Technique for mapping fast speech

Jesse Lawrence and Lucia da Silva

Communication Dynamics Laboratory, University of British Columbia

Introduction

What

The goal of this research is to synthesize intelligible fast rate speech from recordings of slower speech. We are particularly interested in using prosody (F0, amplitude, duration) to generate speaking rates 3, 4, and even 5 times normal rate.

Why

Linear compression is the simplest method for producing fast rate speech from normal rate recordings. However, unfamiliar fast rate speech generated by linear compression becomes unintelligible at compression rates higher than about 2:1 ([1], [2]). In order to synthesize fast speech that will be intelligible to listeners unfamiliar with the speaker, a techquique is needed that incorporates the non-linear relation between naturally produced normal and fast speech.

How

We will model the transform from natural normal rate speech to natural fast rate speech. We have been investigating various platforms to build upon, in particular TANDEM-STRAIGHT (TS) [3], an elegant speech analysis, manipulation, and synthesis system. TS does fine-grained manipulation of F0, amplitude, and the temporal structure of speech signals. However, accessing these parameters is problematic. Therefore data-driven paramerization of TS is an important pre-condition for the modeling we hope to accomplish.

Questions

Initially, we address 3 questions qualitatively:

1. How does the prosody of natural fast speech differ from natural normal-rate speech?
2. How does the prosody of TS generated fast speech differ from natural fast speech?
3. Can we apply the prosody of natural fast speech to TS generated fast speech?

Prosody in normal and fast rate speech

“If we ever had a slinky it tended to get tangled up or snap in half.”

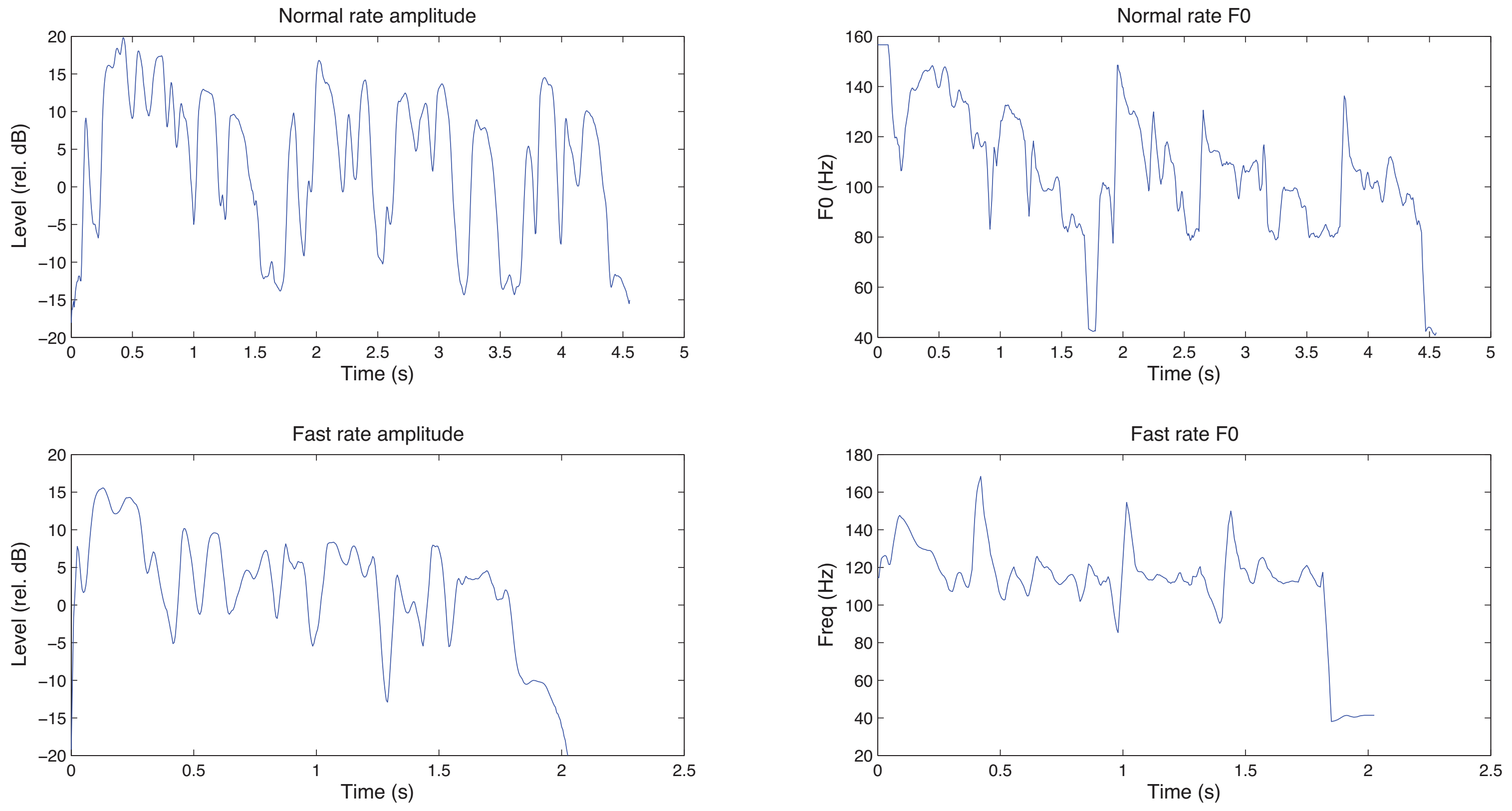


Figure 1: Comparison of the F0 and amplitude contours from natural normal rate speech and natural fast speech.

Observations

1. Amplitude of the fast speech is lower than normal rate speech.
2. The amplitude envelope changes shape at fast rate: After an initial peak, fast speech amplitude decreases and has less dynamic range than normal rate speech.
3. The pauses present in normal rate speech disappear as rate increases.

Prosody in fast speech and TS generated fast speech

Using TS, fast-rate sentences are generated from normal rate sentences by means of a pre-determined ratio between the measured durations for normal and fast rate sentences.

“If we ever had a slinky it tended to get tangled up or snap in half.”

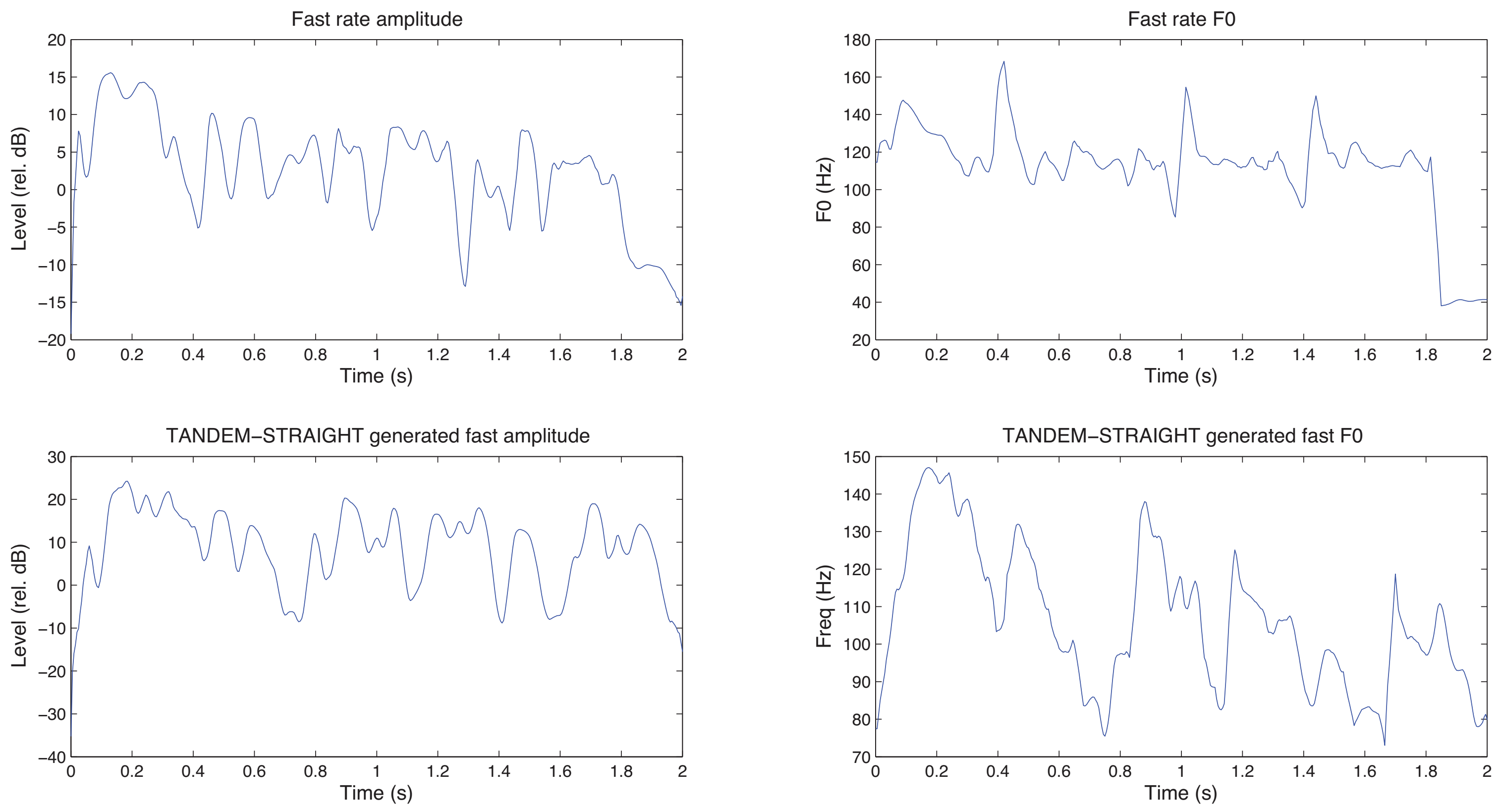


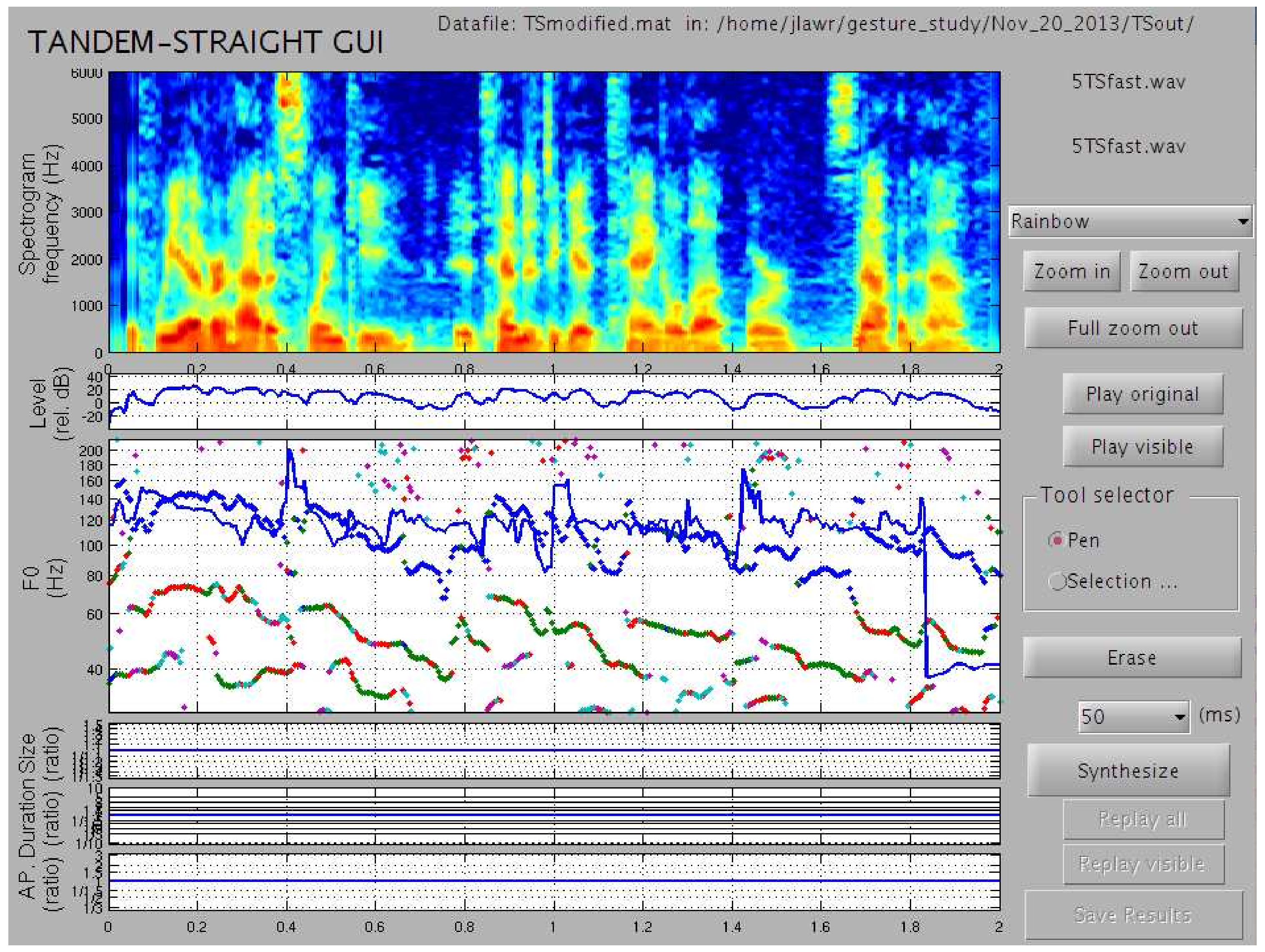
Figure 2: Comparison of the F0 and amplitude contours from natural fast speech and fast speech generated from normal rate speech by TS.

Observations

1. For both F0 and amplitude, values of the TS generated fast speech and the normal rate speech are quite similar (Figure 1).
2. The pause regions present in the normal rate speech are still present in the TS generated fast speech (Figure 1). TS can remove pauses with an extra processing step.

TANDEM-STRAIGHT generated fast speech with natural fast speech prosody

Here we show the results of integrating the F0 value of the natural fast rate speech into the TS generated fast speech.


TANDEM–STRAIGHT GUI

Datafile: TSmodified.mat in: /home/jlawr/gesture_study/Nov_20_2013/TSout/

5TSfast.wav

5TSfast.wav

Figure 3: Integration of F0 and amplitude contours from naturally produced fast speech into TS generated fast speech.

Integrating the F0 of natural fast speech into the TS generated fast speech improves the naturalness of the output. However, we have not been able to suppress the F0 candidate provided by TS which adds noise to the output.

Next steps

- Quantitative comparison of F0 and amplitude for normal rate and fast rate speech, and for fast rate and TS generated fast rate speech.
- When integrating natural F0 and amplitude into TS generated fast speech, eliminate the fast speech F0 candidate provided by TS.
- Perceptual evaluation of the intelligibility of TS fast speech using natural fast speech F0 and amplitude.

References

References

[1] N. Guttman G. Fairbanks and M.S. Miron. Auditory comprehension in relation to listening rate and selective verbal redundancy. *Journal of Speech and Hearing Disorders*, 22:23–32, 1957.

[2] Lucia da Silva, Adriano V. Barbosa, and Eric Vatikiotis-Bateson. Comprehending speech at artificially enhanced rates. In *Proceedings of Meetings on Acoustics*, Montreal, June 2013. Acoustical Society of America.

[3] Hideki Kawahara. Development of exploratory research tools based on TANDEM-STRAIGHT. *APSIPA ASC 2009*, 2009. URL `http://eprints.lib.hokudai.ac.jp/dspace/handle/2115/39651`.