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

Rajiv M. Reddy
Department of Electrical Engineering
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 Introduction

Spontaneous speech contains high rates of disfluencies like repetitions, repairs, filled pauses, etc. It is estimated that about 10% of all spontaneous utterances contain disfluencies (Nakatani and Hirschberg, 1994). Disfluency can be characterized by a three-region surface structure illustrated in figure 1 (Shriberg, 2001).

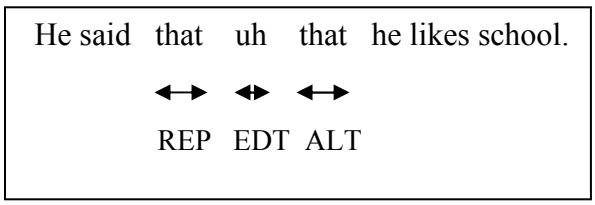

Figure 1. Illustration of the three-region structure of a "repetition-same-disfluency"

The three regions are the Reparandum (REP), the Edit (EDT) and the Alteration (ALT). The first region of disfluency, the REP, is the segment that is being replaced. At the end of the REP there is an interruption point where the speaker realizes that there was an error in the REP and decides to repair it. The next phase is the EDT phase which contains the region between the interruption point and the onset of repair. This region may be a silent pause or a filled pause (um, uh, I mean). The last region is the ALT which represents the repair of the REP and the resumption of fluent speech (Shriberg, 2001).

Disfluencies present a challenge for automatic speech recognition since they are often unmodeled in speech recognition systems (Nakatani and Hirschberg, 1994). A model that could detect disfluencies would decrease the errors in speech recognition systems. In this experiment, "repetition-same-disfluencies" are modeled using Stem-ML (Soft Template Mark up Language). In "repetition-same-disfluencies," the transcriber perceived the REP and the ALT to be the same.

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. (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.

2 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, 2003). In our system, the REP and ALT of each disfluency are assumed to be instantiations of the same stress tag. Different utterances use different stress tags. The *stress* tag is defined by a number of parameters. The most important to this experiment are the *strength* and the number of points that are trained. The parameters that are irrelevant to the experiment were set to 0 and have hence been taken out of the original Stem-ML equations (Kochanski and Shih, 2003).
The final synthesized f0 is given by

$$f_0(t) = p(t) * range + base \tag{1}$$

where *p(t)* is the normalized F0 relative to the range of the speaker (Kochanski et al., 2003). *p(t)* is a compromise between the articulatory effort (G) and the weighted error (R).

$$p(t) = \arg\min_{p(t)} \left(G + R\right) \tag{2}$$

The articulatory effort (G) is represented by

$$G = \sum_t \left(\dot{e}_t^2 + \tau^2 \ddot{e}_t^2\right) \tag{3}$$

where τ is a constant and the dots represent derivatives. The weighted error (R) is the sum of the errors in each tag weighted by the tag *strength* s_k (Shih and Kochanski, 2003).

$$R = \sum_{k \in \text{tags}} s_k^2 r_k \tag{4}$$

and r is the error for each template.

$$r_k = \sum_{t \in \text{tag k}} \frac{\cos\left(type.\pi/2\right)\left(\left(e_t - \overline{e}_t\right) - \left(y_{k,t} - \overline{y}_k\right)\right)^2 + \sin\left(type.\pi/2\right)\left(\overline{e}_k - \overline{y}_k\right)^2}{} \tag{5}$$

where type is set to 0 in all the experiments.

$$\Rightarrow r_k = \sum_{t \in \text{tag k}} \left(\left(e_t - \overline{e}_t\right) - \left(y_{k,t} - \overline{y}_k\right)\right)^2 \tag{6}$$

where e_k is the normalized F0 contour, $\overline{e}_k$ is the average normalized F0 contour of that tag, y_k is the normalized target F0 contour and $\overline{y}_k$ is the average normalized target F0 contour for that *stress* tag (Kochanski et al., 2005).

Strength

Strength controls the interaction of accent tags with their neighbors. If the strength tag is low the smoothness of the synthesized F0 contour is more important than the accuracy (Shih and Kochanski, 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 25th percentile of the F0 in each file and the range is estimated as the difference between the 25th and 75th percentile of the F0 contour in that file.

Word Scale

This parameter is represented by the variable *wscale*. It is the ratio of the length of the segment on which the tag is being placed to the length of the original F0 contour. This allows us to place tags on two segments of different lengths since it stretches or compresses the stress tag according to the length of the segment i.e. REP/ALT. The *wscale* of the alteration stress tag is set to 1 and the wscale ratio for the REP is set to duration (REP) / duration (ALT).

Step to tags

The *Step to* tag forces the phrase curve to have a certain frequency at the tag's position. It is specified as a fraction of the speaker's range (Kochanski and Shih, 2003).

Center Shift

This parameter allows the stress tag to be shifted within the ALT or REP to give a better fit. If it is

set to 0, Stem-ML is forced to follow the start and end given by the transcriber. If it is allowed to be used as a parameter that is learnt by Stem-Ml, it moves the stress tag around to minimize the error in the fit.

3 Data

The database used for these experiments is a subset of Swtichboard. It is the same data set that was transcribed for the Cole et al. (2005) paper. The data contain 71 two minute blocks of speech 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" 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.

4 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.

Parameters

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. *Ctrshift* was set to 0 or learnt by Stem-ML.

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.

Implementation

The experiment was implemented in Python. The data came from 2 sources. The first source was the .wav files which gave the f0's and the other source was .TextGrid files which gave the word boundaries, disfluency markings and disfluency types. The data from the .TextGrid files were extracted by using a Praat script that was generated by the Python script for each file. This was stored in variables in the Python script. The f0 was obtained from the .wav files by running the *get_f0* script on them. With this information model files were created for each individual case from the Python script and saved as .pf files. The Python script then ran *resid_for_opt.py* which is the Stem-Ml python script to learn the parameters for the disfluency case. After the script terminates the fitted f0 is plotted. If the machine learning script *resid_for_opt.py* terminated improperly due to errors, nothing will be plotted. If this happens than the file name and the disfluency case number is saved onto the log file and it is stated that the file crashed. If the learning script goes to completion and terminates properly we will get a proper plot. Then the file name and disfluency case number is saved onto the log file along with the calculated E (Error energy as calculated by Stem-ML) value and the MSE (Mean Square Error) value. E is defined by

$$E = \sum_{t \in disfluency} \left(\frac{e_t - y_t}{\sigma} \right)^2 - \log\left(\sigma^2\right) * dof \tag{7}$$

where e_t is the synthesized F0 in erbs, y_t is the target F0 in erbs, σ is the standard deviation and *dof* is the degrees of freedom. The E values are normalized by the degrees of freedom to compare results from different models (Kochanski, 2006).

Numerous experiments were conducted and the E values were gathered. The experiments included varying the strength values, setting the ctrshift to a particular value, learning 4 points instead of 3, etc. It was found in all the experiments that fluency produced better results than disfluency which was contrary to the hypothesis. Towards the end of the year, the MSE values were calculated as well for some of those experiments to find that the MSE values showed an even larger deviation from the hypothesis. Since it is still unclear if the E values are accurate measures to differentiate between fluency and disfluency only the MSE values have been reported here

5 Results

The mean, median and standard deviation of the RMS pitch errors for fluency and disfluency cases are shown in Table 1 and Table 2 for two experiments with different parameter values.

	Fluency Errors		Disfluency Errors	
	RMS (Hz)	**Norm. E (Erbs)**	**RMS (Hz)**	**Norm. E (Erbs)**
Mean	11.47	0.37	18.29	0.45
Median	7.91	0.13	15.62	0.47
StdDev	9.13	0.60	14.24	0.37

Table 1: RMS pitch error for fluent speech cases and disfluent speech cases after training with strength of the stress tag= 8. The *step_to* tags are forced to jump to the same frequency at the beginning of each tag.

	Fluency Errors		Disfluency Errors	
	RMS (Hz)	**Norm. E (Erbs)**	**RMS (Hz)**	**Norm. E (Erbs)**
Mean	11.53	0.38	18.02	0.77
Median	8.21	0.13	13.71	0.39
StdDev	8.70	0.60	13.41	0.96

Table 2: RMS pitch error for fluent speech cases and disfluent speech cases after training with strength of the stress tag= 8. The *step_to* tags of REP and ALT are allowed to jump to different frequencies that are learnt during machine learning and minimization of the error**.**

Experiment	**Fluency Errors** (Erbs/sample)		
	Mean	**Median**	**StdDev**
(a)	0.1625	0.1177	0.1141
(b)	0.5343	0.1983	0.7562
(c)	0.3679	0.2201	0.3528
(d)	0.2429	0.0920	0.2729
	Disfluency Errors (Erbs/sample)		
	Mean	**Median**	**StdDev**
(a)	0.3555	0.2283	0.3419
(b)	0. 4006	0.2581	0.4499
(c)	0.3692	0.2396	0.4058
(d)	0.3973	0.2870	0.4318

Table 3: Normalized E values for fluent speech cases and disfluent speech cases for a few different experiments:
(a) The EDT region is forced to be learned with the same strength(=8) as the REP and ALT stress tags. Also the wscale parameter is set by the user according to the length of the ALT and REP.
(b) The EDT region is forced to 0 irrespective of whether it is a voiced or silent EDT. The wscale parameter is learnt by stem-ML
(c) The EDT region is learnt with strength 1 while the stress tags are at strength 8, the wscale parameter is learnt by Stem-ML
(d) Same as experiment (b) except that 4 points are learnt instead of 3

Contrary to the hypothesis, a lower average RMS pitch error per sample is found in the fluent word pairs than in the disfluent pairs. There is a high standard deviation, which means that the error rates are not concentrated at a particular range, and are instead distributed more uniformly. In the few cases that higher means are found fluency, it is due to a few extreme values in the results. Figures 2 and 3 contain some examples of the fitting for fluency and disfluency cases.

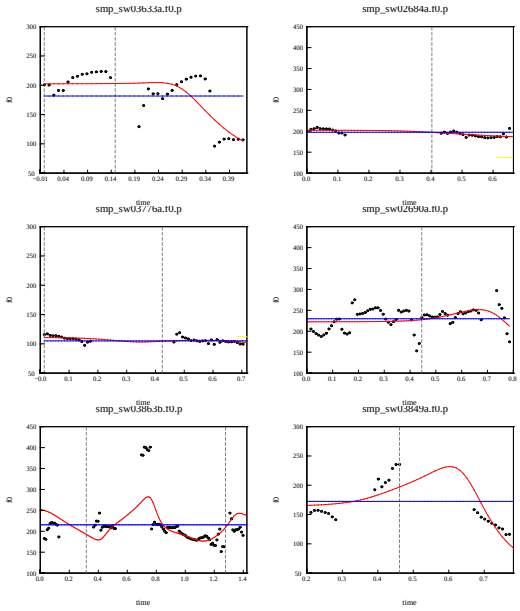

Figure 2. These are some plots of disfluent speech cases from different experiments. The big black dots represent the original F0, the flat line represents the baseline F0 and the curved line represents the learnt F0.

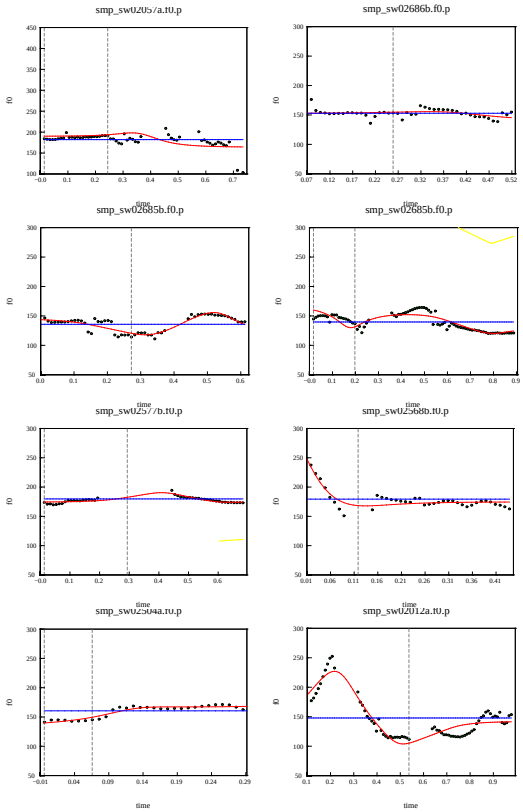

Figure 3. These are some plots of fluent speech cases from different experiments. The big black dots represent the original F0, the flat line represents the baseline F0 and the curved line represents the learnt F0.

6 Discussion and Conclusion

We found fluent pairs to have a lower pitch error than disfluent cases even though the model was constructed to force the REP and ALT to mimic each other. We cannot differentiate between fluency and disfluency by RMS pitch error. It is not possible to demonstrate, experimentally, that the F0 contour of reparandum mimics that of alteration. Rather, it seems that any two consecutive words have similar pitch contours since 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. In order to explore the hypothesis that switchboard is monotone, I calculated F0 standard deviation as a percentage of F0 mean in each utterance file. 31 out of 71 files have an F0 standard deviation that is less than 16% of the mean value. The histogram is shown in figure 4.

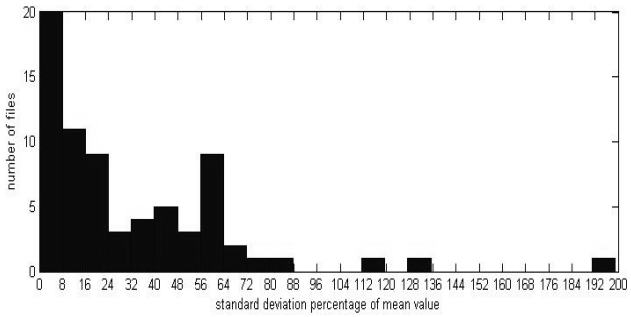

Figure 4. Histogram of the standard deviation percentage of the mean value in the switchboard file.

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.

In conclusion, there were no cues detected by Stem-ML that could be used to differentiate between repetition-same-disfluency and fluent speech.

7 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 and Greg Kochanski for their valuable comments.

8 References

C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech." *Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1603-1616. 1994.

C. Shih, G. Kochanski, "Modeling of Vocal Styles Using Portable Features and Placement Rules," *International Journal of Speech Technology*, vol. 6, no.4, pp. 393-408. October 2003.

G. Kochanski, C. Shih and H. Jing, "Quantitative measurement of prosodic strength in Mandarin", *Speech Communication*, vol. 41, no. 4, pp. 625-645, November 2003.

G. Kochanski, C. Shih and H. Jing, Erratum to "Quantitative measurement of prosodic strength in Mandarin", *Speech Communication*, vol. 47, no. 3, pp. 394, November 2005.

G. Kochanski, C. Shih, "Prosody modeling with soft templates", *Speech Communication,* vol. 39, no. 3-4, pp.311-352, February 2003.

G. Kochanski, C. Shih, "Stem-ML: Language independent prosody description", *ICSLP*, vol. 3, pp. 239-242. Beijing, China. 2000

G. Kochanski, personal communication, March 2006

J. Cole, M. Hasegawa-Johnson, C. Shih, H. Kim, E. Lee, H. Lu, Y. Mo, T. Yoon, "Prosodic parallelism as a cue to repetition and error correction disfluency", *DiSS*, pp. 53-58. 2005

E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the I.P.A.*, vol. 31, no. 1, pp. 153-169. 2001