Lecture 36: April 19, 1995

Text-To-Speech Conversion Technology

Lecturer: Michael H. O'Malley

Scribe: sangjin@eecs.berkeley.edu

1. What is TTS (Text-To-Speech)?

TTS conversion transforms linguistic information stored as data or text into speech. It is widely used in audio devices for blind people. It converts electronic mail to voice mail for audio access by phone. It also permit field personnel to access large text databases, like part inventories, by telephone.

Most TTS systems consist of two major elements, a sound generation mechanism and a mechanism converting text into speech parameters.

2. Text to speech Mechanism

The process of converting text into speech parameters break down into a number of stages. They are text normalization, word pronunciation, prosodies, phonetic rules, and voice tables.

2.1 Text Normalization

The text normalizer converts everything in the text stream to letters. Actual texts have a great deal of symbolic material such as numbers, abbreviations, acronyms, and information signaled by graphic layout. For example, "1990" becomes "nineteen ninety" and "Dr." becomes either "doctor" or "drive", as appropriate.

2.2 Word Pronunciation

This module combines a dictionary with sophisticated pronunciation rules. The dictionary handles words with syntactic or context-dependent effect, as well as words not regular enough to justify a rule. Some words in English are not pronunciated in accordance with the basic rules for English pronunciation. The word "of," for example, is the only case where a final "f" is pronounced "v". To say these words correctly, the system stores a phonemic transcription of their exact pronunciation in an exception dictionary. The dictionary is also needed to store grammatical information about particular words for use by other modules in the system.

2.3 Prosodies

The prosodic component of speech involves such phenomena as rhythm, intonation, and the emphasis or de-emphasis of particular words. For any given text, a human speaker might produce thousands of possible prosodic interpretations, but an even larger number would not be considered natural under reasonable circumstances.

2.4 Phonetic Rules

The phonetic rule module takes phonemes as input to produce a detailed description of the sounds of the utterance. The rules in this module do things such as assign a duration to each phoneme according to context, for example, lengthening the final vowel in a phrase by up to 50 percent. Other rules remove parts of phonemes, for example, deleting the aspiration of a stop as /k/ in a number of different contexts. Phonetic rules play the major role of describing the co-articulation between phonemes. Each phoneme strongly influences the parameters in the adjacent phonemes. In some cases, this influence might extended over several adjacent phonemes.

2.5 Voice Tables

The voice table modules convert phonetic description provided by the phonetic rules into numeric targets for use by the voice model. The phonetic rules in conjunction with the voice tables are the primary determinants of voice intelligibility. One major function of the voice tables is to handle differences in bandwidth. Speech researchers have usually assumed a minimum speech bandwidth of 5 to 6 kHz. But telephone system designers have traditionally employed a bandwidth of 3.5 kHz. Different sets of voice tables are used for different frequency ranges. For example of applications involving telephones, some of the frequency components above 4 kHz can be mapped to a lower frequency, potentially contributing to overall speech quality and intelligibility.

3. Sound Generation Device

In order to produce a sound, some type of sound-generation mechanism whose function is analogous to that of the human vocal tract is needed.

The voice generation module turns phonemes into many smaller speech units, which are then converted to voice parameters. Then the interrupt driver takes frames and sends them to the output hardware at regular intervals of from 5 to 25 milliseconds. Finally various kinds of output hardware are used to generate sounds.

4. Reference

Michael H. O'Malley, "Text-To-Speech Conversion Technology", IEEE Computer, pp. 17-23, August 1990.