Chapter 11

Text to Speech System

Text to Speech (TTS) systems vary widely in performance and usability. Essentially all use speech recorded by "speech talents". This recorded speech is then used to piece together sound from text. How this piecing together is done distinguishes one technology from another.

There are two technologies primarily in use, Unit Selection and Hidden Markov Models (HMM's).

Unit Selection has whole words and phrases stored in memory. These are then selected to assemble the most natural sounding speech from text. The prosody can be excellent, but at the cost of a large memory footprint. One of the better sounding voices in Festival uses ~150 MB of storage.

The system used in the CPA is "Flite+HTS". http://www.sp.nitech.ac.jp/demo/flite+hts_engine/. It balances memory size against performance, it is tunable, and is open source. And, it can have very good prosody as evidenced by the examples on their website.

Before delving into the details of the TTS employed in the CPA, it is instructive to examine a very simple, but flexible, TTS system. This system uses the Linear Predictive Coding method to construct word sounds for words that are found in a pronunciation dictionary. This code is based on a brief tutorial provided by Casey Chesnut at http://www.generation5.org/content/2004/ttSpeech.asp.

Figure II-1.1 outlines this process. First, a sentence is separated into its individual words. The pronunciation (sequence of phones) of each of these words is found in the pronunciation dictionary. The dictionary used here is is the Carnegie-Mellon University (CMU) cmudict.0.7a. This is a plain text file that may be read using most word processors. This is the entry for the word "example":

EXAMPLE IH0 G Z AE1 M P AH0 L

Next, diphones are constructed. The reason for this step is that developers of TTS systems found that working with diphones produces smoother speech than phones.

The phone digit suffixes are ignored in this step. A pause "pau" is inserted at the beginning and end of each word.

pau-ih ih-g g-z z-ae ae-m m-p p-ah ah-l l-pau

The diphones are then looked up in the LPC table. Here we use CMU cmu_us_kal16.txt (16 bit) or cmu_us_kal.txt (8 bit). These tables are derived from a recordings of a male voice "Kevin" (Prof. Kevin A. Lenzo?, a colleague of Prof Allen Black, at Carnegie-Mellon University, both world leading experts in synthetic speech).

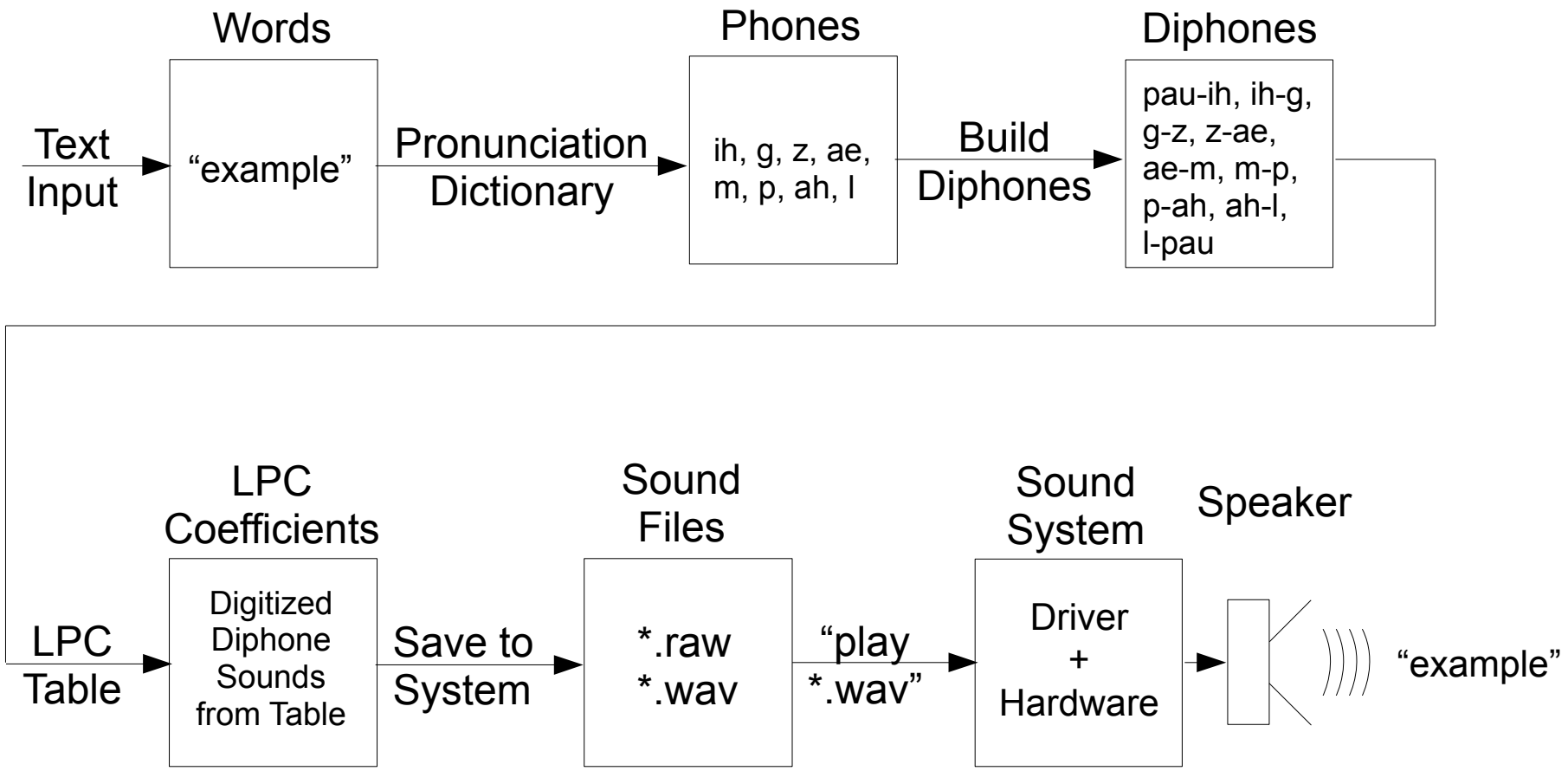
Linear Predictive Coding is a speech compression technique. The Wikipedia entry for LPC is

http://en.wikipedia.org/wiki/Linear_predictive_coding. It is definitely worth the read.

The LPC TTS program on the CD-ROM is located in the SyntheticSpeech directory

The TTS system employed in the CPA is Flite+HTS. As the name suggests, this is a combination of Flite (Festival Lite) and HMM TTS synthesis developed by Nagoya University. A pdf slide presentation is available at http://www.sp.nitech.ac.jp/~tokuda/tokuda_iscslp2006.pdf. Prof Alan Black of CMU wrote the code for Flite+HTS. If HTS is used with University of Edinburgh's HTK, then the licensing restriction of HTK must be followed. Flite+HTS is more liberal in its allowed usage.

The code used in the CPA has some minor modifications made by the author. Additionally, the top level source code was reformatted to improve readability.



Speech Synthesis using Linear Predictive Coding

Figure II-1.1