Modeling intonation in Text-to-Speech synthesis with phrasal prosodic features
Project Proposal
CS224S/LINGUIST285

Kevin Garbe (*kgarbe*), Aleksander Główka (*aglowka*) & Leon Lin (*leonl*)

May 1, 2017

1 Task

Intonation, the pitch pattern of a sentence, conveys linguistic information beyond that conveyed by words. In this project we aim to model phrasal intonation in Text-to-Speech synthesis using prosodic features based on a combination of phrase grouping, pitch, and duration. A TTS system such as CMU Flite Synthesizer transcribes each word separately without controlling the overall pitch contour, which makes the utterance sound unnaturally flat *[click for example]*. However, in natural speech pitch marks phrasal boundaries and signals whether the utterance is a declarative or a question. In (1), for example, the pitch not only declines throughout the utterance marking it as a declarative, but also shows intermediate rises and declinations marking intermediate phrases as illustrated in (1).

(1) [$_\phi$ I have often observed] [$_\phi$ that in married households]
[$_\phi$ the champagne is rarely of a first-rate brand].

Our goal is to implement phrasal prosodic features that combine information about phrase groupings with pitch and duration in order to produce more naturalistic TTS synthesis.

2 Dataset

The Festival Speech Synthesis System was originally developed at the University of Edinburgh and currently has a free software license. Festvox, and in particular the Flite Synthesis Engine, is a project from Carnegie Mellon University built on top of Festival. We will use the tools and modules from Festival and Festvox in implementing our system. In particular, we will use the TTS readings from Flite as the baseline of comparison. The Blizzard Challenge provides 6.5 hours of read speech in British English from children's books. This speech should have a more limited vocabulary and, as it is read rather than conversational, could have greater emphasis than is natural.

3 Approach

We will begin by comparing recorded human read speech from the Blizzard Corpus to Flite TTS of the same script. We will analyze the differences in the intonation and pauses that mark subsections of sentences in the read speech relative to Flite's TTS and attempt to draw relevant features that show those differences. This will serve as the error metric by which we evaluate a larger amount of TTS speech. Next, we will attempt to find prosodic features that can be both derived from pure text and that correspond to the features we found. Using the tools from Festival and Festvox, we will modify the speech intonations and pauses of the TTS systematically with supervised learning to match that of the recorded speech as much as possible.

4 Evaluation

Since our domain of interest is above the word level, we cannot use Word Error Rate as an evaluation metric. We will conduct an error analysis of Festvox versus human speech on a small text sample (1 page of continuous text). We will select the most common intonational synthesizer errors and develop a weighted scoring system for them (e.g. how many phrasal groupings are there in the utterance and how many did the system fail to mark intonationally?). After we implement the features in Flite and train them on an annotated sample of the Blizzard corpus, we will test our system on unseen text and score it compared to human speech in Blizzard.