‘READ SPEECH NORMALIZATION’ (RSN): A METHOD TO STUDY PROSODIC VARIABILITY IN SPONTANEOUS SPEECH

Lena Zipp & Volker Dellwo

Phonetics Laboratory, English Department, University of Zurich, Switzerland
lena.zipp@es.uzh.ch; volker.dellwo@uzh.ch

ABSTRACT

A method, the ‘read speech normalization’ method (RSN), is proposed by which the variability of prosodic parameters (rhythmic durational and intonation) can be compared across different conditions of spontaneous speech. In an experiment using a customized variety of the map task method, spontaneous speech was elicited from two speakers of English in a formal and an informal situation. Sentences from the spontaneously spoken formal and informal situations were afterwards read by the same speakers. The formal and informal conditions were then compared in terms of their differences to the read speech version (read speech normalization). Results showed that meaningful differences could be observed between formal and informal speech from the read speech normalized sentences that could not be obtained by comparing the two spontaneous speech conditions directly.

Keywords: prosody, sociophonetics, spontaneous speech, normalization method

1. INTRODUCTION

It is a well-known problem in experiments focusing on prosodic variables like rhythm and intonation that there is a high between-utterance variability depending highly on the grammatical and lexical characteristics of an utterance (for speech rhythm see [2, 6]). For example, an utterance consisting of a main and a subordinate clause will have a different intonation structure and different rhythmical patterns from an utterance consisting of a main clause only. This becomes a particular challenge in studies in which prosody is compared between different categories of spontaneous speech, as it is hardly possible to elicit utterances which are grammatically and lexically identical. Spontaneous speech, however, is typically required in many areas of interest in prosody, e.g. when comparing prosodic variables between different speaking styles, sociophonetic situations, emotions or speech pathologies. In other words, the analysis of speech prosody is most interesting in cases in which it cannot be studied well.

One way of addressing this is by collecting very large sets of data for the categories to be compared [5]. This method, however, is extremely time intensive in studies in which extensive data editing is required. Another way of addressing this is to control the grammatical and lexical variability using read speech (possibly the most common way). This, however, is not applicable in many experimental set-ups as speech with varying emotional or stylistic characteristics typically requires to be elicited spontaneously.

Here we propose a method which may be a way out of this methodological dilemma in experimental prosodic studies. Instead of directly comparing the prosody of spontaneously produced utterances between different conditions, we compare their difference to a respective read version of each utterance. Thus, this method requires each participant to read out a transcript of their spontaneous utterances. For each spontaneous utterance, the difference to the read utterance is then calculated for a particular prosodic variable. The comparison between different spontaneous speech conditions (e.g. different formality levels) is then expressed in terms of the differences between a condition and the reading condition. We call this method ‘read speech normalization’ (RSN).

We have piloted the method studying the variability of prosodic intonational and durational parameters between formal and informal speech. In sociolinguistics, variation along a range of formal and informal speaking styles has been shown to be systematic for a great number of variables, including (segmental) phonetic, morphosyntactic, or lexical variables. These different formal and informal settings are usually created using different modes (e.g. reading vs. free speech), but also according to the audience design paradigm dependent on different interlocutors [1]. We investigated sociophonetic stylistic variation in spontaneous speech directed to two different interlocutors, one familiar and one unknown, with the latter additionally projecting social distance.

2. EXPERIMENT

2.1. Method

2.1.1. Subjects

Two male native speakers of English took part in the experiment. Male speaker one (age 25-29) was Scottish accented, male speaker two (age 35-39) Irish accented. Both speakers spoke their standard variety of English.

2.1.2. Material and apparatus

A map task [3] was used to elicit goal-directed spontaneous speech from subjects. Two participants were given identical sketches of maps, of which one contained a specific route and the other did not. The participant with the routed map (the speaker) had to explain the route to the other participant (the interlocutor). This map task was redesigned from typical map tasks - which are commonly used to study collaborative discourse - by removing mistakes that encourage interaction, by giving long place names to locations (e.g. ‘village where no-one is younger than 65’) and by adding elements that have to be identified descriptively. This step is not required for performing RNS, but it was deemed an effective way of limiting the variability between lexical items and grammatical constructions. Participants performed the map task in a quiet room sitting opposite each other at a table. A visual barrier in the middle of the table prevented them from being able to see each other’s maps. Participants were recorded using a Zoom H2 recorder. A stereo signal (16 bit, 44100 samples/second for each channel) was recorded from two microphones (one for each speaker). Speech recordings were then transferred to mono files and only utterances from the speaker were extracted in which the interlocutor was not audible.

2.1.3. Procedure

The map task was carried out twice with each speaker, first in an informal and then in a formal setting. The informal setting was created by having the speaker perform the map task with a friend, the formal setting with the experimenter (first author). For the informal setting, speaker and friend were told that this was a practice run preceding the actual experiment; they were not aware that this exercise was part of the experiment. The experimenter created a formal situation by maintaining a formal interaction style and projecting social distance (i.e., wearing a formal work outfit). Participants were informed about the experimental steps during a post-experimental debriefing, and their consent was taken.

From the elicited spontaneous speech material 10 utterances of more than 10 syllables were selected for each condition and each speaker (40 total); only utterances without false starts, hesitations or pauses were chosen. The two speakers who performed both map tasks (informal and formal) were invited for a second recording session during which they read the 20 utterances they produced under the previous spontaneous speech conditions.

2.1.4. Data analysis

From the 80 utterances (2 speakers * [20 read + 10 informal + 10 formal utterances]) five randomly selected utterances for each speaking style for each speaker (40 utterances in total) were annotated (segment durations) using Praat [4].

Upon auditory inspection of the speech material from the formal and informal conditions it was found that there may be differences in the way subjects used intonational and durational variables, in particular vocalic durations. For this reason we applied two measures, (a) mean duration and standard deviation of vocalic durations and (b) mean fundamental frequency and standard deviation to each utterance. As mean fundamental frequency varied between subjects, the standard deviation was normalized using the coefficient of variation (σ*100/ mean). Duration measurements in (a) were based on the annotated data (40 utterances), Intonation measurements in (b) were extracted automatically from the sound files using Praat and were based on the entire data set (80 utterances).

2.1.5. Read speech normalization (RSN)

RSN was applied by calculating the difference between the mean of a speaker's spontaneously elicited utterance (formal or informal) and the mean of his respective read utterance for a prosodic variable.

2.2. Results

2.2.1. Results for durational variability

Figure 1 contains the distributions of mean vocalic durations for the three speaking styles (formal, informal and read). The box-plots in the figure suggest that distributions across the three conditions are very similar (whiskers = total range, boxes = inter-quartile range, line = median). A univariate ANOVA (style * mean vowel duration) highly supports this view (F[3,39]=.01; p=.99).

Figure 1: Box-plot showing the distributions of vowel segment durations under the three conditions formal, informal and read speech.

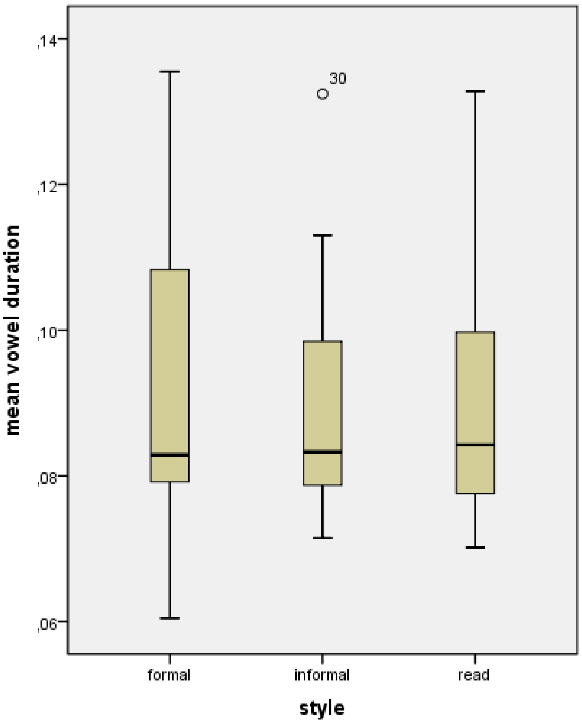
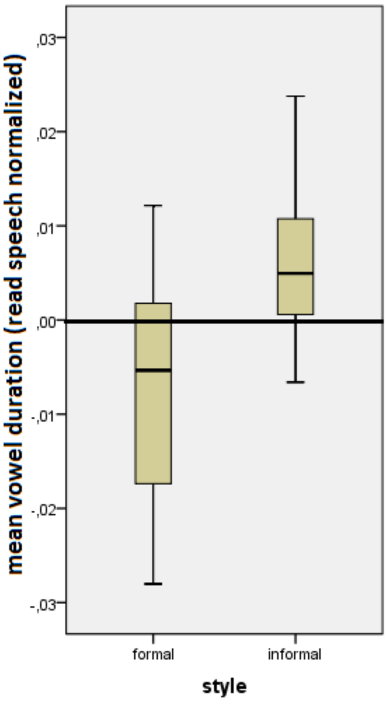


Figure 2 contains the distributions of the RSN mean vocalic durations (y-axis) for the formal and informal speaking style (x-axis). At zero there is no difference between the read version and the spontaneous version under observation (for better visibility the zero value has been highlighted across the box-plot with a black horizontal line). Values below zero indicate that the utterance average vocalic durations are shorter than in read speech, above zero they are longer. Two different observations can be made from Figure 2:

(a) There is a difference between each of the styles and read speech. Formal speech typically has shorter vowels, informal speech longer vowels than read speech. This effect, however, is only marginally significant in case of the informal condition (one-sample t-test: t[9]=2.15; p=.05) and not significant for the formal condition (t[9]=-1.68; p=.13). The difference is not obtainable from Figure 1 where we find no significant difference between any of the spontaneous speech conditions and read speech. When the differences between spontaneous and read speech are calculated for each individual sentence, however, meaningful differences are obtainable (Figure 2).

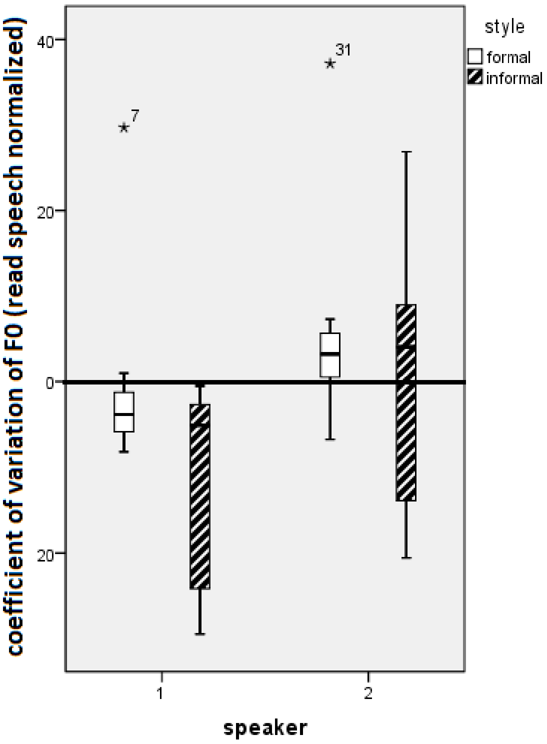
(b) There is a difference between the two spontaneous conditions in comparision to read speech. Formal speech shows shorter vowels compared to read speech than informal speech. This difference is significant (independent samples t-test: t[19]=2.67; p=.015). The data reveals significant results where a direct comparison of the variable between read and formal speech suggests that there are no differences (Figure 1).

Figure 2: Box-plot showing the distributions of the vocalic durational differences between read speech and the spontaneous speech conditions formal (left) and informal (right).



The results for vocalic duration variability are not presented here as no significant differences were obtainable between any of the groups in the raw and the RSN data. It remains unclear whether the mean vocalic duration is rather a correlate for rate or for rhythm (or another factor) in the present data. Traditional rate measures such as the number of syllables per second, however, did not arrive at a similar result.

Figure 3: Box-plot showing the distributions of the read speech normalized intonational variability between the two speakers (x-axis) for the formal (white) and informal (striped) conditions.



2.2.2. Results for intonation

Results for intonation, mean F0 and coefficient of variation, showed that there was no significant effect for either of the variables across the two subjects. However, looking at subjects individually, we found that there were strong between-subject differences. Figure 3 shows the results for the RSN F0 variability. It is apparent that the variance is higher in the case of informal speech compared to formal speech, again an effect which we were not able to obtain from comparing the conditions directly with each other. An ANOVA with factorial design (2x2; speaker*style) revealed no interaction between speaker and style (p=.4) but significant main effects for both factors (speaker: F[1,39]=6.4; p=.015; style: F[1,39]=4; p=.05). This shows that variability between speakers is possible but that both speakers show similar patterns: the variability of intonation is higher in informal than in formal speech compared to read speech.

3. DISCUSSION

In the present paper we presented a method to normalize the variability of spontaneous speech across different conditions (read speech normalization; RSN). Using pilot data from two male speakers we showed that spontaneous speech elicited in a formal and informal situation varies in the prosodic domains of duration and intonation when individual sentences are compared to read speech, but does not vary when the prosodic measurements are compared directly between conditions. As such the procedure normalizes for grammatical and lexical influences on prosodic parameters which prevent different sentences from being compared.

The present study is based on little data and an exemplary choice of variables. We are planning to evaluate the method on larger datasets. Given that some prosodic parameters (e.g. F0 variability) can be processed at least semi-automatically, a collection of a larger data-set seems feasible within a relatively short time.

4. REFERENCES

[1] Bell, A. 2006. Speech accommodation theory and audience design. In Brown, K. (ed.), *Encyclopedia of Language and Linguistics* (2nd ed.), Oxford: Elsevier, 11, 648-651.

[2] Dellwo, V. 2010. *Influences of Speech Rate on the Acoustic Correlates of Speech Rhythm: An Experimental Phonetic Study Based on Acoustic and Perceptual Evidence.* Ph.D. Dissertation, Universität Bonn. *http://hss.ulb.uni-bonn.de:90/2010/2003/2003.htm*

[3] Design of the HCRC Map Task Corpus *http://groups.inf.ed.ac.uk/maptask/maptask-description.html*

[4] Praat *http://www.fon.hum.uva.nl/praat/*

[5] Tortel, A., Hirst, D. 2008. Rhythm and rhythmic variation in British English: Subjective and objective evaluation of French and native speakers. *Electronic Proceedings of Speech Prosody* Campinas, Brasil, 359-362.

[6] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., Mattys, S. 2010. How stable are acoustic metrics of contrastive speech rhythm? *J. Acoust. Soc. Am.* 127(3), 1559-1569.