Modeling Prosody Automatically in Concept-to-Speech Generation

Shimei Pan
Department of Computer Science
Columbia University
1214 Amsterdam Ave, Mail code 0401
New York, NY 10027
pan@cs.columbia.edu

A Concept-to-Speech (CTS) Generator is a system which integrates language generation with speech synthesis and produces speech from semantic representations. This is in contrast to Text-to-Speech (TTS) systems where speech is produced from text. CTS systems have an advantage over TTS because of the availability of semantic and pragmatic information, which are considered crucial for prosody generation, a process which models the variations in pitch, tempo and rhythm. My goal is to build a CTS system which produces more natural and intelligible speech than TTS. The CTS system is being developed as part of MAGIC (Dalal *et al.* 1996), a multimedia presentation generation system for health-care domain.

My thesis emphasizes investigation and establishment of systematic methodologies for automatic prosody modeling using corpus analysis. Prosody modeling in most previous CTS systems employs handcrafted rules, with little evaluation of the overall performance of the rules. By systematically employing different machine learning techniques on a speech corpus, I am able to automatically model prosody for a given domain. Another focus of my thesis is on system architecture. There are two concerns when designing a CTS system: modularity and extensibility. The goal is to design a flexible CTS system so that new prosody generators, natural language generators and speech realization systems can be incorporated without requiring major changes to the existing system. Designing a CTS system to facilitate multimedia synchronization is another focus of this research.

I have conducted initial investigations on different prosody models using a speech corpus collected from a medical domain. Different machine learning techniques were explored. For example, a classification based rule induction system and a generalized linear model are used in identifying and combining salient prosody indicators. Hidden Markov Models are also used to automatically derive probability models to predict a sequence of prosodic features from a sequence of language features. Preliminary results (Pan and McKeown 1998) show that the output features of a general-purpose natural language generator, FUF/SERGE, are useful in improving the performance of prosody models. Recent results also indicate that the semantic informativeness of a word is an effective predictor of pitch accent assignment. In the future, I plan to investigate the effects of more discourse and semantic features, such as given/new, semantic focus, rhetorical relations, and build a more comprehensive prosody model. Both subjective and objective evaluations will be provided for the final comprehensive prosody models.

In order to design a flexible CTS architecture, I employ a SGML-based markup language as a standard interface between CTS components. The elements defined in the markup language are typical language and speech features. As a result, different prosody generators, natural language generators and speech synthesizers can be integrated in the CTS system through this interface.

In order to design a CTS system in a multimedia context, I have modified the sentence planner to produce different paraphrases to facilitate the coordination with spatial constraints from graphics. Similarly, in the lexical selection and prosody generation module, special considerations are incorporated to facilitate temporal coordination with other media.

Acknowledgments

This work has been advised by Kathleen McKeown and Julia Hirschberg and is supported by NSF under Grant NO. IRI 9528998 and the New York State Science and Technology Foundation under Grant No. NYSSTF CAT 97013 SC1.

References

M. Dalal, S. Feiner, K. McKeown, S. Pan, M. Zhou, T. Höellerer, J. Shaw, Y. Feng, and J. Fromer. Negotiation for automated generation of temporal multimedia presentations. In *Proceedings of ACM Multimedia'96*, pages 55–64, 1996.

S. Pan and K. McKeown. Learning intonation rules for Concept-to-Speech generation. In *Proceedings of COLING/ACL'98*, Montreal, Canada, 1998.