Analysis of Pitch Contours in Repetition-Disfluency using Stem-ML

Rajiv M. Reddy
Department of Electrical Engineeing
University of Illinois
Urbana, 61801, IL, USA
rreddy@uiuc.edu

Mark A. Hasegawa-Johnson
Faculty of Electrical Engineering
University of Illinois
Urbana, 61801, IL, USA
jhasegaw@uiuc.edu

Abstract

F0 analysis-by-synthesis methods are used in order to test the hypothesis that the pitch contour in the alteration segment of disfluency tends to mimic the pitch contour in the reparandum segment of that disfluency. Reparandum-Alteration pairs selected by transcribers as having perceptually similar F0 contours were compared to arbitrarily selected fluent word-pair sequences using Stem-ML. All word-pair sequences had similar pitch; disfluent pairs were not more similar than others.

1 Acknowledgments

This research was supported by REU funding under NSF grant IIS 04-14117. Conclusions are those of the authors, and are not endorsed by the NSF. Special thanks go to Chilin Shih for her valuable comments.

2 Introduction

Soft Template Mark up Language (Stem-ML) is a tagging system that is used to describe intonation and prosody in human speech. These tags are used in automated training of accents shapes and parameters from acoustic databases (Kochanski and Shih, 2000). Stem-ML is used to synthesize pitch contours of disfluent speech in this experiment.

Cole et al. [DISS 2005] proposed that "the frequent occurrence of parallel prosodic features in the reparandum (REP) and alteration (ALT) intervals of complex disfluencies may serve as strong perceptual cues that signal the disfluency to the listener." The goal of this research is to test whether prosodic features in the REP and ALT (specifically, F0) resemble one another. The preliminary impression from looking at the data is that the REP and ALT seem similar; if so, this similarity might be used to detect disfluencies.

3 Stem-ML parameters

Certain features of Stem-ML described below are the most relevant to understanding how the hypothesis was tested.

Stress Tags: The *stress* tag specifies the local F0 contour of a period of time normally corresponding to a syllable or word (Kochanski and Shih, 2000). In this case it corresponds to REP or ALT segments of a disfluency. The *stress* tag is defined by attributes like type, atype, strength and the number of points that are trained.

The pitch target y consists of the phrase component added to the stress tag.

(1)

where P is the phrase curve, Y is the shape of the stress tag (specified by interpolating a small finite number of points), and $atype * s^{|atype|}$ is a scale factor for the tag's pitch range where *s* stands for the strength. (Kochanski and Shih, 2003).

Strength: *Strength* controls the interaction of accent tags with their neighbors. If the strength tag is low the smoothness of the synthesized pitch is more important than the accuracy (Kochanski and Shih, 2003).

Base and Range: The base and range are speaker dependant constants. To reduce the number of parameters Stem-ML needs to learn, the base and range are calculated outside of Stem-ML. The base is estimated as the mean value of the F0

in each file and the range is estimated as the difference between the 25th and 75th percentile of the F0 in that file.

4 Data

The database used for these experiments is a subset of Swtichboard. It is the same data set that was transcribed for [Cole et al., DISS'05]. The data contain 71 two minute blocks of data with added transcription tiers including disfluency type (repetition, repair, …), disfluency segment (REP, EDIT, ALT), and perceived relationship between REP and ALT pitch contours (same, stress, phrase boundary, …). Tokens marked repetition-same-disfluency cases were extracted; these are repetition disfluencies in which the REP and ALT F0 contours were perceived by the transcriber as sounding the same. The REP and ALT segment markings bound the domain of *stress* tags in Stem-ML. For comparison, fluent word pairs were extracted: a fluent word pair contains any two words uttered sequentially during normal fluent speech.

5 The Experiment

To test the hypothesis that REP mimicked ALT, Stem-ML models with tags are created to represent the disfluent speech. Stem-ML is forced to learn the same *stress* tag (pitch contour) for the reparandum and alteration. If REP mimics ALT, we should get a lower RMS pitch error value per sample in disfluent word pairs as compared to fluent word pairs.

The *strength* of the *stress* tags are varied to see the effect of changing the strength on the RMS of the pitch error per sample. We set the *stress* tag to learn 3 points i.e. the 25th, 50th and 75th percentile for each placement on the pitch curve.

The model is used to learn the pitch contour of each REP/ALT pair, and the RMS error per sample for each disfluent pair is calculated. To compare these values with fluent speech we run the same model on randomly selected consecutive words of fluent speech. The results are shown in Table 1.

6 Results

	Fluency	**Disfluency**
Mean	11.47 Hz/Sample	18.29 Hz/Sample
Median	7.91 Hz/Sample	15.62 Hz/Sample
StdDev	9.13 Hz/Sample	14.24 Hz/Sample

Table 1. RMS pitch error for fluent speech cases and disfluent speech cases after training with strength = 8.

Contrary to the hypothesis, a lower average RMS pitch error per sample is found in the fluent word pairs than in the disfluent pairs. This goes against the hypothesis that the F0 contour of reparandum mimics that of alteration. Rather, it seems that any two consecutive words have similar pitch contours. Notice that, when using one fluent word to predict the next word's F0, we incur an RMS error of only 11.47Hz. One possible conclusion is that the Switchboard database is primarily monotone: 31 out of 71 files have an F0 standard deviation that is less than 16% of the mean value. Thus the lower average RMS pitch error per sample for fluency cases may be due to the fact that a large part of the database is spoken in monotone; disfluency does not reduce the difference between successive words because all word pairs have similarly flat F0 contours. Hence, there were no cues detected by Stem-ML that could be used to differentiate between repetition-same-disfluency and fluent speech.

7 References

Chilin Shih, Greg Kochanski, 2003. Modeling of Vocal Styles Using Portable Features and Placement Rules, *International Journal of Speech Technology*, 6(4):393-408.

Chilin Shih, Greg Kochanski, 2003. Prosody modeling with soft templates. *Speech Communication* 39(3-4):311-352.

Chilin Shih, Greg Kochanski, 2000. Stem-ML: Language independent prosody description. ICSLP, 3:239-242. Beijing, China.

Jennifer Cole, Mark Hasegawa-Johnson, Chilin Shih, Heejin Kim, Eun-Kyung Lee, Hsin-yi Lu, Yoonsook Mo, Tae-Jin Yoon, 2005. Prosodic parallelism as a cue to repetition and error correction disfluency, *DiSS*, 53-58.