Interspeech 2024
1-5 September 2024, Kos, Greece

Uh, *um* and *mh*: Are filled pauses prone to conversational convergence?

Mathilde Hutin[1], *Junfei Hu*[1], *Liesbeth Degand*[1]

[1]University of Louvain, Institute for Language and Communication, Belgium

mathilde.hutin@uclouvain.be, junfei.hu@uclouvain.be, liesbeth.degand@uclouvain.be

Abstract

Filled pauses are very frequent words, such as *uh*, *um* or *mh* in English. They have been shown to serve many purposes in interaction management, yet whether these conversational functions are intended by the speaker to ease the interaction (hearer-oriented) or just interpretations by the interlocutor of otherwise unintentional vocalizations (speaker-oriented) remains an open question. To participate in this debate, we investigate whether filled pauses converge in terms of form (*uh*, *um* or *mh*) in the course of 5 conversations made up of 12 asymmetrical sub-conversations. Results show that each conversation displays a different pattern in the choice of the filled pause, yet none of them shows any clear evidence that participants align with their interlocutors. This indicates that, unlike other frequent words from the lexicon, filled pauses do not converge, at least in their form, and that their use is more likely speaker-oriented than hearer-oriented.

Index Terms: filled pauses, disfluency, convergence, discourse management

1. Introduction

Filled pauses, sometimes also called "vocalizations" or "hesitations", designate very frequent utterances such as "uh", "um" or "er" in English, "äh" and "ähm" in German, "euh" or "euhm" in French, etc. They have been shown to qualify as words in written language [1] but their status in spoken language is still debated: Some suggest that filled pauses in spontaneous speech are "on a cline of wordhood" [1] while others suggest that they are full conventional words in English [2, 3] or German [4].

In any case, in written language, filled pauses have been shown to convey extra-propositional meaning such as attitude or style [1] or to add subtext [4]. In spontaneous speech, they have been shown to fulfill similar functions and more, such as signaling major syntactic or discursive boundaries [5, 6], heightening attention [7], coordinating turn-taking [8, 9], holding the floor [10, 11], informing the listener about meta-cognitive states of the speaker [2, 12], etc. However, it is not clear whether these speech-specific functions are intended by the speaker, i.e., filled pauses are consciously used to convey this kind of meta-linguistic information, or whether they are the result of the interpretation by the interlocutor of what are mere expressions of buffering or speech planning by the speaker. The first scenario is what we will call the "hearer-oriented" use of filled pauses, the second "speaker-oriented".

In this study, we propose to participate in this debate by testing how filled pauses converge in terms of form (i.e., "uh" *vs* "um" *vs* "mh") in English dialogues. Convergence is a linguistic phenomenon whereby participants in an interaction tend toward a common speaking style [13, 14] – a mechanism which has been shown to help establish a shared conceptualization among speakers [14], to improve the success of communication [15, 16, 17] and to be favored by highly frequent words ([17] despite [18]). The rationale behind our operation is that, if filled pauses are hearer-oriented, they should be prone to convergence, as they are frequent words and thus ideal loci for implementing such conversational device. On the other hand, if filled pauses are speaker-oriented, they are considered as non- (or less-) interactional items and thus they would be less prone to convergence.

Past literature is surprisingly scarce regarding convergence of filled pauses in spontaneous speech, considering the potentially far-reaching insight its investigation could bring to the study of discourse management and to both communication studies and cognitive linguistics in general. Regarding their phonetic patterns, Beňuš and colleagues investigate filled pauses only indirectly, as turn-initial single-word utterances among others (along with "okay" or "mhm") and show that the convergence of their patterns are informative of the dominance relationships between interlocutors [8]. They also investigate three-speaker asymmetrical conversations from Supreme Court's arguments and show that the convergence of the acoustic characteristics of filled pauses during the whole session does not correlate with the justices' votes, however local convergence does correlate with the success of the interaction [19, 20]. Regarding the frequency of occurrence of filled pauses, it is investigated in Hebrew spontaneous speech, where the rates of filled pauses in asymmetrical dialogues tend to converge [21], as well as in non-native German speech, where the proportions of filled pauses are shown to be highly idiosyncratic and depend only marginally on the interlocutor [22]. Regarding the form of the filled pause, i.e., vocalic (the "uh" type), vocalic-nasal (the "um" type) or only nasal (the "mh" type), it has been investigated only in non-native German speech, where it is also idiosyncratic and mostly independent of the interlocutor [22].

In the present study, we participate in this field of research by investigating the use of various forms of filled pauses, i.e., "uh", "um" and "mh", in five task-oriented, same-gender dyadic conversations in English. We present these conversations, i.e., how they were recorded, transcribed and analyzed in Section 2. We then present the overall patterns for each conversation in Subsection 3.1 before presenting more fine-grained results on the alignment of the vowel-to-nasal ratio in Subsection 3.2. Finally, we conclude and discuss the results in Section 4.

2. Data and Methodology

2.1. Task

Data elicitation consists of speech from two-participant (dyadic) conversations resulting from a description-drawing

10.21437/Interspeech.2024-1168

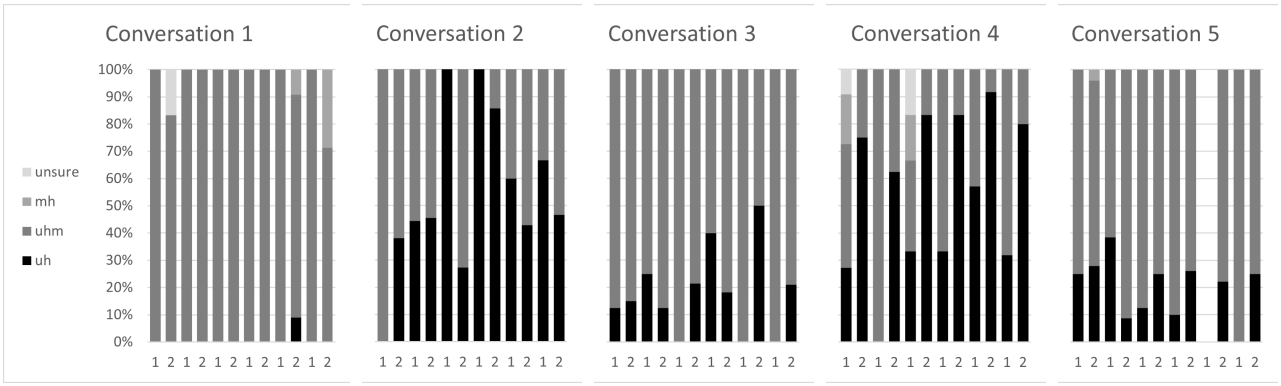

Figure 1: *Rates of various forms of filled pauses ("uh", "um" and "mh", as well as unclear cases ("unsure")) per speaker (1 or 2) per trial per conversation.*

task involving 2D object images adapted from [23]). To that extent, two previously unacquainted participants are paired into one conversational pair, or dyad. Participants in a dyad always share the same gender to avoid unbalanced propensity to alignment within a dyad, given that women have been shown to actively align with the linguistic performance of male interlocutors while the reverse is not true [24, 25].

Participants are instructed to alternatively describe the picture of an object to their dyadic partner and draw an object described to them by their partner, meaning that each participant in each conversational pair alternates between the role of 'describer' (main speaker) and 'drawer' (secondary speaker).

In total, each participant describes six objects and draws six other. Each participant's audio signal is captured exclusively by the corresponding AKG C520 microphone. The audio signal passes through an RME FIREFACE UC and is then transmitted back to the corresponding XF 105 HD camera.

2.2. Participants

The task is completed by 10 native speakers of American English, either students at the University of Aix-Marseille or residents in Aix-en-Provence, in France.

Among them, 6 are women and 4 are men. Their ages range from 21 to 47 years old (M=31.3). In addition to English, all participants speak a second language (e.g., French or Spanish), however none has been exposed to the second language before the age of 4, and they exclusively speak English at home.

2.3. Data set

The gathered data thus comprises 5 task-oriented dialogues, 3 female conversations (F(1), F(2) and (F(3)) and 2 male ones (M(4) and M(5)). Each conversation is made of 12 trials, 6 where Speaker 1 is the main speaker, and 6 where Speaker 2 is.

In the present study we take into account only the speech by the primary speakers (i.e., Speaker 1 and Speaker 2 alternatively). The data thus amounts to $\sim$6.3h of recordings, i.e., 76,950 word tokens. Among them, 647 are unconnected filled pauses[1].

[1]The dataset is part of an ongoing project and will be made publicly available upon project completion.

2.4. Methodology

The filled pauses are manually annotated in the recordings using Praat (version 6.2.14 [26]). The starting and ending points of the filled pause are based on vocal fold vibrations, and the boundary between the vowel and the nasal part are determined based on where the first formant changes significantly.

It may happen that the filled pause is produced in the continuation of words like "a", "like" or "then". In these cases, the filled pause may actually refer to a case of lengthening of the last phone of the word rather that to an actual filled pause. To avoid grouping together different manifestations of hesitation, we consider here only cases where the filled pause is uttered disconnectedly from the preceding and following speech.

Finally, the durations of the vocalic and nasal parts of each filled pause have been automatically extracted using Praat [26]. The vowel-to-nasal ratio is calculated as the duration of the vowel divided by the duration of the nasal. To avoid dividing the vowel in "uh" and the nasal in "mh" by zero, we assign a value of 1 to the duration of the nasal component in "uh" and a value of 1 to the duration of the vowel component in "mh".

3. Results

3.1. Overall use of *uh*, *um* and *mh*

Fig. 1 shows that, overall, dyads have very differing uses of filled pauses. The first (female) dyad displays a massive preference for *um*, the third (female) and fifth (male) dyads also display such a preference, although less overwhelmingly, while dyads 2 (female) and 4 (male) show more variation. One notable point is that speakers tend to favor the vowel-only filled pause *uh* or the vowel-nasal one *um*, but none favors the nasal-only filled pause. Most of them (7/10) do not use it at all.

It is also worth noting that, in Conversation 2, Speaker 2 is the one introducing *uh*, but Speaker 1 is the one using it the most (100% of her filled pauses in Trials 5 and 7), and although the overall pattern suggests an increase in the use of *uh*, the overall impression is that it is not due to alignment, at least not at the local level (from one trial to the next), as indicated by the discrepancy of the uses between Trials 4 (45.45%), 5 (100%, δ=55.55%), 6 (27.27%, δ=$-$72.73%) and 7 (100%, δ=72.73%).

In Conversation 3, there is also variation in the patterns of the speakers, with Speaker 1 realizing 0% *uh*s in 3 out of 6 trials, while Speaker 2's use of *uh* ranges from 12.50% to 50.00% of her filled pauses. The pattern does not indicate that the two

Figure 2: *Values for vowel-to-nasal ratio for each speaker in each trial in each conversation. The y-axis represents the values of the ratio, and the x-axis indicates the number of the trial. Grey represents Speaker 1 (S1), yellow represents Speaker 2 (S2). A bold line connects the mean values for each speaker in each trial. A bold dashed line indicates that the first of the two connected trials lacks data. The statistical values are the outcomes of a Mann-Whitney U Test of vowel-to nasal ratio for each comparative pair in dialogues: The fine black line connects two adjacent trials of S1 followed by S2; The fine purple line connects two adjacent trials of S2 followed by S1; The fine blue line connects the first and last trials of the same speaker; The asterisks indicate the level of statistical significance, with * indicating $p \leq 0.05$, ** indicating $p \leq 0.01$, *** indicating $p \leq 0.001$, and 'NA' indicating that data are not available.*

speakers converge in their choice of filled pauses.

In Conversation 5, the *um*s clearly dominate the choice of the filled pause's form, but it is also notable that, although speakers use little *uh*s, Speaker 1 seems to use much less than Speaker 2, especially at the end of the dialogue, indicating, if anything, divergence rather than convergence.

The most interesting behavior is to be found in Conversation 4, where there is a clear difference between the usages of the two speakers. The rates of *uh*s are consistently almost twice higher for Speaker 2 than for Speaker 1, throughout the conversation (from δ=26.19% between trials 8 and 9 to δ=75.00% between trials 2 and 3). This dialogue displays the clearer pattern of two speakers having each a specific pattern and never showing any kind of alignment, either locally (with the preceding trial) or in general (across the 12 trials).

3.2. Dialogical alignment of the vowel-to-nasal ratio

In this subsection, we present more granular results regarding not only the form of the filled pause (*uh*, *um* or *mh*), but the

exact ratio between the duration of the vowel and the duration of the nasal.

Fig. 2 shows both the intra-speaker (size of the boxes) and inter-speaker (Mann-Whitney U test) variation in the vowel-to-nasal ratio for each speaker, trial by trial.

In the first (F(1)) as well as in the second (F(2)) conversations, Speaker 2 only significantly diverges from Speaker 1 in Trial 6 compared to Trial 5 (U=6, p=0.008 for F(1) and U=109, $p<0.001$ for F(2)) and Speaker 1 diverges from Speaker 2 in the following Trial 7 compared to Trial 6 (U=71, p=0.025 for F(1) and U=1, p=0.003 for F(2)). It is notable, though, that in F(1), the first speaker diverges from herself between her first and last trial (U=29, p=0.043). This may indicate that she adapts her natural production to actively converge with her interlocutor. Unfortunately, Speaker 1 in F(2) does not produce any filled pause in Trial 1, and thus it is impossible to calculate whether she too changes her usage to align with her interlocutor, although impressionistic observations of Fig. F(2) indicate that she may.

In the third conversation F(3), the two speakers remain aligned throughout the conversation and none of them shows a statistically significant difference between their first and last trials, indicating either that they may have been naturally aligned, or that Speaker 2 aligned her productions immediately after Trial 1 and remained aligned with her interlocutor throughout the conversation without shifting back to her natural way of speaking.

The fourth conversation (M(4)), shows that, as expected, the two speakers often diverge in their vowel-to-nasal ratio, especially at the beginning and at the end of the conversation. In Trial 2, Speaker 2 diverges significantly from Speaker 1's first trial (U=19, p=0.003); Then in Trial 3, Speaker 1 diverges from Speaker 2 (U=36, p=0.004); And in Trial 4, Speaker 2 diverges again from the preceding trial by Speaker 1 (U=19, p=0.012). The two speakers then seem convergent for 6 trials (half the conversation), despite the fact that they show very diverging patterns in Fig. 1: This may be due to the fact that, although Speaker 2 uses less nasal filled pauses (*um* or *mh*) than Speaker 1, he compensates with very long nasals in the few *um*s he utters, or conversely that Speaker 1 realizes his few *uh*s with very long vowels, thus compensating for the fact that he produces less vowel-only filled pauses. Whether such compensation patterns actually arise however will remain a topic for future studies. Finally, towards the end of the conversation, the two speakers diverge again: In Trial 11, Speaker 2 diverges from Speaker 1's Trial 10 (U=211.5, p=0.004) and in the next Trial 12, Speaker 1 diverges from the preceding trial (U=18, p=0.023). This may indicate that, if the two speakers converged after four divergent trials, remaining convergent demands efforts that last only temporarily: More data from more (and longer) conversations would be needed to test this hypothesis. As is, it can be noted that none of the speakers diverge from themselves between their first and last trials, indicating that they have an internally similar use of filled pauses at the beginning and at the end of the conversation.

Finally, regarding the fifth conversation (M(5)), it can be seen that the two speakers are divergent at the beginning of the conversation (U=150, p=0.038 for Trial 2 compared to Trial 1; U=92, p=0.031 for Trial 3 compared to Trial 2; and U=240, p=0.003 for Trial 4 compared to Trial 3), but soon both display too few filled pauses for results to be significant. Again, none of the speakers diverge from themselves between their first and last trials, indicating that neither changed their production habits to align with their interlocutor.

4. Discussion

In this study, we investigate whether the various functions of filled pauses are hearer-oriented, i.e., deliberately used by the speaker to convey meaning, or speaker-oriented, i.e., mere artifacts of thought in the making, with no intention from the speaker. To explore this question, we monitored the use of autonomous filled pauses by 10 speakers in 5 same-gender dyads.

Our results indicate that, first, the choice of the form of the filled pause (either *uh*, *um* or *mh*) seems highly idiosyncratic, some speakers favoring *uh*, others *um*, but none favoring *mh* over all others. Second, the comparison of the vowel-to-nasal ratio in the filled pauses of speakers shows that speakers either converge or diverge naturally, but that none, except one, change their use of filled pauses to adapt to the interlocutor. From second speakers, this may be due to the fact that they adapt (either positively, thus causing convergence, or negatively, thus enhancing divergence) right after the first trial. From the first speaker however, this shows that only one participant out of five adapts her speech to maintain or enhance convergence with her interlocutor. These results indicate that filled pauses do not seem prone to conversational convergence, thus advocating in favor of the analysis of filled pauses as speaker-oriented rather than hearer-oriented, at least in task-oriented speech.

In further steps of our investigations, we plan on measuring the phonetic characteristics of the vowels and nasals in the filled pause to investigate whether the lack of lexical convergence is coupled with a lack of phonetic convergence. We also intend to gather more data from more participants, to test whether the observed pattern is confirmed. This would also allow us to test more fine-grained hypotheses, such as whether participants modulate the duration of their vowels or nasals to compensate their differing use of filled pauses. Moreover, it would be interesting to conduct a similar data collection with longer conversations to test whether convergence, if or when it happens, is indeed only temporary, as hinted by the results on one of our dyads, and if so, how long it lasts.

It is also possible that we do not find convergence in the form of the filled pause because each form actually has a different meaning, as hinted by [2], or because various filled pauses are actually on various stages on the "cline of wordhood" [1]. In these cases, longer conversations may also allow us to explore whether the various filled pauses indeed announce different types of delay in speech, or have a different lexical status, and with more tokens, it could also be possible to investigate the convergence within vowel-only, vowel-nasal or nasal-only filled pauses.

Finally, a possible evolution of this work would compare speech in task-oriented dialogues, which induces a higher cognitive load, and natural speech in informal settings, to see whether task or cognitive load affect the use and possible convergence of filled pauses.

5. Acknowledgements

This work was supported by an F.R.S.-FNRS research grant to M. Hutin (project PPaDisM), and by the Marie Skłodowska-Curie Actions program (MSCA-ITN-ETN) COBRA "Conversational Brains" (grant number: 859588) and U of Louvain Special Research Funds FSR Project 2023.

The Interspeech 2024 organisers would like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

6. References

[1] G. Tottie, "From pause to word: uh, um and er in written American English," *English Language and Linguistics*, vol. 23, no. 1, p. 105–130, 2019.

[2] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027702000173

[3] M. Kirjavainen, L. Crible, and K. Beeching, "Can filled pauses be represented as linguistic items? Investigating the effect of exposure on the perception and production of um," *Language and Speech*, vol. 65, no. 2, pp. 263–289, 2022, pMID: 34028288. [Online]. Available: https://doi.org/10.1177/00238309211011201

[4] I. Rehbein, "Filled pauses in user-generated content are words with extra-propositional meaning," in *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, E. Blanco, R. Morante, and C. Sporleder, Eds. Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 12–21. [Online]. Available: https://aclanthology.org/W15-1302

[5] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, vol. 30, no. 4, pp. 485–496, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378216698000149

[6] L. Kosmala and L. Crible, "The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence," *Language and Speech*, vol. 65, no. 1, pp. 216–239, 2022, pMID: 33977820. [Online]. Available: https://doi.org/10.1177/00238309211010862

[7] M. Corley and O. W. Stewart, "Hesitation disfluencies in spontaneous speech: The meaning of um," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 589–602, 2008. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2008.00068.x

[8] Štefan Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378216611001469

[9] B. Jiang, E. Ekstedt, and G. Skantze, "What makes a good pause? Investigating the turn-holding effects of fillers," 2023.

[10] A.-B. Stenström, "Lexical items peculiar to spoken discourse," *The London-Lund Corpus of Spoken English. Description and Research*, 1990.

[11] E. Gilmartin, C. Vogel, and N. Campbell, "Disfluency in chat and chunk phases of multiparty casual talk," in *Proceedings of the Disfluency in Spontaneous Speech Workshop (DiSS 2017)*, 2017, pp. 25–28.

[12] S. Brennan and M. Williams, "The Feeling of Anothers Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers," *Journal of Memory and Language*, vol. 34, no. 3, pp. 383–398, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749596X85710170

[13] H. Giles, N. Coupland, and J. Coupland, *Accommodation theory: Communication, context, and consequence*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, 1991, p. 1–68.

[14] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, pp. 1482–1493, 1996.

[15] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of personality and social psychology*, vol. 76, p. 893–910, 1999. [Online]. Available: https://doi.org/10.1037//0022-3514.76.6.893

[16] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, A. Zaenen and A. van den Bosch, Eds. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 808–815. [Online]. Available: https://aclanthology.org/P07-1102

[17] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, ser. HLT-Short '08. USA: Association for Computational Linguistics, 2008, p. 169–172.

[18] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, perception psychophysics*, vol. 79, p. 637–659, 2017. [Online]. Available: https://doi.org/10.3758/s13414-016-1226-0

[19] Beňuš, R. Levitan, and J. Hirschberg, "Entrainment in spontaneous speech: The case of filled pauses in supreme court hearings," in *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 793–797.

[20] Štefan Beňuš, A. Gravano, R. Levitan, S. I. Levitan, L. Willson, and J. Hirschberg, "Entrainment, dominance and alliance in supreme court hearings," *Knowledge-Based Systems*, vol. 71, pp. 3–14, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705114002184

[21] V. Silber-Varod, D. Amit, and A. Lerner, "Tracing changes over the course of the conversation: A case study on filled pauses rates," in *Proc. Speech Prosody 2020*, 2020, pp. 754–758.

[22] M. Belz, M. Müller, and C. Mooshammer, "How consistent are non-native speakers in their usage of filler particles when talking to native speakers?" in *Proc. Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, 2023, pp. 53–57.

[23] M. Rasenberg, W. Pouw, A. Özyürek, and M. Dingemanse, "The multimodal nature of communicative efficiency in social interaction," *Scientific Reports*, vol. 12, no. 19111, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-22883-w

[24] 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. [Online]. Available: https://doi.org/10.1177/026192702237958

[25] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, E. Fosler-Lussier, E. Riloff, and S. Bangalore, Eds. Montréal, Canada: Association for Computational Linguistics, Jun. 2012, pp. 11–19. [Online]. Available: https://aclanthology.org/N12-1002

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 6.2.06," 1992–2022. [Online]. Available: https://www.praat.org