODUCTION

The spoken language generation system described here is a component of the MAGIC system (Multimedia Abstract Generation for Intensive Care), a testbed system for generating multimedia presentations which inform caregivers about the status of a patient who has undergone Coronary Artery Bypass Graft (CABG) surgery. Three media generators are involved: a 3D *graphics generator*[15], a *speech generator*, and a *text generator*. The speech, graphics, and text are all generated on the fly. In this paper, we show how the design of a spoken language generation system can facilitate the production of a coherent multimedia presentation by providing information to a *media coordinator*, a component which negotiates between media generators.

In this paper, we focus on three topics:

1. Augmenting the functionality of the language generation system to support coordination between spoken references and the graphical objects they refer to.
2. Predicting the location and relative strength of pauses to facilitate temporal synchronization between media.
3. Augmenting the functionality of the language generation system for the generation of speech as opposed to written language.

2. RELATED WORK

Our research builds on techniques developed for language generation [9], incorporating components that handle the tasks of content planning, lexical selection, and syntactic sentence generation. Our speech generator uses the FUF/SURGE package [6], a sentence generation system containing a robust English grammar. Unlike previous work in language generation, our work focuses on the development of techniques for producing spoken language. We used Lucent Bell Laboratories' *Text-To-Speech (TTS)* system for speech synthesis.

For a general speech generation system, a major task is generating appropriate prosody for speech output. Most research in this area is related to TTS synthesis [1, 8, 10], which uses stored text as input. A major problem for TTS systems is that they must analyze the underlying linguistic structure from the text during the text-to-speech conversion and then assign prosody based on the results [3]. However, the results of this analysis are usually not accurate enough to predict prosody fully. In order to avoid such limitations, a smaller amount of research has been done in generating speech from concepts, termed *Meaning-To-Speech or MTS* [5, 12, 14]. Like other MTS systems, MAGIC builds a full semantic and syntactic representation for the text as part of the generation process. This kind of information has been used to produce better prosodic structures in several MTS systems, such as the given/new distinction or contrast has been used to determine accentual patterns and intonational contours [5, 12]. Our work differs from other MTS systems both in the use of a large-scale language generation subsystem, as well as in our focus on generating speech that is compatible with requirements from other media.

Of the few systems that also coordinate speech with graphics, either a unified generator is used to produce speech and graphics simultaneously and thus, no specific negotiation between speech and graphics is necessary [11], or speech controls the process of synchronization, thereby simplifying the language generation task [12]. In both of these systems, the flexibility in coordination is more limited than in MAGIC.

3. COORDINATING SPEECH WITH GRAPHICS

In MAGIC, the communicative goal is to convey a patient's status to caregivers. Text, animated graphics and speech are used simultaneously to achieve this goal. Since multiple items may be displayed graphically at any one time, MAGIC uses synchronized speech and highlighting to help the user focus on the current topic and find the illustrated information referred to by speech. In order to achieve a coordinated presentation, our work uses negotiation [4] to arrive at an ordering of spoken references that is compatible with the order of highlighting. We want to use a highlighting order that is regular and does not jump around the screen; at the same time, speech should be natural. In addition to coordinating ordering, spoken references must be synchronized in time with highlighting, so that an item stays highlighted while it is being verbally described. In this section, we describe how the speech generator produces different orderings to make the task of media coordination easier.

Input to the speech generator is represented in a hierarchical presentation plan produced by MAGIC's high level content planner [4]. The task for the speech generator is to determine how to order the *basic information units* of the plan in speech, where a basic information unit corresponds to the lowest-level goal, or smallest unit, within the plan. Ordering in speech is influenced by how information units are distributed across sentences, by the words selected to lexicalize information units, and by the ordering of words within a sentence. Since there is considerable flexibility in ordering due to paraphrasing, the speech generator produces a representation of possible orderings, ranked according to preference. These orderings are then passed to the media coordinator for negotiation with graphics as described in [4]. For example, three possible orderings of a patient's demographics information can be produced:

1. Ms. Walker is a 50 year old anorexic, hypokalemic female patient of doctor Longman undergoing CABG.
2. Ms. Walker is a 50 year old female patient of doctor Longman undergoing CABG. She has a history of anorexia and hypokalemia.
3. Ms. Walker is a 50 year old female. She has a history of anorexia and hypokalemia. She is a patient of doctor Longman undergoing CABG.

MAGIC's speech generator determines the possible orderings in two stages, speech content planning and lexicalization. First, the speech content planner distributes information units among sentences. It attempts to place as many information units into a single sentence as possible, using modifiers (e.g., adjectives) since this will result in fewer words in the output (see [13] for aggregation techniques).

After this, the lexical chooser selects appropriate words for each information unit. On selection of the verb, the overall sentence structure is determined; the choice of verb controls the type of verb arguments that will appear and it is at this point that a mapping between information units and verb arguments are made. Lexical choice is done in a top-down recursive manner. From the lexicalized, skeletal sentence structure, a default ordering of words (and thus information units) is determined. Any remaining grammatical variation that remains is represented using partial order constraints (e.g. ordering elements within a list) and this avoids fully generating each candidate sentence before the media coordinator determines the compatible ordering.

After the surface structure of a sentence has been produced based on the compatible ordering, the sentence is sent to the speech synthesizer to determine durations. The duration, start and end time of each reference to an information unit are generated based on the time information of each phoneme produced by the speech synthesizer. This information will be used to synchronize speech with graphics actions.

4. PAUSE AND DURATION PREDICTION

If the duration of spoken references is too short, synchronized highlighting will occur too quickly in a blinking fashion. To avoid this, the duration of the corresponding spoken references must be increased. This can be done either by predicting where pauses can be added to the speech without comprising intelligibility and naturalness or by overlapping highlighting with spoken material between references.

For the purpose of predicting possible additional pause locations, we employ a variant of the algorithm proposed in [2]. Briefly, in their algorithm, a phonological phrase is defined as all the material up to and including the head of a syntactic phrase, following Gee and Grosjean [7]. Two kinds of rules are used to determine pause location and relative strength respectively. Location rules effectively derive phonological words and phrases. Boundary salience rules are used to group phonological phrases into prosodic phrases with pause strength index. Although, the accuracy of this algorithm is not totally satisfactory, we use this as the first step towards predicting pauses and then modify pause length based on constraints for determining a reasonable duration.

The speech generator produces a sentence tagged with both the information structure derived from the presentation plan[1] and pauses at the locations which are the primary phrase boundaries predicted by Bachenko's and Fitzpatrick's algorithm. If any predicted phonological phrase boundaries occur within a phrase referring to a basic information unit, its relative strength is diminished.

[1] We represent both basic information units and *intermediate information units*, which are a group of basic information units. Intermediate information units consist of more than one basic information unit. For example, "a medical history of anorexia and hypokalemia" is an intermediate information unit which includes two basic information units: anorexia and hypokalemia.

The speech generator provides several ways to adjust speech durations that can be used if there are conflicts between the speech and graphics constraints on duration. First, the speech generator computes length *ranges* for each possible pause at each primary phrase boundary. In this domain, anywhere in the range of $0.02 \times C$ to $0.08 \times C$ seconds per pause provides satisfactory results, where C is the ratio of the real speech rate to the default rate. Through experimentation, $0.05 \times C$ seconds was selected as the default pause length which is added to a sentence at the primary prosodic boundaries. As a result, the speech generator can adjust the duration of each spoken reference by modifying the duration of surrounding primary phrase boundaries by as much as $\pm 0.03 \times C$

Another way to adjust the time constraints is by changing the speaking rate, expressed through C above. Again through experimentation, we determined that C can vary from 0.5 to 1 without significantly affecting the speech quality.

Finally, we can also adjust duration by considering the non-referential words which occur in between spoken references to graphical objects. Non-referential words are those which do not correspond to any basic information units in the presentation plan, or which correspond to a unit which is not depicted graphically. For example, in the sentence *"(Ms. Walker) is a (50 year old) (female) patient of (doctor Longman) (undergoing CABG). She has (a medical history of (anorexia) and (hypokalemia))"*, the parentheses indicate the boundaries of the information units and the words which are not within any pair of parentheses are non-referential words.

The speech generator must in addition determine whether non-referential words are more naturally linked with the spoken reference that precedes or follows. For example, in the sentence above, "is a" are two non-referential words which are spoken after the reference to *name* and before the reference to *age*. Highlighting *name* and *age* should be synchronized with the references to *name* and *age* respectively. In synchronizing graphical highlighting with non-referential words, the graphics generator can turn off highlighting between references, can retain highlighting for the preceding reference while non-referential words are spoken (e.g., *name* remains highlighted while "is a" is spoken), can switch to highlight the graphical object referred to by the following reference (e.g., *age* is highlighted while "is a" is spoken), or can use some combination of these methods. In this example, turning off highlighting for a short time is undesirable from graphics' point of view.

Information about prosody and language structure is used to provide input for synchronizing the non-referential words. Here, "a" should be synchronized with highlighting of the spoken reference that follows since it is part of the following noun phrase. Based on the result of applying Bachenko's verb balancing rule for pause prediction[2] on "is", the position after "is" gets higher salience as a prosodic phrasal boundary. The speech generator uses this pause strength index to indicate that it is preferable to speak "is" during the highlighting of *name* (i.e., with the preceding reference) while "a" can be spoken during the highlighting of *age*.

Currently the speech generator provides the durations for highlighting actions to the media coordinator based on phrasal and syntactic boundaries as described. This situation is not ideal since the graphics generator also has constraints on duration of highlighting, which are not explicitly represented. We are currently investigating having the speech and graphics generators provide their preferences to the media coordinator for negotiation.

5. LANGUAGE GENERATION FOR SPEECH

We usually think of speech as more casual than written language. Through user studies with eventual end users of MAGIC, we identified specific ways in which references can be more casual in speech than in text. Our user studies also revealed that speech should be shorter than any text used in the accompanying illustration. Speech takes time to be heard but the user can read the accompanying text at her own pace while listening to speech. As a result, a major goal for the speech generator is to produce language that is as concise as possible. This is realized both through the generation of references and through the generation of sentence structure.

In generating references to objects that are displayed graphically, the speech generator can use a short, casual reference as long as the full, unambiguous reference is displayed using a textual label in the accompanying illustration. For example, the system generates the full-length term "Ventricular Pacemaker" for the textual label displayed in the illustration. But for speech, a shorter term "pacemaker" is used to refer to the same information. The speech generator achieves this by representing multiple forms of referring to the same concept in its lexicon and choosing one based on the media being used. The choice of spoken reference is also constrained by the textual reference that is generated; any modifiers that are included in text can be omitted in speech.

Another difference is that usually a well-organized sentence with complex structure (such as example (1) in Section 3) is perfectly acceptable in written language but hard to understand in speech. However, in order to generate concise speech (and as noted, in our user studies, our busy caregivers specifically requested concise speech), MAGIC uses the strategy of loading information into small phrases (e.g., modifiers) resulting in overall fewer words, but more complex sentences. We need to balance conciseness versus excessive complexity. Currently, we use the pause predicting techniques to

[2]The verb balancing rule places a prosodic phrasal boundary either before or after the verb depending on whether combining the verb with the phonological phrase to the left results in a phrase with more phonological words than that to the right of the verb.

add pauses at proper locations in complex sentences, so that such sentences are subdivided into balanced prosodic phrases of acceptable intelligibility.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a model of spoken language generation to support coordination required in multimedia systems. Our work provides methods for generating both different possible orderings of information units and variations of durations for speech which can be synchronized with graphical highlighting. We exploit accurate syntactic and semantic information available to us during sentence planning to compute prosodic features; these are used in turn to improve speech quality and adjust duration for the purpose of media synchronization. In addition, we integrated a large scale language generation system with speech synthesis, identifying new constraints on language production for speech.

We are extending our prototype system in several ways:

1. Identifying how a variety of prosodic features can be substantially altered because of the accompanying graphical presentation.
2. Modifying the content and wording of what is said based on the amount of time available to speak. This can involve either deleting information (or words) or adding in additional relevant information (or words) when coordination with graphics indicates additional time is available.
3. Incorporating additional constraints on syntactic and lexical choice that are appropriate for speech.

7. ACKNOWLEDGMENTS

MAGIC is a system involving a large number of people at Columbia University. In addition to the authors, who are responsible for the text and spoken language generator, the MAGIC team includes James Shaw (media-independent content planning and text organization); Steve Feiner, Michelle Zhou (graphics generation); Mukesh Dalal, Yong Feng (knowledge representation and temporal constraint satisfaction); and Jeannie Fromer, Tobias Hollerer (media coordination). We thank Vasilis Hatzivassiloglou for extensive comments on an earlier version of the paper. This research is supported in part by DARPA Contract DAAL01-94-K-0119, the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York State Science and Technology Foundation) and GER-90-2406.

8. REFERENCES

1. J. Allen, S. Hunnicutt, and D. Klatt. *From text to speech: the MITalk system.* Cambridge University Press, Cambridge, 1987.
2. J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170, 1990.
3. K. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Morristown, New Jersey, 1988.
4. M. Dalal, S. Feiner, K. McKewon, S. Pan, M. Zhou, T. Hollerer, J. Shaw, Y. Feng, and J. Fromer. Negotiation for automated generation of temporal multimedia presentations. Submitted, 1996.
5. J. Davis and J. Hirschberg. Assigning intonational features in synthesized spoken discourse. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, Buffalo, New York, 1988.
6. M. Elhadad. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation.* PhD thesis, Columbia University, 1993.
7. J. P. Gee and F. Grosjean. Performance structure: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15:411–458, 1983.
8. M. Liberman and A.L. Buchsbaum. Structure and usage of current Bell Labs Text-To-Speech programs. Technical Memorandum 11225-850731-11, AT&T Bell laboratories, 1985.
9. K. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press, Cambridge, England, 1985.
10. A. Monaghan. *Intonation in a Text-To-Speech conversion system.* PhD thesis, University of Edinburgh, 1991.
11. J. Neal, C. Thielman, Z. Dobes, S. Haller, and S. Shapiro. Natural language with integrated deictic and graphics gestures. In *Proceedings of Speech and Natural Language Workshop*, pages 410–423, Cape Cod, Massachusetts, 1989.
12. S. Prevost. *A Semantics of Contrast and Informaiton Structure for Specifying Intonation in Spoken Language Generation.* PhD thesis, University of Pennsylvania, 1995.
13. J. Shaw. Conciseness through aggregation in text generation. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics*, pages 329–331, Cambridge, Massachusetts, 1995.
14. S. Young and F. Fallside. Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustical Society of America*, 66:685–695, 1979.
15. M. Zhou and S. Feiner. Data characterization for automatically visualizing heterogeneous information. Submitted, 1996.