MAS.632 Conversational Computer Systems
Fall 2008

Problem Set 6

1. Synthetic vs. recorded speech

   Digitized speech can be sped up on playback, and synthesized speech can be spoken faster as well. The SOLA algorithm removes segments approximating pitch periods during playback, to minimize distortion and maintain a constant pitch.

   Speech synthesizers also generally support speed control. Of course one could take the synthesized audio waveform and run the SOLA algorithm on it. Is there a smarter way to increase the speed of synthesized speech? Think about how the synthesizer generates speech.

   Considering the differences in how they produce speech, for which (of digitized and synthesized speech) would you expect the greatest ADDITIONAL DECREASE of intelligibility as you increase playback speed? How would you measure this decrease in intelligibility?

2. Speech Recognition

   a) Why are LPC coefficients a popular representation of speech for recognition purposes?

   b) Why might it be useful to build a template by having the user train it by speaking the word to be recognized multiple times?

   c) In order to achieve speaker independent recognition, we could imagine getting several dozen subjects, having them each train a template, and then averaging all these templates together to obtain a representative speaker independent template. Is this a good technique? How can we improve it? How should we "average"?

   d) Why is keyword spotting more difficult than connected speech recognition? Compare the difficulty of detecting the telephone number a caller leaves in my voice mailbox versus trying to recognize a number spoken into a speech recognition telephone on my desk. (Ignore issues of channel characteristics, i.e., that the telephone line may have been noisy and have reduced bandwidth compared to the signal reaching the handset in my hand.)