ACOUSTIC VARIATION IN SPEECH: CONTRASTING INITIAL AND LATER STAGES OF CONVERSATIONS SHOWING OPINION CONVERGENCE AND DIVERGENCE

Charlize Ma [*1], Jahurul Islam [†1], Effie Kao [‡1], Raechel Kitamura [♦1], Stephanie Wang [♯1], Marcell Maitinsky [♩1], and Bryan Gick [♠1]

Department of Linguistics, University of British Columbia, Vancouver, Canada

1 Introduction

The phenomenon of speech accommodation is influenced by various factors, including interlocutor traits, social identity, context, power dynamics, and speech goals [1]. Additionally, a speaker's perception of positive or negative alignment with their interlocutors can also influence speech patterns [2, 3]. The duration of a conversation is another important factor that can impact speech, as individuals may modify their speech patterns as the interaction progresses. Despite extensive research on many of these factors, the effects of speakers' opinion convergence or divergence and the duration of interaction on speech characteristics are relatively new areas of interest [4]. While some studies have explored the impact of duration on speech features (e.g., F0) in human-robot computer games, finding no evidence of accommodation [5], there is a lack of research on how the duration of interaction specifically interacts with the expression of explicit opinion convergence or divergence.

To address this gap, the current study aims to investigate whether the duration of interaction influences the acoustic characteristics of speech produced by individuals during instances of opinion convergence (agreement) or divergence (disagreement) in relation to their interlocutors. By examining this relationship, the study seeks to provide valuable insights into the dynamics of speech accommodation and shed light on how the duration of interaction may shape the expression of opinion alignment or contrast.

2 Methods

2.1 Data

The data came from speakers from a YouTube channel called the *Ellen Fisher Podcast*. Speech from several individuals was extracted from three polarized conversations. The topics included "Plant vs. Animal Regenerative Farming", "Pro-Life vs. Pro-Choice" and "Vegan vs. Animal Foods." The videos did not specify the speaker's language background, but all appeared to be native-level American English speakers (M:4, F:4). Speakers' presumed ages were around 30-60 years, and all speakers were knowledgeable in their respective fields, so they were presumed to have a high education level.

[*] sy.charlize@gmail.com
[†] jahurul.islam@ubc.ca
[‡] effiekao@student.ubc.ca
[♦] raechelk@student.ubc.ca
[♯] smwang@student.ubc.ca
[♩] mlmtnsky@student.ubc.ca
[♠] gick@mail.ubc.ca

The videos had two different discussion structures. In the "Plant vs. Animal Regenerative Farming" video, the four speakers were led by a host to take turns speaking for a few minutes each time. In the other two videos, the two speakers were allowed to speak freely throughout the whole discussion. In each video, the host would encourage turn-taking between interlocutors and guide the conversation by introducing new topics and posing questions.

2.2 Annotation and Processing

From each episode of the selected podcasts, two chunks (each about 20 minutes long) were extracted: the first one was from the initial 30 minutes of the episode while the second one was about one hour into the conversation. Audio was extracted from each video clip and then all the audio clips were transcribed using Praat TextGrids [6]. Next, the phone-level transcriptions were generated using the Montreal Forced Aligner [7]; any errors in the automatic alignment were manually corrected.

Chunks of speech were then manually inspected to identify and code three types of events representing an opinion category: 1) when a speaker expresses an explicit agreement (convergence) with an interlocutor, 2) when a speaker expresses an explicit disagreement (divergence) with an interlocutor, or 3) when a speaker produces a "neutral" statement.

2.3 Analysis

We extracted the fundamental frequency (F0), the first formant (F1) and the second formant (F2) at the midpoint for the monophthongal vowels in the speech using FastTrack [8] which is a Praat-based toolkit for measuring vowel formants in an optimized way. The formant values were normalized in R [9] using the Lobanov method [10]. Statistical analysis of the data was performed using R to investigate the effect of the conversation stage on F0, F1 and F2 within each opinion category.

3 Results

Figures 1, 2 and 3 present the distributions of normalized F0, F1, and F2, respectively, across opinion categories and conversation phases. The x-axis represents whether the data points belong to an event where the talker expressed an agreement ("converged"), a disagreement ("diverged") or neither ("neutral"). Phases of the conversation are marked using colours ("part1" = towards the beginning; "part2" = towards the middle or end).

As Figures 1 and 2 reveal, the distributions of F0 or F1 values across conversational stages are very similar, which is

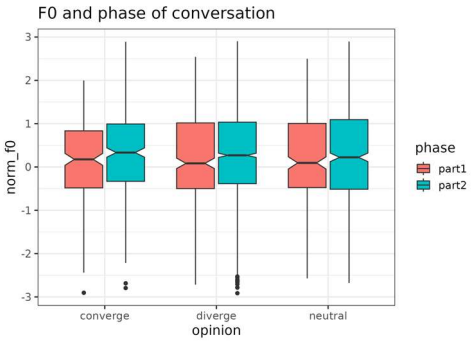

Figure 1: F0 and conversation stage

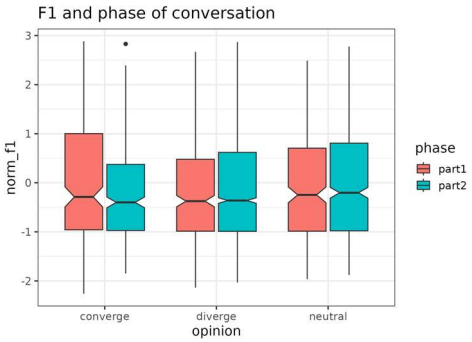

Figure 2: F1 and conversation stage

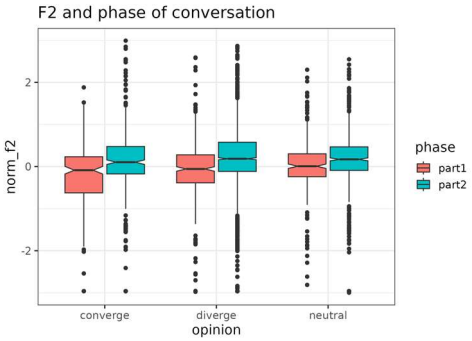

Figure 3: F2 and conversation stage

consistent for the events of both convergence and divergence. This indicates that speakers kept using similar ranges of pitch and vowel heights across the whole conversation and did not vary their pitch or vowel height no matter how far along they are and whether they expressed an opinion or remained neutral in the conversation. Contrarily, Figure 3 reveals that conversational stage. More specifically, vowels in the second phase of the conversations tended to have higher F2 values indicating more fronted vowels.

To investigate the effect on conversational phase on acoustic features within each opinion category, separate linear mixed-effects models were fitted using the R package *lmerTest* [11] for "converge", "diverge", and "neutral" opinion categories. In each model, fixed effects were included for the variables "phase" and "vowel" and random intercepts were included for individual "speakers". P-values were obtained via t-tests using Satterthwaite's method, as provided by the *lmerTest* R package [11].

Results of the linear mixed-effects models revealed a significant effect of conversational phase on F2, both in the events of convergence ($t(387) = 4.35$, $p < 0.001$) and divergence ($t(32) = 4.49$, $p < 0.001$). In both cases, F2 had a significant increase in part 2 of the conversation. Crucially, the effect of phase was not statistically significant in the neutral condition ($t(97) = 1.16$, $p = .25$). No such patterns were observed for F0 and F1.

4 Conclusion

Results of this study revealed that there was a significant effect of the phase of conversation on F2 in the event of convergence or divergence of opinion; similar effect was not confirmed for neutral speech. This suggests that speakers used specific strategies (i.e. making vowels frontier) when they had an opinion to express. For F0, our results are consistent with human-robot interactions [5], showing no accommodations. These results add a new dimension to our understanding of speech accommodation in addition to the affecting factors reported in previous studies (as in [1-3]) by providing evidence that opinion convergence or divergence interacts with the duration of interaction in natural speech.

Acknowledgments

This study was supported by NSERC.

References

[1] Pardo, S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Vocal accommodation in speech communication. *J. of Phon.*, *95*.

[2] Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*(4), 437-456.

[3] Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. of Phon.*, *40*(1), 177-189.

[4] Ma, C., Kao, E., Kitamura, R., Wang, S., Islam, J., De Boer, G. & Gick, B. (2023) Relations between Opinion Convergence, Acoustic Convergence and Movement Convergence in Interlocutors. In *Proceedings of the International Symposium on Phonetics & Cognitive Sciences of Language*, 62-63.

[5] Ibrahim, O., Skantze, G., Stoll, S., & Dellwo, V. (2019). Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing. *Interspeech*.

[6] Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.10.

[7] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*.

[8] Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, *7*(1). https://doi.org/10.1515/lingvan-2020-0051

[9] R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[10] Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *J. of Acou. Soc. of America*, *116*(5), 3099-3107.

[11] Kuznetsova A, Brockhoff PB, & Christensen RHB. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *J.of Statistical Software*, *82*(13), 1-26.