INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Prosodic convergence with spoken stimuli in laboratory data

Margaret Zellers

Department of Linguistics: English, University of Stuttgart, Germany
margaret.zellers@ifla.uni-stuttgart.de

Abstract

Accommodation or convergence between speakers has been shown to occur on a variety of levels of linguistic structure. Phonetic convergence appears to be a very variable phenomenon in conversation, with social roles strongly influencing who accommodates to whom. Since phonetic convergence appears to be strongly under speaker control, it is unclear whether speakers might converge phonetically in a laboratory setting. The current study investigates accommodation of pitch and duration features in data collected in a laboratory setting. While speakers in the study did not converge to spoken stimuli in terms of duration features, they did converge to an extent on pitch features. However, only some information-structure contexts led to convergence, suggesting that even in a laboratory setting, speakers are aware of the discourse implications of their production.

Index Terms: speech production, prosody, convergence, accommodation, discourse structure

1. Introduction

Recently the question of accommodation or convergence between speakers in a conversation has been receiving an increased amount of attention. Convergence has been shown to occur on the levels of syntax and semantic structure (e.g. [1], [2], [3]) as well as in the phonetic characteristics of spoken turns (e.g. [4], [5], [6]). However, these different types of convergence appear to behave differently, and have led to different theories about the mechanisms underlying convergence.

Syntactic and semantic structures have been demonstrated to prime structures occurring later in a conversation; for example, the order of the direct and indirect objects in English sentences can be influenced by hearing previous sentences with one or the other construction (cf. [3], *inter alia*). On the basis of these results, [3] propose that convergence on different levels of linguistic structure is a result of automatic priming effects; that is, conversational participants unconsciously use the same structures because these structures have already been activated. Their Interactive Alignment Model assumes that it is not parsimonious for speakers and listeners to maintain multiple representations of concepts, and that automatic priming processes simplify communication by keeping the number of active representations to a minimum. In this model, conversational participants have control over their convergence behaviors only insofar as they may make a decision about whether an interlocutor is someone that they wish to align with or not; if the social conditions are propitious, accommodation will necessarily occur.

These findings may be contrasted with those of [4] and others on phonetic, and particularly prosodic, accommodation. Accommodation on this level appears to be a much more variable phenomenon, with social roles influencing who accommodates to whom, and to what degree. Even when listeners are specifically asked to mimic what they hear, as in studies by [7], [8], and [9], what they specifically choose to mimic can vary, and may be different based on the time frame of imitation; for example, shadowing tasks may lead to more direct phonetic imitation than more delayed mimicry or conversational convergence ([8]). The results of studies on phonetic and prosodic mimicry are not in line with a mechanistic model such as that proposed by [3], but are more congruent with Accommodation Theory ([10], [11]), which proposes that convergence in conversation is strongly under the control of conversational participants, can be mediated by speaker or listener expectations, and does not necessarily have to involve complete matching. Furthermore, [5], [6], [7], [12] and [13] report that phonetic accommodation is not simply a linear process, but that it occurs dynamically over the course of conversations. In particular, [4], [7] and [14] point out that one possible reason for the variation in degree of prosodic convergence is the degree to which conversational participants are involved in or attentive to the conversation. [15] find that convergence in accommodation rate is correlated with participants' ratings of their interlocutor's likeability. Furthermore, [16], [17], and [4] report gender differences in accommodation, with speakers, especially female speakers, more likely to converge to male interlocutors than vice-versa.

Research into convergence is primarily conducted (with good reason) on spontaneous, conversational speech. Experimental investigations of similar phenomena are primarily psycholinguistic investigations of priming, as reported by [1], [2], and [3] *inter alia*. However, phonetic convergence may also be of relevance in phonetic production studies, especially when stimuli are presented auditorily instead of (or in addition to) visually. Since phonetic convergence appears to be less automatic than syntactic convergence, it is unclear to what degree production experiment participants might converge to stimuli that they hear. However, the degree to which they converge could be important in the interpretation of production data, since in such studies the researcher needs to be able to account for sources of variation in the data beyond that resulting from the different experimental conditions. Thus, the current study investigates the degree to which speakers in a production experiment (unrelated to the topic of convergence) accommodate to the phonetic features of auditorily-presented stimuli.

2. Methodology

2.1. Data

The data analyzed in this study were collected as part of a larger project on the realization of Contrastive Topics (CTs) in first- and second-language speakers ([18]). The current subset of data comprises recordings of 15 native speakers of Southern German producing utterances with three different CT types. The three conditions can be summarized as follows (for more detail, see [18], [19]):

http://dx.doi.org/10.21437/Interspeech.2016-238

- **Context-preserving**: Question requires a multi-part answer
- **Context-changing, move-insertion**: Question expects a single-part answer, but response is multi-part, indicating that the question was somehow insufficient
- **Context-changing, strategy shift**: Question expects a single-part answer, but response answers a different question, which should provide the requested information by means of a pragmatic inference

Participants read aloud sentences in response to stimuli questions which were presented both visually on a computer screen and auditorily through speakers; target and distractor utterances were presented in a semi-randomized order. Each participant produced a total of 24 target utterances, 8 each in the three different CT conditions, which are used in the current analysis. The prosodic characteristics of the participants' target utterances are compared with those of the stimuli recordings, which were produced by a female native speaker of Southern German.

2.2. Prosodic Labelling

The data were automatically aligned to labels using [20]'s aligner. The segmentations were then visually inspected and hand-corrected when necessary. For the fundamental frequency (F0) analysis, two pitch points were hand-labeled on the basis of visual inspection in Praat ([21]). For all files, the lowest reasonable F0 valley (excluding microprosody and creak) was labeled as Low (L), and the highest reasonable non-final F0 peak was labeled as High (H). Final H boundary tones (i.e. H%) are not included as part of the pitch range, since [22] provides evidence that listeners do not take boundary tone height into consideration when calculating the value of pitch range of a turn.

Three prosodic features are addressed in the current study. The first is the height of the highest non-final F0 peak in the utterance, measured in semitones (st) in order to normalize across speakers. The second prosodic feature is the pitch range of the utterance, also measured in st. The third prosodic feature measured is speech rate, in syllables per second. Any silent intervals in the utterances are ignored in this calculation. All values were extracted automatically using a Praat script; measurement errors such as octave errors were hand-corrected.

2.3. Hypotheses

Depending on the extent to which speakers in the study converge to the production of the auditory stimuli, several conditions are possible. Note that not all of these hypotheses are necessarily mutually exclusive, and that interactions among several sources of variation are likely.

2.3.1. Hypothesis set 1: Influence of convergence

a. Participants will produce relatively higher F0 peaks in response to stimuli with relatively higher F0 peaks

b. Participants will produce relatively wider F0 spans in response to stimuli with relatively wider F0 spans

c. Participants will produce a relatively faster speech rate in response to stimuli with a relatively faster speech rate

d. Participants' productions will resemble stimulus productions to a greater degree later in the experiment than earlier

2.3.2. Hypothesis set 2: Influence of information structure

a. Participants will speak more slowly in Strategy Shift contexts than in Context-Preserving or Move-Insertion CTs (S. Zerbian, personal communication)

b. Participants will produce higher pitch peaks and wider pitch ranges in Move-Insertion and Strategy Shift CTs than in Context-Preserving CTs, since there is a greater disjunction from the first speaker's assumptions in these cases

2.3.3. Hypothesis set 3: Individual differences

a. Some speakers will converge to the stimuli in terms of prosodic form, while others will not (cf. [4], [7], [15])

b. Female speakers will be more likely to converge to the stimuli than male speakers (cf. [4], [16], [17])

3. Results

Hypotheses were tested with linear mixed models in R ([23]) using the package lme4 ([24]). All p-values were calculated using the R package lmerTest ([25]), testing at $\alpha = .05$.

3.1. Pitch

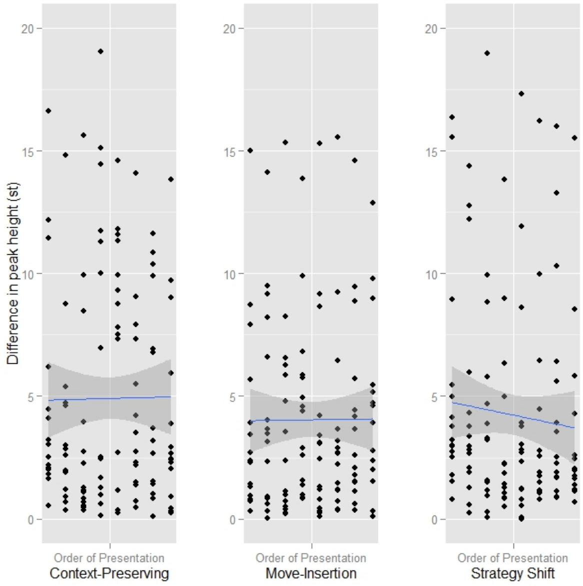

Figure 1: *Plot of absolute value of difference between stimulus peak height and participant peak height for each CT type. The x-axis represents order of presentation for the eight stimuli within each CT category.*

3.1.1. Pitch peak height

The statistical model gives evidence that speaker convergence with the spoken stimuli interacts with the CT category of the item they were producing; while the main effect for pitch peak height actually appears to slightly diverge from the stimulus peak height, there is convergence in Strategy Shift items, which improves in relation to the presentation order (i.e. higher convergence later in the experiment); this is illustrated in Figure 1.

Table 1: *Linear mixed model for pitch peak height. Note that the very high R^2 is likely due to the strong influence of the fixed factor gender and the random factor speaker identity on predicting F0.*

Category	Est.	SE	df	t-val	p-val
(Intercept)	107.65	3.52	315.8	30.62	.000
MoveIns	-18.68	13.87	334.7	-1.35	.179
StratShift	-36.12	11.01	318.1	-3.28	.002
StimHiPitch	-0.09	0.03	334	-2.55	.011
PresOrder	-0.003	0.06	334.9	-0.05	.962
GenderMale	-9.82	1.75	13	-5.61	.000
StimHi*MvIn	0.21	0.14	334.6	1.47	.142
StimHi*StSh	0.34	0.11	317.5	3.04	.003
PrsOrd*MvIn	-0.06	0.08	332.9	-0.70	.485
PrsOrd*StSh	0.17	0.08	334.1	2.22	.027

$R^2 = 0.938$

Formula: highpitch $\sim$ stimulushighpitch * CTcategory + presentationorder + gender + CTcategory:presentationorder + (1| speaker) + (1| stimulusrisefall)

Statistics are reported in Table 1. Although participant gender also played a role in predicting pitch peak height, there was no interaction between participant gender and stimulus peak height or CT category of the item, indicating that male and female speakers did not show different convergence patterns.

3.1.2. Pitch range

Unlike in the case of pitch peak height, speakers did not appear to differentiate the CT categories using pitch range. There was also no evidence of convergence in pitch range to the stimulus pitch range, despite the fact that participants modified their peak heights.

3.2. Speech Rate

Participants did not appear to converge with the stimulus speaker in terms of speech rate. However, they did modify their speech rate production on the basis of the CT categories, with speech rate being relatively slower in Context-Preserving CTs than in the two Context-Changing CT types, as shown in Figure 2; although there was substantial overlap in the categories' speech rates, a linear mixed model indicates that the differences between the means are nonetheless significant, cf. Table 2. Although a random intercept for speaker improves the model, random slopes did not, indicating that participant behavior followed more or less the same pattern in all cases.

Table 2: *Linear mixed model for speech rate.*

Category	Est.	SE	df	t-val	p-val
(Intercept)	6.15	0.16	29.2	39.27	.000
MoveIns	0.16	0.06	324.1	2.59	.010
StratShift	0.28	0.12	85	2.32	.023

$R^2 = 0.589$

Formula: speechrate $\sim$ CTcategory + (1| speaker) + (1| itemsylcount)

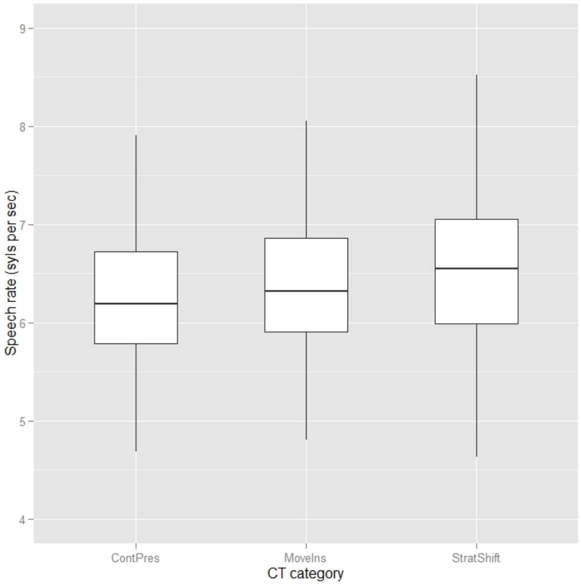

Figure 2: *Plot of speech rate for each CT type.*

4. Discussion

The current study investigated factors influencing the prosodic production of sentences in a lab setting, with a specific eye to identifying whether experiment participants might converge to the prosodic production of auditory stimuli. The results in relation to the hypotheses presented in section 2.3 are somewhat mixed.

With regards to the hypotheses about convergence, hypotheses 1a and 1d were partially confirmed, while hypotheses 1b and 1c were not supported; in the Strategy Shift CT context, participants' pitch peaks converged with the pitch peaks of the stimuli, with the differences between the peak heights reducing over the course of the stimulus presentation. With regards to the hypotheses about information structure, while an effect was found for speech rate in the Strategy Shift category, it was in the opposite direction to that predicted by hypothesis 2a; that is, Strategy Shifts were produced with an overall faster rate than the other CT turns. The results are difficult to interpret in terms of hypothesis 2b, since there was an interaction with convergence, but it is at least clear that while 2b may be true in terms of peak height in Strategy Shifts, it is not the case in Move-Insertion contexts. Finally, with regards to the hypotheses about individual differences, hypotheses 3a gains some support from the fact that the random factors for speaker were significant in the statistical models, but hypothesis 3b is not supported, since gender only became relevant in terms of absolute pitch differences between stimuli and participants.

4.1. Interpreting the speech rate results

Possibly the most surprising finding in the current study was the increased, rather than decreased, speech rate for Strategy Shift CTs. Although this prediction was based on a native speaker's intuition, it also had grounds in literature about topic changes, which indicate that slower speech rate is associated with new-topic utterances ([26], [27]). Although CTs need not lead to a topic change, they could be used to achieve such topic changes in conversation, particularly a stepwise topic change (cf. [28])

in which the topic change is accomplished by using an already-present conversational referent as a kind of pivot around which to move from one topic to a new one. However, in the current study, participants did not have to continue speaking after producing their CT utterances, and they may therefore have not treated them as new-topic beginnings.

One possible confounding factor in the measurement of the speech rate was that Context Preserving and Move Insertion contexts were generally produced with two intonational phrases, while Strategy Shifts tended to have only one intonational phrase, which was longer than the individual intonational phrases in the first two contexts, but shorter than the full durations in those contexts. To an extent, the fact that both the Move Insertions and Strategy Shifts were produced with a faster speech rate than the Context Preserving CTs argues against this finding being solely a result of the intonational phrasing. Nonetheless, several methods of addressing speech rate variation due to phrase length were attempted in the statistical modeling phase, but none achieved substantially different results than those reported above using the full syllable count for each item. A visual inspection of all of the data with speech rate plotted against syllable duration suggests that some speakers modified their speech rate to be faster with longer stimuli, while other speakers did not. Thus the non-modifying speakers may have cancelled out the effect in the statistical model. More analysis is required to determine whether this is in fact the case.

4.2. Does this convergence matter?

The effects reported here for convergence to the experimental stimuli are fairly small. However, they are worth drawing attention to from the point of view of considering the degree to which participants are attentive to and engaged in the linguistic task at hand. The question of participant attentiveness in experimental settings is often raised in considering the quality and generalizability of experimental results. Recently, work by [29], [30], and [31], among others, has raised the issue of the degree to which inattentive participants' responses can skew results. They suggest that up to 46% of survey responses are affected by inattention, although they point out that this number may differ when experiments are carried out in a laboratory ([31]).

A common criticism of laboratory data in phonetics and phonology is that participants may not produce speech in real life in the same way that they produce it in the lab. One reason for this is that the laboratory setting is considered artificial, and thus participants do not condition their language use to any social demands—or at least not the same social demands that would arise in conversation. However, the effect of convergence reported here suggests that this is not entirely true. [32] report that conversational participants are more likely to match their prosody across contrastive forms in a conversation when they are attempting to minimize or disguise some kind of potential disjunction. The Strategy Shift CTs are such a case of possible disjunction, where a question is answered indirectly at best. By converging prosodically in this CT context, the participants can thus be interpreted to be attentive at least to some extent to the social implications of such a construction, even within the laboratory setting. Although the convergence in this case was fairly minimal, the entire recording procedure for each participant only lasted about 20 minutes, and there were also only eight stimuli per participant in the Strategy Shift context; a longer experiment with more stimuli in this condition might attain a greater degree of phonetic convergence.

4.3. Influence of laboratory context

Speech researchers in more interaction-oriented fields may be tempted to dismiss laboratory data as lacking context, while researchers using laboratory data may dismiss spontaneous speech as too messy. Instead, the results of the current analysis can offer a third perspective: that different aspects of speech production are prioritized in different contexts. Stronger effects of convergence would no doubt be found in a more spontaneous setting, but this should not mean that we assume that no formal structure underlies that speech. Similarly, stronger formal effects of the discourse structure may be found in laboratory speech, but as we have seen in the current data, this does not mean that the interactive features are completely eliminated. Thus a balanced approach to prosody research requires both kinds of data in order to gain a more well-rounded view of the phenomena in question.

5. Conclusions

This study reports on the effects of prosodic convergence with spoken stimuli in a laboratory setting, finding a modest effect of such convergence in interaction with other discourse-structural factors influencing the prosodic production of such turns. It thus provides evidence that even in laboratory settings, speakers can and do take social or interactive context into account in their productions. Thus phonetic investigations cannot simply assume that laboratory data is free from the influence of such context, even though its effects may be different. Instead, more effort should be made to identify and quantify such effects in order to improve our ability to correctly interpret our data.

6. Acknowledgements

This research was carried out as part of project A7 of the SFB 732 (University of Stuttgart), funded by the German Research Foundation (DFG). I am grateful to Sabine Zerbian and Nadja Schauffler for helpful discussion on these topics, as well as to our research assistants Sabrina Beer, Yulia Pavlyuchenko, and Nicole Stirm for their assistance with data collection and processing.

7. References

[1] W. J. M. Levelt and S. Kelter, "Surface form and memory in question answering," *Cognitive Psychology*, vol. 14, pp. 78–106, 1982.

[2] J. K. Bock, "Syntactic persistence in language production," *Cognitive Psychology*, vol. 18, pp. 355–387, 1986.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[4] J. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustical Society of America*, vol. 119, pp. 2382–2393, 2006.

[5] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[6] T. Gijssels, L. Staum Casasanto, K. Jasmin, P. Hagoort, and D. Casasanto, "Speech accommodation without priming: the case of pitch," *Discourse Processes*, 2015.

[7] C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring dynamics of mimicry by means of prosodic cues in conversational speech," in *Proceedings of ICPhS XVII, Hong Kong*, 2011, pp. 1294–1297.

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

[9] M. D'Imperio, R. Cavone, and C. Petrone, "Phonetic and phonological imitation of intonation in two varieties of Italian," *Frontiers in Psychology*, vol. 5, p. 1226, 2014.

[10] N. Coupland and H. Giles, "Introduction: the communicative contexts of accommodation," *Language & Communication*, vol. 8, no. 3, pp. 175 – 182, 1988.

[11] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: communication, context, and consequence," in *Contexts of Accommodation: Developments in Sociolinguistics*, H. Giles, J. Coupland, and N. Coupland, Eds. Cambridge, UK: Cambridge University Press, 1991, pp. 1–68.

[12] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011, pp. 3081–3084.

[13] F. Bonin, C. De Looze, S. Ghosh, E. Gilmartin, C. Vogel, A. Polychroniou, H. Salamin, A. Vinciarelli, and N. Campbell, "Investigating fine temporal dynamics of prosodic and lexical accommodation," in *Proceedings of 14th Interspeech, Lyon, France*, 2013.

[14] A. Gravano, Š. Beňuš, R. Levitan, and J. Hirschberg, "Three tobi-based measures of prosodic entrainment and their correlations with speaker engagement," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, Dec 2014, pp. 578–583.

[15] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of Interspeech 2013, Lyon, France*, 2013.

[16] R. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[17] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender differences in vocal accommodation:: The role of perception," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 422–432, 2002.

[18] S. Zerbian, G. Turco, N. Schauffler, M. Zellers, and A. Riester, "Contrastive topic constituents in German," in *Proceedings of Speech Prosody, Boston, USA*, 2016, pp. 345–349.

[19] V. Gast, "Contrastive topics and distributed foci as instances of sub-informativity: A comparison of English and German," in *Comparative and Contrastive Studies of Information Structure*, C. Breul and E. Göbbel, Eds. Amsterdam: John Benjamins, 2010, pp. 15–50.

[20] S. Rapp, "Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models—an aligner for German," in *Proceedings of ELSNET Goes East and IMACS Workshop Integration of Language and Speech in Academia and Industry*, Russia, 1995.

[21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/, 2016.

[22] C. Gussenhoven, "Intonation and interpretation: phonetics and phonology," in *Proceedings of Speech Prosody, Aix-en-Provence*, 2002, pp. 47–57.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[24] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[25] A. Kuznetsova, P. Bruun Brockhoff, and R. Haubo Bojesen Christensen, *lmerTest: Tests in Linear Mixed Effects Models*, 2016, r package version 2.0-30. [Online]. Available: http://CRAN.R-project.org/package=lmerTest

[26] J. Hirschberg and B. Grosz, "Intonational features of local and global discourse structure," in *Proceedings of the Speech and Natural Language Workshop*, Harriman, NY, 1992, pp. 441–446.

[27] M. Zellers, "Prosodic detail and topic structure in discourse," Ph.D. dissertation, University of Cambridge, 2011.

[28] G. Jefferson, "On stepwise transition from talk about a trouble to inappropriately next-positioned matters," in *Structures of social action: studies in conversation analysis*, J. Atkinson and J. Heritage, Eds. Cambridge, UK: Cambridge University Press, 1984, pp. 191–222.

[29] D. M. Oppenheimer, T. Myevis, and N. Davidenko, "Instructional manipulation checks: detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, vol. 45, pp. 867–872, 2009.

[30] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data," *Psychological Methods*, vol. 17, pp. 437–455, 2011.

[31] M. R. Maniaci and R. D. Rogge, "Caring about carelessness: participant inattention and its effects on research," *Journal of Research in Personality*, vol. 48, pp. 61–83, 2014.

[32] M. Zellers and R. Ogden, "Exploring interactional features with prosodic patterns," *Language and Speech*, vol. 57, no. 3, pp. 285–309, 2013.