TAL2018, Sixth International Symposium on Tonal Aspects of Languages
18-20 June 2018, Berlin, Germany

Context dependent and time-course dependent prosodic analysis

Vered Silber-Varod[1], *Hamutal Kreiner*[2] *Noam Amir*[3]

[1]Open Media and Information Lab (OMILab), The Open University of Israel
[2]Linguistic Cognition Lab, Ruppin Academic Center, Israel
Dept. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University, Israel
vereds@openu.ac.il, hamutalk@ruppin.ac.il; noama@post.tau.ac.il

Abstract

In this preliminary study, we examine the change in prosodic parameters along personal interviews comprised of six questions. Our purpose is to demonstrate the importance of examining prosodic discourse context when studying conversational prosody. Findings show that as the interview unfolds, speakers tend to increase their speaking rate in longer IPUs while their intensity is lowered, and its instability increases. It is suggested here that the meaningful context that might explain the behavior of the acoustic parameters is the level of intimacy for one group of features, and the level of fluency and "easy going" topic for a different group of features. This preliminary study shows that context dependent and time-course dependent prosodic analysis are associated with the discourse content and can assist in understanding the discourse as a whole and the interaction in which it occurred.

Index Terms: Prosody of discourse, interview, acoustic features, intimacy, discourse context.

1. Introduction

One of the most important language skills for humans is the ability to adjust speech to the context. Speakers use a specific speaking style along a whole conversational unit, which is usually longer than a single utterance, be it a monologue or a dialogue. Nevertheless, prosodic research and prosodic typology mostly deal with the prosodic word (PrWd) and utterance levels, aiming to explore syntactic-prosodic constituency relation [1], following the Prosodic Hierarchy model ([2] and [3]). Thus, although it is plausible to assume that the phonetic realization of prosodic patterns is sensitive to the larger scopes of the linguistic context, these scopes are seldom investigated. In this paper we examine the time-course of prosodic parameters of speakers in a specific discourse context.

We define the *discourse context*, as the overall unit of spoken interaction that the speaker was engaged in ([4] and [5]). In the realm of discourse analysis, typically focusing on verbal expressions such as wording or phrasing, the notion of context is a key factor. Nevertheless, the scope of context is still controversial, and different approaches to data analysis propose different views. While most approaches involve a micro-level analysis of stretches of written or spoken texts, scholars are varied on the extent of contexts in which utterances should be analyzed [5]. [6] suggested the term *procedural consequentiality* to conceptualize the mechanism that links between the *context* and the *consequences* for the talk. According to [6], this mechanism "procedurally" connects between the context and what actually happens in the talk, instead of having a list of characterizations of the interaction that do not inform us about the production and perception of the details of its conduct. DiFelice Box [7] uses the notion of *procedural consequentiality* to demonstrate how context unfolds within the course of a classroom interaction. In [7], the discourse unit is a math lesson in a classroom. Seeking the context was also conceptualized as *Frame analysis* [8] and *Framing in Discourse* [9].

In the current study, we examine the prosodic patterns along a context unit of an interview and analyze interdependencies between prosody and the specific discourse context in which speech was produced. Specifically, we analyze prosody in the context of personal interviews. Recent studies examined the relationship between verbal content and prosody concordance as a cue of emotion regulation strategies during an interview [10]. The findings showed that the concordance between the acoustic and the verbal characteristics of emotionally related contents is strongly associated by the degree of attachment of the speakers to their family. Secure speakers showed high matching while dismissing speakers showed discrepancy between the valence of verbal expressions and prosody. However, in [10], only responses 3 and 4 out of 20 consecutive questions asked in an Adult Attachment Interview (AAI) protocol [11] were analyzed. This means that only the early stages of the interview were analyzed. Our assumption is that different concordance between verbal content and prosody might be revealed in later phases, due to context-dependent prosodic phenomenon such as accommodation and convergence. In this preliminary study, we examine how prosodic parameters change along personal interviews comprised of six questions. Our purpose is to demonstrate the need to examine the discourse context when studying conversational prosody.

2. Material and methods

2.1 Participants

Fourteen Hebrew female speakers participated in the present study, with an age range of 26–35 (mean=30). Participants were told that the aim of the research was to examine interpersonal interaction and communication differences. They were further informed that the interview would be recorded. All participants signed an Informed Consent form. The speakers participated in the study for course-credit, and voluntarily engaged in the interview.

2.2 The interview

The interviewer was a male student in the second year of an MA program in clinical psychology. Participants did not know him prior to the interview. The spoken interaction between the interviewer and the interviewee was minimal, as the interviewer asked a question, and the participant answered as much as

10.21437/TAL.2018-43

she/he wished, without any other prompt or interference from the interviewer. This left much room for the participant to manage responses according to her/his will.

The interview comprised six questions (as detailed below). The interview began with two casual "small talk" questions and the following four questions were designed to gradually increase the expected degree of self-disclosure:

Q1: What do you think about the weather today?

Q2: What do you think about reality shows on TV?

Q3: Tell me about your hometown and the neighborhood where you were raised.

Q4: Tell me about a meaningful person in your life.

Q5: What part does/has this person play/played in your life?

Q6: Tell me about positive and negative qualities of this person.

All interviews were recorded with a head-mounted Sennheiser MKE 2 microphone digitized with an Icicle 48V external sound card connected to a computer. The microphone was positioned at a fixed distance from the speaker's mouth, and the recording was carried out with a sampling frequency of 48 kHz, 16 bit sample resolution. This setting was previously mentioned in [12].

Figure 1 presents the average response (R) duration for each question (error bars reflect the variations among speakers). It is evident that the response duration increases gradually, except for R5. This deviation might be due to a "leakage", or merge, of R4 and R5. Average session duration is ~2 minutes (range from 70.437 seconds to 216.783 seconds).

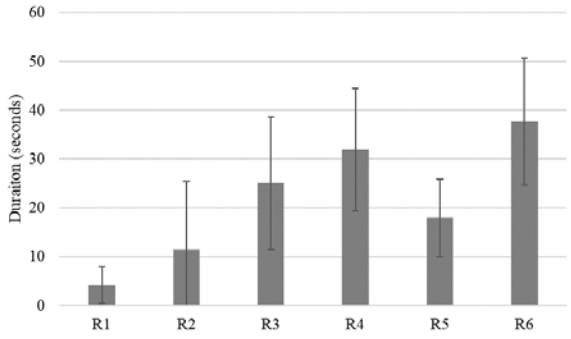

Figure 1: *Average response duration of each response (R).*

The goal of the study is to examine how acoustic and prosodic features change over time. Moreover, we sought to explore the relationship of prosodic features with response duration and with Inter-Pausal Unit (IPU) duration. Although the sessions are rather short, we hypothesized that changes will occur due to the intensive subversion towards intimacy, which characterizes the current interviews.

2.3 Acoustic and prosodic features

Twelve prosodic features were calculated, associated with the three main prosodic characteristics of rhythm, pitch and intensity:

Rhythm:

1. Duration of mean Inter-Pausal Unit (IPU) [13], with a minimum silent pause threshold of 100 milliseconds.
2. Mean fluent speech rate (Syllable per Second, SPS). Elongated syllables were discarded.

Pitch (normalized to the individual overall mean of each speaker):

3. Mean pitch (semitones (ST))
4. Mean Pitch (ST) variability (standard deviations)
5. Mean Pitch slope (Hz/sec)
6. Mean Pitch (ST) inter-Percentile. (10 to 90 percentiles)

Intensity (normalized to the individual overall mean of each speaker):

7. Mean Intensity (dB).
8. Mean intensity (dB) variability (standard deviations)
9. Mean Intensity slope (Pa/sec).
10. Mean Intensity inter-Percentile (dB)(10 to 90 percentiles)

Others:

11. Spectral slope (dB/Hz). Slope of the Long Term Average Spectrum (LTAS),
12. Mean jitter (%).

These acoustic features were calculated using custom written MATLAB software, apart from jitter and spectral slope, which were calculated using PRAAT software [14]. Each measure was first calculated for each IPU, then averaged across all IPUs for each participant separately, and then averaged across all participants. Outliers beyond three standard deviations from the overall mean value per each speaker were excluded.

3. Results

To explore the time course of the interview we analyzed first distribution of prosodic peaks along the interview. Second, we analyzed the relationship between different prosodic parameters along the interview.

3.1 Minimum and maximum peaks

Minimum and maximum peaks along the interview can be informative regarding changes in the course of the interview. Hence, we next examined the prosodic peaks of the different prosodic features as reflected in the minimum and maximum values of each trend-line. In the following we highlight only several trends. A summary is presented in Table 1.

Only Mean Intensity is highest at R1, while Intensity slope, Duration and Rate are lowest in R1.

IPU duration and speech rate, as well as other features, including Mean Intensity Variability, Intensity interpercentile, and Spectral Tilt exhibit their highest peaks in R2. Intensity variability and Intensity interpercentile show similar trend lines to the two rhythm parameters (Figure 2), which appears as a burst-like production of the speakers in the initial phase of the interview (R1 and R2) and a declination towards mean values for the rest of the session. On the other hand, Pitch Slope and Mean Jitter have their lowest values in R2. This makes seven extreme features for R2.

Highest values in R3 occur in four pitch features: Mean Pitch, pitch interpercentile, pitch slope, and pitch variability. These features demonstrate a similar trend line from beginning to end of the session, however, only two exhibit high correlation – Pitch variability and pitch interpercentile (Corr = 0.983). The

peak on R3 might be explained as a wider range of pitch, and higher values of mean pitch in R3.

R4 have the highest values for features Jitter and Intensity Slope and the lowest for Mean Intensity and Mean Pitch. Taken together, these findings suggest that instability of pitch, as reflected in speakers' jitter ratios, is strongest in R4, whereas voice intensity, as reflected in mean intensity and spectral tilt, tends to become lower in the course of the interview.

To further explore the general patterning of peaks, we summarized in Table 1 the minimum and maximum peaks as a function of the position of the response along the interview. In the Total column, it is shown that R2 was the most extreme, in terms of acoustic parameters with seven extreme parameters: Two parameters with minimum values and five parameters with maximum values. R3 has six extreme values, R1 and R4 has four each, R6 has only three extreme *minimum* values, while R5 has none.

Table 1: *Summary of minimum and maximum peaks per response.*

Response	Minimum	Maximum	Total
R1	3	1	4
R2	2	5	7
R3	2	4	6
R4	2	2	4
R5	0	0	0
R6	3	0	3

3.2 Relations between the prosodic parameters

Table 2 presents the correlations between the mean values of the twelve features across all answers. As can be seen on the table, different measures of pitch variability (Pitch variability, pitch slope and pitch inter-percentile) are highly correlated, so as different measures reflecting intensity variability. Based on these observation we have selected three different measures that represent speech rate, speech intensity and jitter and examined how they fluctuate along the interview, and to what extent these fluctuations correspond to IPU durations. Correlation coefficient measurements between the mean values of the twelve features for each pair of answers show that R2, R5 and R6 have almost perfect positive correlation (above 0.99). Overall, correlation between pairs of questions is positive and relatively high (lowest correlation of 0.598 is between R2 and R3. Figure 2 shows the relationship between IPU duration and the Rate parameter, as can be seen these two measures fluctuate along the interview in high correlation (R=0.911) indicating that speech rate increases as IPU duration increases. In the following, we explain this correlation by resonating to the responses' content: The first response (about the weather) was short in most of the cases, and with low speech rate. Example responses were [sababa] 'Great', [na'im] 'Nice'. The second question (about reality shows) provoked an enthusiastic response, as reflected in its length and its speech rate. Example responses were: [ʃeze mefager aval lifamim ani tsofa beze] 'it is stupid but sometimes I watch it', or [ani mexura lerealiti <laugh> tov <laugh>] 'I am addicted to reality shows <laugh>, well <laugh>' (laughs were not part of the analysis). The third and fourth responses demonstrate a convergence to the mean IPU length and speech rate of the speakers in this corpus. Responses to the third question included utterances such as […hu haya al hakarmel haya nof shel yam…] 'It was on Mount Carmel, with a view to the sea', or […hekarnu et kulam…] 'we knew everybody'. In the fourth response, subjects talked about varied personal connections, from first-degree family members, to their partners, individual friends, group of friends, and even their school instructors. R5 was an extension of R4. The sixth response was the longest, on average (Figure 1), probably due to slower speech rates associated with longer IPU duration (Figure 2). R6 included utterances such as […harbe peamim yesh lanu xilukey deot legabey kol miney dvarim…] 'in many occasions we disagree about certain matters', or [… ani yexola lehityaets ita yesh la xoxmat xayim…] 'I can consult her, she has life wisdom'. In general, these results suggest that speech rate and response duration are strongly related to the content of the question.

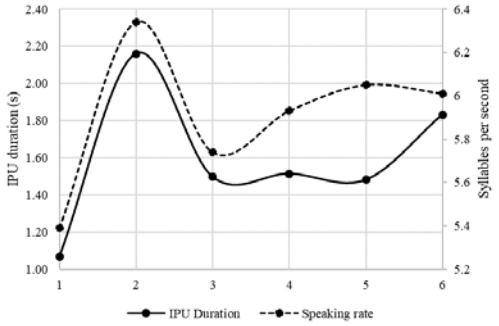

Figure 2: *IPU duration and speaking rate trends (Correlation = 0.911).*

Duration was found negatively correlated (R=-0.581) with Mean Intensity (Figure 3). This finding suggests that the shorter the IPU the higher the energy invested in producing it.

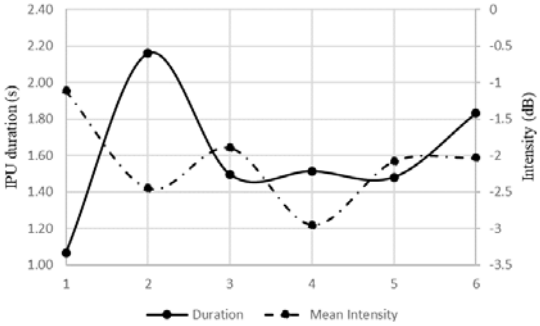

Figure 3: *IPU duration and Mean Intensity (Correlation = -0.581).*

Figure 4 presents the relationship between Jitter and Duration along the interview. As can be seen, there is only a low negative correlation (R=-0.183) between these parameters.

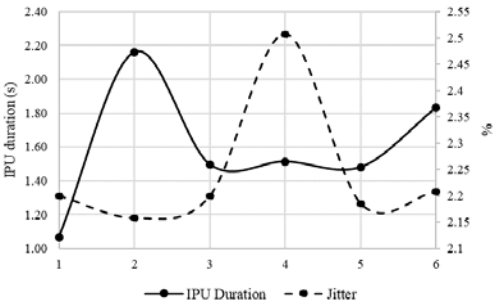

Figure 4: *IPU duration and Jitter trends (Correlation = -0.184).*

Table 2: *Correlation values between mean values of each of the twelve prosodic-acoustic features.*

Feature	1	2	3	4	5	6	7	8	9	10	11
1. IPU Duration											
2. Speaking rate	0.911										
3. Mean pitch	-0.559	-0.582									
4. Mean Pitch variability	0.007	-0.104	0.221								
5. Mean Pitch slope	-0.430	-0.500	0.367	0.756							
6. Mean Pitch inter-percentile	-0.026	-0.174	0.241	0.983	0.693						
7. Mean Intensity	-0.581	-0.721	0.731	-0.156	-0.040	-0.030					
8. Mean intensity variability	0.688	0.761	-0.583	-0.042	-0.636	-0.005	-0.429				
9. Mean Intensity slope	0.126	0.166	-0.403	0.080	0.449	-0.067	-0.688	-0.334			
10. Mean Intensity inter-percentile	0.758	0.794	-0.622	0.081	-0.528	0.106	-0.516	0.985	-0.231		
11. Spectral Tilt	0.191	0.117	0.174	0.688	0.098	0.763	0.087	0.504	-0.584	0.536	
12. Mean jitter	-0.184	-0.064	-0.533	0.134	0.440	0.046	-0.625	-0.172	0.781	-0.109	-0.378

4. Discussion

Our preliminary study showed that context dependent and time-course dependent prosodic analysis can shed light on the evolution of the discourse and the turning points along this evolution. As the interview unfolds, speakers tend to increase their speaking rate in longer IPUs while their intensity is lowered, and its instability increases. It is suggested here that the procedural consequentiality [6], i.e., the meaningful context that might explain the behavior of the acoustic parameters is the level of intimacy (with R4 as a pivot) for one group of features on the one hand, and the level of fluency and "easy going" topic (with R2 and R3 as pivots) for a different group of features on the other hand. Taken together the data suggests that whereas some prosodic parameters are fluctuate in high correlation along the interview, others are negatively correlated and yet others are relatively independent.

Our findings further extend previous findings that showed interesting fluctuations in the correspondence between verbal content and prosody along interviews [10]. Whereas in [10] only three prosodic parameters (Mean F0, F0 variability (range), and speech rate) were analyzed, here we reported twelve acoustic features. Nevertheless, since this study is limited regarding number of participants, and their gender (only females), further research is required to replicate, establish and extend these findings.

Importantly, the findings show that, unlike utterance level prosody, where initial units and final units tend to carry higher valence of prosody due to phonological phenomenon (anacrusis, initial rise, final lengthening, etc.) ([15] and [16]), the discourse unit level of analysis is less sensitive in its edges (start and end of the interview), or at least it can be said that edges attract minimum values while more prosodic valence is present in its middle phases.

These findings have important theoretical and practical implications. At the theoretical level, the present findings demonstrate that the phonetic realization of prosodic patterns is sensitive to the larger linguistic context. Consequently, when analyzing prosody, the contextual scope of the discourse unit should be considered. Future research that will further establish these findings and extend our understanding of discourse level prosodic patterns may entail important modifications in models of Prosodic Hierarchy ([2] and [3]).

Our findings suggest that speech-based dialogue systems, typically designed at the utterance level, should be designed to allow emulation of prosody that corresponds to the discourse context as well. As suggested by the emerging field of Conversation Intelligence ([17] and [18]), emulation of prosody that correspond to the time-course and discursive content might pave the way to novel data analyses of natural conversations that will promote the knowledge on human-human and human-machine verbal interactions.

5. Acknowledgements

We are thankful to Mr. Ronen Lovett and Ms. Niv Schleider who assisted in data collection and acoustic pre-processing. This study was supported by an internal research grant of Ruppin academic center.

6. References

[1] E. Selkirk, “On prosodic structure and its relation to syntactic structure,” *presentation given at MIT Linguistics 50th Anniversary*, December 10, 2011. Available at: http://ling50.mit.edu/wp-content/uploads/Selkirk-Slides.pdf

[2] E. Selkirk, “On prosodic structure and its relation to syntactic structure,” in T. Fretheim, (ed.), *Nordic Prosody II*. Trondheim: Tapir, 1978/81, pp. 111–140.

[3] M. Nespor, and I. Vogel, “Prosodic domains of external sandhi rules,” in H. van der Hulst and N. Smith, (eds.), *The Structure of Phonological Representations* (pp. 225–256), Part I, Dordrecht: Foris, 1982.

[4] C. Gordon, *Making meanings, creating family: Intertextuality and framing in family interaction*, New York: Oxford University Press, 2009.

[5] D. Delprete, and T. E. Tarpey, “Text and Context: The Role of Context in Discourse Analysis,” *Working Papers in TESOL and Applied Linguis*tics, vol. 10, no. 1, 2010. doi: 10.7916/D88D07VR

[6] E. A. Schegloff, “Whose text? Whose context?,” *Discourse & Society*, vol. 8, 165–87, 1997.

[7] C. DiFelice Box, “Classroom as Context: Procedural Consequentiality in a Secondary English Classroom,” *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, vol. 10, no. 1, 2010. doi: 10.7916/D88D07VR

[8] E. Goffman, *Frame analysis: An essay on the organization of experience*. Cambridge, Boston: Northwestern University Press, 1974.

[9] D. Tannen (Ed.), *Framing in discourse*. Oxford University Press, 1993.

[10] M. Spinelli, M. Fasolo, G. Coppola, and T. Aureli, “It is a matter of how you say it: Verbal content and prosody matching as an index of emotion regulation strategies during the Adult Attachment Interview,” *International Journal of Psychology*, 2017.

[11] C. George, N. Kaplan, and M. Main, “The adult attachment interview,” (Unpublished protocol). Department of Psychology. University of California, Berkeley, CA, 1985.

[12] V. Silber-Varod, H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, “Do social anxiety individuals hesitate more? The prosodic profile of hesitation disfluencies in Social Anxiety Disorder individuals,” *Speech Prosody 2016*, 2016, pp. 1211–1215.

[13] B. Bigi, and D. Hirst, “SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody,” *Speech Prosody 2012* (pp. 1-4), 2012.

[14] P. Boersma, and D. Weenink, Praat: doing phonetics by computer [Computer program]. Version 6.0.35, retrieved 16 October 2017 from http://www.praat.org/

[15] Nakajima, Shin’ya, and James F. Allen. "A study on prosody and discourse structure in cooperative dialogues." *Phonetica*50, no. 3 (1993): 197-210.

[16] Cruttenden, A., *Intonation*, Cambridge: Cambridge University Press, 1997.

[17] E. J. Glaser, *Conversational Intelligence: How Great Leaders Build Trust and Get Extraordinary results*. Bibliomotion, Incorporated, 2013.

[18] V. Silber-Varod, “Is human-human spoken interaction manageable? The emergence of the concept Conversation Intelligence,” *Online Journal of Applied Knowledge Management (OJAKM)*, International Institute for Applied Knowledge Management, 2018.