Prosodic Similarity – Evidence from an Imitation Study

Hansjörg Mixdorff[1], *Jennifer Cole*[2], *Stefanie Shattuck-Hufnagel*[3]

[1] Department of Computer Science and Media, Beuth University of Applied Sciences, Berlin
[2] Department of Linguistics, University of Illinois at Urbana-Champaign
[3] Speech Communication Group, Research Laboratory of Electronics, MIT
mixdorff@beuth-hochschule.de, jscole@illinois.edu, sshuf@mit.edu

Abstract

Unlike audio recording devices, a human speaker imitating a heard utterance or reading a sentence aloud must formulate a cognitive representation of the linguistic object to guide the phonology and phonetics of the spoken output. The current study used two different production tasks to explore the prosodic aspect of these representations: an imitation experiment in which speakers heard and then imitated spontaneous utterances from a Maptask corpus, and a read enactment task in which speakers read the same sentences aloud from a video display. For each task, the resulting utterances were compared for similarity a) to the original Maptask utterance and b) to each other. Similarity measures included perceptual accent and boundary labels and syllable durations, as well as Fujisaki model-based *F0* parameters. The imitations showed strong agreement with the stimulus utterances both in their phonological structure (perceptually labeled accents and boundaries), and in several phonetic cues to prosody from measures of duration and *F0*. Furthermore, agreement between imitated utterances and the original spoken stimulus was higher than between different imitations. Finally, read and enacted utterances were substantially different from the original spoken stimulus, in terms of their phonology and *F0* characteristics, though duration patterns were less variable. Overall, these results are consistent with the view that listeners extract the prosodic form of an utterance in terms of both phonological features and phonetic cues, and that the syntactic and semantic content of the text is not sufficient to determine a reliable prosodic outcome across subjects.

Index Terms: prosody, spontaneous speech, spoken imitation, phonetics and phonology, Fujisaki model

1. Introduction

Prosody in spontaneous speech exhibits remarkable variability both in the distribution of phonological prosodic features (pitch-accents and boundaries) and in their phonetic realization. This study uses two imitation tasks to investigate listeners' sensitivity to the phonological and phonetic encoding of prosody, examining the imitation of phonological prosodic categories and the imitation of their phonetic cues. Our research question concerns the degree to which speakers reproduce the phonological prosodic features and their phonetic cues. There are two possible outcomes. First, speakers may reproduce both the phonological and phonetic prosodic characteristics of the utterance as produced by the original speaker, or second, they may reproduce only the phonological characteristics using their own phonetic proclivities. For instance, a speaker may imitate an utterance by producing prosodic breaks and prominences at the same locations as in the original utterance, but implemented with different phonetic cues.

In the current study we adopt several different approaches to measuring prosodic similarity. On the phonological level, we examine prosodic agreement based on perceptual judgments of accent and boundary locations. On the phonetic level, among many phonetic cues to phonological prosodic features, we select quantitative measures of *F0* and syllable duration previously adopted for the assessment of the prosodic quality of synthetic as well as L2 speech [2]. Comparison of F0 contours as a phonetic cue to prosody is done based on parameterization of the F0 contour using the Fujisaki model [3], which provides a quantitative measure of the interval and timing of 'tone switches' (changes in the slope of F0 from rising to falling, or vice-versa) in relation to syllable onset and offset times and duration. This model reproduces *F0* from three components: a base frequency *Fb*, a phrase component and an accent component. Here we are mainly concerned with the alignment of the accent component, i.e. the response of the model to step-wise accent commands, defined by onset and offset times *T1* (*F0* rise) and *T2* (*F0* fall), and amplitude *Aa*. By evaluating the agreement of tone switches found in imitations of the same sentence, we hope to find correlates of prosodic similarity in the phonetic implementation of pitch accents in *F0* patterns.

2. Speech Material and Method of Analysis

32 utterances of spontaneous speech (7-15 words each) were extracted from 4 speakers in the American English Maptask corpus [4], and presented in auditory form to 10 participants, who were asked to "repeat the words and the way the utterance was said". This instruction was intended to elicit a reproduction of the utterance without focusing the speaker's attention too strongly on the phonetic detail. In particular, we did not intend the imitator to mimic the physiologically determined characteristics of the original speaker's voice, such as pitch range, aspects of voice quality, etc. The participants were requested to repeat each stimulus three times in succession. Stimulus utterances and imitations were transcribed for the location of prominences (accents) and boundaries (henceforth A/B annotation) using ToBI criteria, enabling comparison of the matched imitation and stimulus utterances for agreement in the location of these prosodic characteristics [5].

For the current study, a subset of 16 sentences produced by two of the Maptask speakers and the corresponding third imitation by six experimental subjects were chosen for analysis from the materials of Experiment 1. In addition, read-enacted versions of the same sentences were elicited from a different set of speakers. Preliminary results from six of these speakers are presented here. In this reading-aloud enactment task, participants were presented with written versions of the

sentences underlying the 16 stimulus utterances from the Maptask speakers, and instructed to read the excerpts aloud, as though they were the speaker in a map navigation activity.

All utterances were manually segmented on the word level and divided into pseudo-syllables at locations corresponding to dictionary syllabification using waveform and spectrographic displays in *Praat* [6]. These pseudo-syllables (henceforth syllables) were based on the text of the stimulus utterance of each sentence and then applied to the imitations with some modification. In some cases this involved associating a single pseudo-syllable in the utterance with two or three lexically specified syllables, due to segmental reductions or articulatory overlap. *F0* values were extracted using the autocorrelation method in *Praat* with a step size of 10 ms, and inspected for errors, and *F0* tracks were subsequently decomposed into Fujisaki-model parameters using automatic methods with manual correction [7,8].

Syllable durations were calculated from the syllable segmentation, and though this method does not resolve all problems in locating syllable boundaries, it is the case that all utterances were segmented according to the same criteria. This segmentation provides a way to line up corresponding intervals of matching paired utterances to compare duration as a correlate of prosody. The vectors of durations for matched utterances were correlated in a pair-wise fashion. Due to occasional slight differences in the wordings of imitations only those syllables present in both the stimulus utterance and its corresponding imitation were included in these analyses.

F0 analysis was done using two methods. First, we examined the correspondence between accent commands (in the Fujisaki framework) and prominent syllables (as determined by our perceptual Accent labels). Here we only consider the rising and falling tone switches aligned with Accented syllables, and their intervals expressed by the accent command amplitude *Aa*. In other words, we only take into account the subset of the *F0* parameters yielded by the Fujisaki model that can be associated with a phonological Accent label. Considering only this subset of tone switches, we calculate the agreement between matched pairs of utterances (from stimuli, imitations, and read utterances) with respect to tone switch location and direction, size of *Aa, and* tone switch alignment given as the accent command onset or offset relative to the syllable boundary.

Second, we calculated the similarity of complete *F0* contours for matched pairs of utterances based on the smooth and continuous log *F0* contours that are the output of the Fujisaki model, which are based on the superposition of the base frequency *Fb*, phrase component and accent component. Based on the syllable segmentation, a linearly time-warped version of the second utterance's *F0* contour is calculated which matches the timing of the first utterance. Both contours are normalized to a mean of zero,, and the root mean square distance (RMSD) as well as the correlation between the normalized *F0* contours are calculated. The *F0* contours are defined in the log *F0* domain, but to facilitate comparison we also converted those values back to a linear scale in the RMSD % value.

3. Results

Agreement in phonological labels for prosodic structure. Results from four subjects are reported here. Cohen's kappa statistics measure pair-wise agreement between each subject's imitations and the stimulus utterances. Agreement for the location of accents and boundaries was substantial to excellent, based on Kappa values for individual subjects from 0.61-0.71 (prominence) and 0.72-0.83 (boundary), showing subjects' sensitivity prosodic phonology of the stimulus. Higher agreement for Boundary suggests cues that are stronger or more reliable than for Accents.

Figure 1 displays bar-graphs of pair-wise agreement between stimulus (target) and imitations, as well as between imitations. We predict higher agreement for subject-stimulus pairs than for pairs of subjects (imitators), if subjects are in fact imitating the stimulus based on their perception of the phonological and phonetic prosodic form, yet two imitators may be reproducing a different subset of the features in the stimulus utterance. Under this scenario, the chance agreement between two imitators should be lower than the agreement between each imitator and the stimulus. This prediction is confirmed for the location of accents, but not for boundary location, suggesting that boundaries may be more reliably cued.

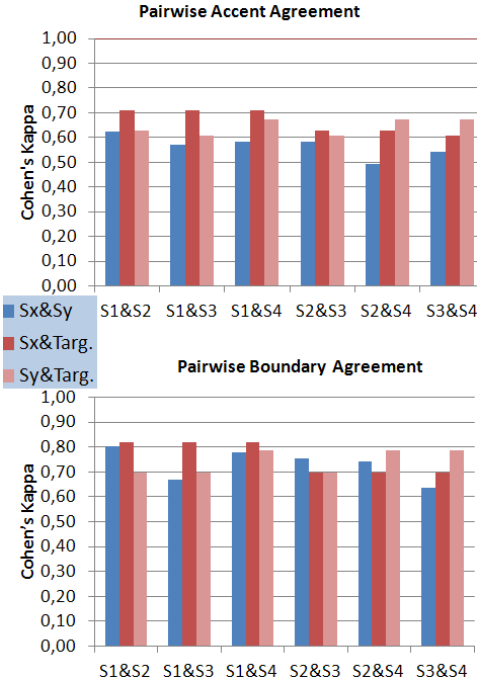

Figure 1: *Pair-wise agreement between stimulus and imitations by subjects 1-4 with respect to accent (top) and boundary (bottom) labels.*

Agreement between A/B annotation and Fujisaki model parameters. Figure 2, in panels (1) to (3), displays examples of Fujisaki model-based analysis of the stimulus utterance (1) *yup, towards the old mill, d' you see the old mill?* and imitations by speaker 6 and speaker 1. The three panels display, from the top to the bottom: the speech waveform, the F0 contour (extracted and modeled), and the underlying phrase and tone commands. The syllable segmentation is indicated by the dotted vertical lines. Syllable texts are provided, together with the associated A/B annotation. We examined the agreement between the A/B annotation with the associated accent and phrase commands of the Fujisaki model. 75% of syllables labeled as Accented were associated with an accent command. However, this percentage varies depending on the speaker in a range between 59% and 86%. 60% of

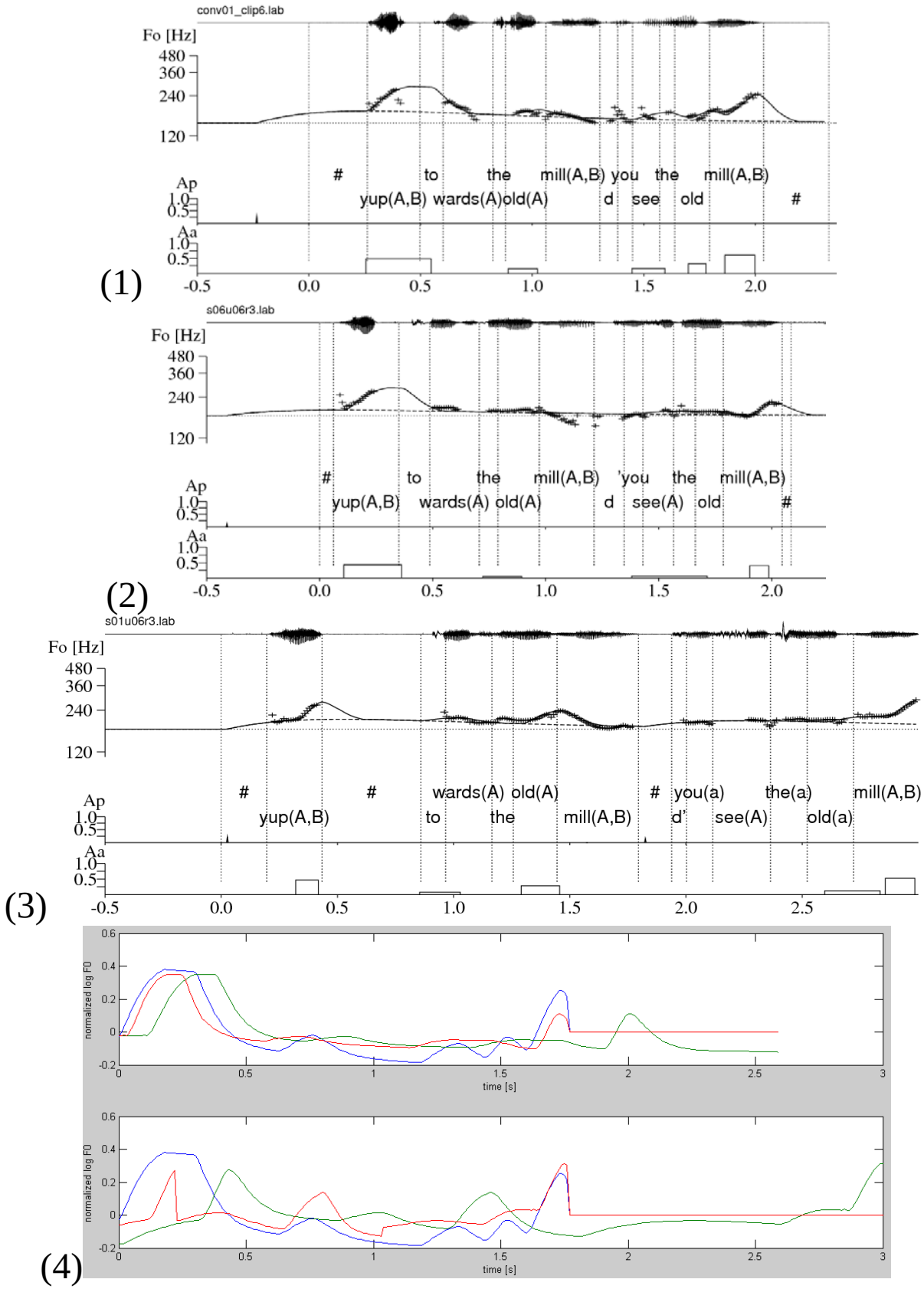

Figure 2: *Examples of Fujisaki model-based analysis. Sentence: "yup, to-wards the old mill, d' you see the old mill?" (1) stimulus utterance, (2)imitation by speaker 6, (3) imitation by speaker 1, and (4) target and time-warped f0 contours: stimulus utterance (blue), original imitation (green), time-warped imitation (red). Top: speaker 6 (r= .926, linear RMSD=8.0%), bottom: speaker 1(r=.536, linear RMSD=15.8%).*

labeled intra-utterance boundaries were aligned with a phrase command in the vicinity. Since utterances are produced in isolation there is no matching phrase command for utterance-final boundaries.

Agreement with respect to *F0* contours and syllable durations. Correlation analyses tested subject-stimulus similarity for measures of *F0* from the Fujisaki model and for syllable duration. These subject-stimulus correlations were significant for all subjects for three measures: syllable duration (Pearson's *r* values between .82-.87), contour type (rise vs. fall; *r* between .48-.74), accent command amplitude *Aa* (*r* between .50-.66) There was only sporadic, weak correlation between subject-stimulus in *F0* turning point alignment (*r* between .19-.41) with respect to the prominent syllable. We calculated mean correlation values for stimulus-imitation pairs on the one hand and matched imitation-imitation pairs on the other. For *Aa* and contour type these values are higher for stimulus-imitation pairs (*r*=.58 and .60, respectively) than when matched imitations are compared with one another (*r*=.51 and .48). For syllabic duration, mean correlations are .85 for stimulus-imitation and .83 for imitation-imitation pairs. These values again suggest a higher similarity between stimulus and imitation than between matched pairs of imitations.

The correlation between subject-stimulus normalized *F0* contours of matched utterances is somewhat variable across subjects. Lower correlations may occur even when two utterances display corresponding *F0* excursions on matched words if there are timing differences in the onset or offset of the *F0* excursions. The RMSD measure is less sensitive to differences in the timing of *F0* excursions, but is also a less sensitive measure when both contours have very small *F0* contours, in which case the RMSD will be small even if corresponding *F0* contours diverge in their onset or offset. The mean correlation between warped *F0* contours for stimulus-imitation pairs is significantly higher (mean/s.d.=.66/.18) than for imitation-imitation pairs (mean/s.d.=.56/.23, p<.01, Mann-Whitney U-test for independent samples). The RMSD, in contrast, is not significantly different (mean/s.d.=8.4/2.7% vs. mean/s.d.=8.7/2.4%). Together, the *F0* correlation and RMSD findings suggests that *F0* contour differences between paired utterances derive from differences in the timing of *F0* excursions and not solely due to major differences in the scaling of corresponding Accent-related *F0* contours.

Turning to the read, enacted utterances, we report here only on results from the analysis of normalized *F0* contours and duration (Analysis of Fujisaki-model parameters and A/B perceptual labelling is in progress.) Mean *F0* contour correlation was .56 (s.d.=.26) between read utterances and slightly lower when read utterances were compared with the original (unheard) stimulus (mean *r*=.51, s.d.=.24). This difference between imitated and read, enacted utterances is significant ($p < .027$, for the Mann-Whitney U-test). The difference in RMSD once again is not significant. These findings tell us that the read, enacted utterances are less similar to one another than are imitated utterances. They are also less similar to the stimulus, unsurprisingly, than the imitated utterances are. As with the imitated utterances, the lack of a significant RMSD indicates that dissimilarity in *F0* involving read utterances is not solely due to large differences in the scaling of *F0* contours. Comparing duration between matched pairs of read, enacted utterances, the mean correlations of syllabic durations of read utterances were fairly high (mean *r*=.81/s.d.=.11), just as we found with duration correlations between imitated utterances. But unlike the imitated utterances, the read utterances showed a lower correlation when compared with the original stimulus (mean *r*=.77, s.d.=.11), and this difference is significant ($p < .050$, for the Mann-Whitney U-test. These results again indicate that the read utterances are more similar to one another than they are to the (unheard) acoustic stimulus. This finding is surprising and will have to re-examined in the light of the fully analyzed data set on completion of this study. Comparison between imitations and read utterances yielded generally lower correlations, mean Pearson's *r*=.423 for *F0* contours and mean r=.78 for syllabic durations, a result which can be expected, as they do not share a common stimulus.

4. Discussion and Conclusions

Our observations showed that, with the exception of the phonological Boundary labels, the agreement between stimulus and imitation utterances was higher than the agreement between pairs of imitated utterances. This finding provides evidence that imitators perceive and reproduce the phonological prosodic features and at least some of their phonetic cues. In contrast, we find that read, enacted utterances are less similar to one another in their normalized *F0* contours, than are imitated utterances. This finding suggests that the syntactic and semantic content of the text is not sufficient to determine a reliable prosodic outcome across subjects. However, the fact that the syllable duration results for read, enacted utterances do not parallel the *F0* contour results suggests the need for future research examining a larger set of phonetic cues to prosody. Future research will extend the data set, perform a closer comparison of pair-wise similarities and continue the search for refined similarity measures.

5. Acknowledgements

We thank Dayna Cueva-Alegria for assistance with data collection and Angelika Hönemann for segmentation and data organization. This work was supported in part by grant NSF IIS 07-03624 to Cole.

6. References

[1] Mo, Y. 2011. *Prosody production and perception with conversational speech.* PhD diss., U Illinois.

[2] Mixdorff, H. and Ingram, J. "Prosodic Analysis of Foreign-Accented English", in *Proceedings of Interspeech 2009*, Brighton, England, 2009.

[3] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", J. of the Acoustical Society of Japan (E) 5(4), 233-241, 1984.

[4] Shattuck-Hufnagel, S. and Veilleux, N.M., "The robustness of acoustic landmarks in spontaneous speech", in *Proceedings of ICPhS2007*, Saarbrücken, Germany, 2007.

[5] Cole, J. and Shattuck-Hufnagel, S., "The phonology and phonetics of perceived prosody: What do listeners imitate?", *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[6] Boersma, Paul. "Praat, a system for doing phonetics by computer". *Glot International* **5:9/10**, 341-345, 2001.

[7] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", in *Proceedings of ICASSP2000*, Istanbul, 3:1281–1284, 2000.

[8] Mixdorff, H. (1/10/2009). *FujiParaEditor*, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html.