ISSP 2024 - 13th International Seminar on Speech Production
13-17 May 2024, Autrans, France

Examining the Link between the Perception and Production of Phonetic Convergence of Laughter in Interaction

Marin Schröer, Bogdan Ludusan

Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany
{marin.schroeer, bogdan.ludusan}@uni-bielefeld.de

Abstract

Convergence, i.e. the process of two people adjusting their behaviour to become more similar to each other, has been found in most aspects of human interaction by now. However, most of the studies investigating convergence have so far considered mostly linguistic aspects while taking into account only production or perception rather than both. In this paper, we expand on previous work to examine paralinguistic phenomena, namely laughter, by integrating an analysis of differences between several acoustic cues extracted from laughter in spontaneous interaction with a perceptual experiment in order to determine their perceptual relevance.

Keywords: laughter, convergence, speech production, speech perception, conversational speech

1. Introduction

In recent years, many studies have investigated the behaviour of speakers in interaction and have found evidence of them adapting their communicative habits to become more similar to each other in a process known as convergence (Pardo 2013). These studies have shown convergence for syntactical (Branigan, Pickering, and Cleland 2000) and lexical (Brennan and Clark 1996; Nenkova, Gravano, and Hirschberg 2008) aspects of speech, among others. Segmental and prosodic features have also been examined for convergence (Pardo 2006; Pardo, Urmanche, et al. 2017; Levitan and Hirschberg 2011).

Some studies have also investigated paralinguistic items, such as pauses (Edlund, Heldner, and Hirschberg 2009), gestures (Holler and Wilkin 2011) and for conversational phenomena, such as laughter (Truong and Trouvain 2012a; Truong and Trouvain 2012b; Ludusan and P. Wagner 2019; Ludusan and P. Wagner 2022). Ludusan, Schröer, and P. Wagner (2022) previously showed, that speakers exhibit convergence of laughter regarding their vowel quality, as measured by comparing distances of F1 and F2 in between speakers at the start and end of the conversation.

Most phonetic studies of convergence can be grouped into two sets: One set studies convergence by calculating and comparing the distances in a given set of acoustic cues, such as spectral moments, formant values, etc. (Levitan and Hirschberg 2011; Pardo, Urmanche, et al. 2017; Gessinger et al. 2017; Ludusan, Schröer, and P. Wagner 2022) The other set looks at the phenomenon from a perceptual perspective by having listeners rate the similarity between audio stimuli (Pardo 2006; Pardo 2013; Babel 2012; Namy, Nygaard, and Sauerteig 2002).

Both of the aforementioned approaches provide important insights into the process of convergence, by either pinpointing acoustical cues that undergo statistical changes throughout the interaction or by showing that listeners perceive convergence to several different phonetic aspects.

Putting both approaches together should thus provide a more complete account of convergence by determining both the amount of change in the investigates acoustic measures, as well as their perceptual relevance. This holistic approach has started to become more popular in recent studies such as Abel and Babel (2017), M. Wagner et al. (2021), Lewandowski and Nygaard (2018), and Pardo, Jordan, et al. (2013).

In their study, Abel and Babel (2017) had pairs of participants perform a cooperative task. Afterwards, a different group of participants had to listen to and rate whether the pair had converged or not. While they found evidence for convergence both in the acoustic features extracted from the dyad itself as well as in the perceptual ratings, they could not establish a clear correlation between the two. They argued that, while the listeners had access to global changes across all acoustic dimensions, the acoustic difference algorithms they employed to analyse the data had access to only one dimension at a time. This would then suggest that not one cue might be important, but rather that perceptual information about convergence comes from an interplay between multiple cues.

Both M. Wagner et al. (2021) and Lewandowski and Nygaard (2018) integrated this in their studies investigating convergence towards both native and non-native accents. Each of these studies had participants produce a set of items and then shadow a different speaker going through the same list. The resulting productions were compared and the distance between several acoustic cues was calculated. They further presented both speakers' productions to a set of listeners who had to judge whether convergence took place (similar to Abel and Babel (2017)). Wagner et al. found vowel duration, speech rate and f0 to be the most converged-to and perceptually salient dimensions. The findings of Lewandowski and Nygaard further corroborate this, with one exception. In their study, the f0 measure only correlated with perceived convergence for non-native speakers converging to native speakers, while vowel quality was correlated with native to non-native convergence.

Pardo, Jordan, et al. (2013) had the same general setup as the previous studies, with a shadowing as well as an AXB perceptual task. While they also investigated lexical factors alongside acoustic ones, they could not establish a link between the lexical and perceived convergence. For the acoustic measures, however, they found vowel spectra, duration and f0 to correlate with perceived convergence, thus being in line with the findings of the two aforementioned studies

Here, we extend these findings by integrating production and perception aspects of laughter convergence in spontaneous interaction by testing the perceptual significance of several acoustic cues shown to differ within conversation.

10.21437/issp.2024-20

2. Methods

2.1. Stimuli

Our stimuli were taken from the DUEL corpus (Hough et al. 2016), consisting of spontaneous dyadic interactions, in which speakers have to either role-play as a border control agent, discuss furnishing an apartment or write an embarrassing film script. In total, the German portion of the corpus contains 19 such dyads. One of the dyads was shown to have both speakers converge in Ludusan, Schröer, and P. Wagner (2022). As they produced far more instances of laughter in the film script condition than in the others, laughter from that condition was considered.

We segmented the laughter instances into syllables and vocalic/consonantal parts in accordance with Trouvain (2003) using VocalToolKit (Corretge 2022), a Praat (Boersma 2002) plugin. and corrected the annotations manually. Afterwards, we extracted vowels from the first and last third of the interaction, and we selected vowels which were egressive, artifact free, and non-fricated to be used as stimuli (n = 38). All but one of the extracted vowels were more than 100ms in length (on average 184ms). All steps were performed using Praat (Boersma 2002).

2.2. The perception experiment

Using the multiple forced choice experiment function provided by Praat (Boersma 2002), 23 participants were presented the stimuli in an AXB paradigm (Goldinger 1998; Pardo, Jordan, et al. 2013), modified so that the X was presented after A as well as after B, essentially resulting in two pairs (AX and BX). The participants (all German-speaking university students) were then asked to rate which pair was more similar. They were able to listen to each stimulus twice and could not go back in order to change their choice. At the beginning of each trial, there was a pause of 0.5s, so the trial would not start immediately after the previous answer was given. Furthermore, there was a pause of 300ms within and of 600ms between pairs in order to separate the stimuli and pairs clearly. The instructions on what the participants should listen for were left intentionally vague (only similarity was stated) in order to get as unbiased of a response as possible.

The stimuli had A taken from either the first or last third from one speaker, B taken from the last third of the same speaker, and X from the first third of the opposite speaker. This was done in order to assess whether the speaker (S1 or S2) had converged to their interlocutor's baseline or not. In total, participants were presented 80 tokens of the structure described above.

2.3. Analysis

We originally extracted the mean F1 and F2 values for every vowel used in the experiment in order to compute the Euclidean distance as described by Equation 1 and determine whether there was a change in vowel quality from the first to the last third for each speaker.

$$d_{f1f2}(A,B) = \sqrt{(F1_A - F1_B)^2 + (F2_A - F2_B)^2} \quad (1)$$

The formant values were normalised using the PhonR package (McCloy 2016) in R (R Core Team 2020), in which all other analyses were also carried out. We also examined, whether the convergence shown in Ludusan, Schröer, and P. Wagner (2022), for the dyad from which the vowels were extracted, could still be found in the stimuli used in this experiment by using a Wilcoxon test to test the difference in vowel quality between on speakers' baseline production and the other speaker at the start and end of the conversation.

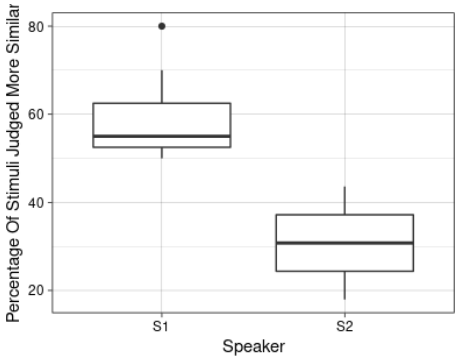

Figure 1: Percentage of how often each speaker in the last third of the interaction was perceived to be more similar to their interlocutors baseline than at the beginning of the conversation. A value significantly higher than 50% represents convergence, a value significantly lower than 50% shows divergence

However, since the preliminary results of the perceptual experiment and the acoustic analysis considering the F1F2 distances did not agree, we further extracted the fundamental frequency (*f0*), the root-mean-square energy of the signal (*en*), the duration of the vowel (*dur*) and the cepstral peak prominence (*cpp*, a measure of voice quality, with lower values of this measure indicating a more breathy phonation). For these features, we calculated absolute distances analogous to the example formula for f0 in Equation 2. We furthermore normalised all acoustic cues by subtracting their mean and dividing by two standard deviations (Gelman 2008).

$$d_{f0}(A,B) = |f0_A - f0_B|. \quad (2)$$

For the acoustic analysis Wilcoxon signed rank tests were used to test convergence, evaluating whether the distances between the speakers at the start of the conversation were greater than the distances between one speaker at the start (1st third) and one speaker at the end (last third) of the conversation. The distance between one speaker at the start and one at the end being smaller than the other indicates some degree of convergence of one speaker towards the baseline of the other, whereas the opposite would suggest divergence.

We fitted generalised mixed effects models in order to determine a link between perception and production, pooling the data from both speakers. The dependent variable was assigned a value of 1 if the raters had picked a pair consisting of vowels from different thirds as more similar (i.e. the convergence case) and was assigned a value of 0 in the divergence case.

All the examined cues were fixed factors in the model, and the rater was chosen as a random intercept. We employed model reduction by first building the largest possible model and then reducing it down step by step, as long as each step reduced the Akaike Information Criterion value of the model.

	dur		en		f0		cpp		F1F2	
	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2	Speaker 1	Speaker 2
Mean	0.0126	-0.0071	-0.2739	1.6496	-7.1395	-88.5311	0.2015	1.6547	43.569	24.7816
SD	0.0226	0.0211	5.7367	4.0579	27.9043	65.096	2.5	2.3238	98.8131	97.84

Table 1: Means and standard deviation for the distance in difference value for each acoustic cue by which the speaker was being compared. Positive numbers indicate larger difference between the baselines, i.e. convergence.

3. Results

The Wilcoxon test performed prior to the perception study showed both speakers converging to their interlocutors' baseline over the course of the conversation in the F1F2 measure when considering all possible combinations of vowels in the stimulus set (S1 converging to S2 $p = 0.034$ and S2 to S1 $p = 0.0021$).

The participants in the perceptual experiment rated the stimuli from speaker S1 as showing convergence (58.9% convergence answer, $p = 2.0e^{-4}$), while those for speaker S2 as showing divergence (30%, $p = 2.8e^{-5}$)(s. fig. 1).

We then examined whether there are differences between the acoustic distances of the two pairs of stimuli. For speaker S1, there were significant differences for *dur* ($p = 0.003$) and for the *f1f2* distance ($p = 0.008$), both indicating convergence. For speaker S2, the acoustic analysis showed a more complex picture, with the values for *en* ($p = 6.8e^{-4}$) and *cpp* ($p = 1.9e^{-4}$) showing convergence, while those for *dur* ($p = 0.033$) and *f0* ($p = 1.0e^{-7}$) showing divergence. The *f1f2* distance showed a trend towards convergence, although it was not significant ($p = 0.087$).

The model fitted to study the relation between raters' perception and stimuli acoustic distances revealed a significant main effect for *dur* ($\beta = 0.688, p = 1.2e^{-7}$), *f0* ($\beta = 1.516, p < 2.2e^{-16}$), *cpp* ($\beta = 0.424, p = 1.6e^{-3}$) and *f1f2* ($\beta = 0.933, p = 4.8e^{-9}$). There was no significant main effect for *en* ($\beta = 0.255, p = 0.060$). One two-way interaction (*en:f1f2*), four three-way interactions (*dur:en:cpp*, *dur:en:f1f2*, *en:f0:f1f2* and *f0:cpp:f1f2*), all but one of the four-way interactions (*dur:f0:cpp:f1f2*), and the five-way interaction were found to be significant.

4. Discussion

When taking into account instances between every combination of the data in the stimulus set, we found significant convergence for both speaker's vowel quality. This is in contrast to the results of the acoustical analysis performed only on the distances between the combination of token included in the perceptual experiment, in which the vowel quality measure was only significant for S1. This is due to the fact not every possible combination of stimuli was used, as this would have made the experiment exceedingly long. Thus, it is possible that vowel quality could have a more important role than shown by our study. Investigating the role of several acoustic cues on perception, our study revealed that, for each of the examined cues, having a higher distance at the beginning of the conversation, compared to at its end, increased the odds of the raters to perceive convergence. While these findings may suggest a straightforward link between production and perception, in the case of phonetic convergence of laughter, a high number of interactions were found to be significant and many of them had a negative estimate. Thus, these results point towards a more complex picture, in which also the interactions between several acoustic cues need to be taken into account. Moreover, based on the fitted model we are able to draw conclusions on the importance of each acoustic cue for the perception of convergence, with the fundamental frequency of the voice playing the most important role, followed by vowel quality (as given by the Euclidean distance between the first two formant values), duration, voice quality (breathiness – as given by *cpp*) and finally, speech intensity.

The different importance ranking of the examined acoustic cues may explain the more complex case we encountered for speaker S2, where the energy of the signal and the cepstral peak prominence measures indicated convergence, while f0 and duration indicated divergence. Considering that the cue that played the most important role in perception f0 showed divergence, it may not be surprising that the raters judged that speaker as diverging. These results do not fully align with those of a previous acoustic study of phonetic convergence (Ludusan, Schröer, and P. Wagner 2022), in which both speakers of this pair showed convergence. The difference is most likely due to the small subset of stimuli that were included in the current study, which might not be representative of the full set considered in the previous work (which analysed a set one order of magnitude larger). In particular, some of the stimuli employed here had high f0 values, diverging from other stimuli, counteracting the convergence (or convergence trends) seen with respect to other measures.

While these results are based on a rather limited data set, they do align well with those of previous studies, that looked at both the production and perception aspects of convergence. As in Pardo, Jordan, et al. (2013) and M. Wagner et al. (2021) and Lewandowski and Nygaard (2018) duration, f0 and to a lesser extent the vowel quality seemed to play the most important role in determining whether listeners perceived convergence or not. Furthermore, it seems that as suggested in Pardo, Urmanche, et al. (2017) and Abel and Babel (2017) taking different cues and the interactions between them into account improves a model's ability to predict whether convergence/divergence are perceived. This further suggests that listeners not only have access to, but also integrate, several cues simultaneously in order to judge convergence. Our results further extend those of these studies, by showing that they hold for non-verbal phenomena as well.

In the future, we intend to address the limitations of this study by extending the analysis to a larger, more diverse dataset. It might further be interesting to try to incorporate even more acoustic cues. Lastly one may examine whether individual listeners vary in how they weigh acoustic cues, as well as whether different speakers tend to converge more strongly to different acoustic cues than others, as similar effects have been found for accent/speaker groups (Lewandowski and Nygaard 2018; M. Wagner et al. 2021)

5. Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 461442180.

6. References

Abel, Jennifer and Molly Babel (2017). "Cognitive Load Reduces Perceived Linguistic Convergence Between Dyads". In: *Language and Speech* 60.3, pp. 479–502. DOI: 10.1177/0023830916665652.

Babel, Molly (2012). "Evidence for phonetic and social selectivity in spontaneous phonetic imitation". In: *J. Phonetics* 40, pp. 177–189. DOI: https://doi.org/10.1016/j.wocn.2011.09.001.

Boersma, Paul (2002). "Praat, a system for doing phonetics by computer". In: *Glot International* 5, pp. 341–345.

Branigan, Holly, Martin Pickering, and Alexandra Cleland (2000). "Syntactic co-ordination in dialogue". In: *Cognition* 75, B13–25. DOI: 10.1016/S0010-0277(99)00081-5.

Brennan, Susan and Herbert Clark (1996). "Conceptual Pacts and Lexical Choice in Conversation". In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, pp. 1482–1493. DOI: 10.1037/0278-7393.22.6.1482.

Corretge, Ramon (2022). *Praat Vocal Toolkit*. https://www.praatvocaltoolkit.com.

Edlund, Jens, Mattias Heldner, and Julia Hirschberg (2009). "Pause and gap length in face-to-face interaction". In: pp. 2779–2782. DOI: 10.21437/Interspeech.2009-710.

Gelman, Andrew (2008). "Scaling Regression Inputs by Dividing by Two Standard Deviations". In: *Statistics in medicine* 27, pp. 2865–73. DOI: 10.1002/sim.3107.

Gessinger, Iona, Eran Raveh, Sébastien Le Maguer, Bernd Möbius, and Ingmar Steiner (2017). "Shadowing Synthesized Speech-Segmental Analysis of Phonetic Convergence." In: *Interspeech*, pp. 3797–3801. DOI: 10.21437/Interspeech.2017-1433.

Goldinger, Stephen (1998). "Echoes of Echoes? An Episodic Theory of Lexical Access". In: *Psychological review* 105, pp. 251–79. DOI: 10.1037/0033-295X.105.2.251.

Holler, Judith and Katie Wilkin (2011). "Co-Speech Gesture Mimicry in the Process of Collaborative Referring During Face-to-Face Dialogue". In: *Journal of Nonverbal Behavior* 35.2, pp. 133–153. DOI: 10.1007/s10919-011-0105-6.

Hough, Julian, Ye Tian, Laura de Ruiter, Simon Betz, Spyros Kousidis, David Schlangen, and Jonathan Ginzburg (2016). "DUEL: A multilingual multimodal dialogue corpus for disfluency, exclamations and laughter". In: *Proc. of LREC*, pp. 1784–1788.

Levitan, Rivka and Julia Hirschberg (2011). "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions". In: *Proc. Interspeech 2011*, pp. 3081–3084. DOI: 10.21437/Interspeech.2011-771.

Lewandowski, Eva and Lynne Nygaard (2018). "Vocal alignment to native and non-native speakers of English". In: *The Journal of the Acoustical Society of America* 144, pp. 620–633. DOI: 10.1121/1.5038567.

Ludusan, Bogdan, Marin Schröer, and Petra Wagner (2022). "Investigating phonetic convergence of laughter in conversation". In: *Proc. of INTERSPEECH*, pp. 1332–1336. DOI: 10.21437/Interspeech.2022-10332.

Ludusan, Bogdan and Petra Wagner (2019). "Laughter Dynamics in Dyadic Conversations". In: DOI: 10.21437/Interspeech.2019-1733.

— (2022). "Laughter entrainment in dyadic interactions: Temporal distribution and form". In: *Speech Communication* 136, pp. 42–52. DOI: doi.org/10.1016/j.specom.2021.11.001.

McCloy, Daniel (2016). *phonR: Tools for phoneticians and phonologists. R package version 1.0-7*. Online: https://cran.r-project.org/web/packages/phonR/phonR.pdf.

Namy, Laura L., Lynne Nygaard, and Denise Sauerteig (2002). "Gender Differences in Vocal Accommodation:: The Role of Perception". In: *Journal of Language and Social Psychology* 21.4, pp. 422–432. DOI: 10.1177/026192702237958.

Nenkova, Ani, Agustín Gravano, and Julia Hirschberg (2008). "High Frequency Word Entrainment in Spoken Dialogue." In: pp. 169–172. DOI: 10.3115/1557690.1557737.

Pardo, Jennifer (2006). "On phonetic convergence during conversation". In: *The Journal of the Acoustical Society of America* 119, pp. 2382–93. DOI: 10.1121/1.2178720.

— (2013). "Measuring phonetic convergence in speech production". In: *Frontiers in Psychology* 4, p. 559. DOI: 10.3389/fpsyg.2013.00559.

Pardo, Jennifer, Kelly Jordan, Rolliene Mallari, Caitlin Scanlon, and Eva Lewandowski (2013). "Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures". In: *Journal of Memory and Language* 69.3, pp. 183–195. DOI: https://doi.org/10.1016/j.jml.2013.06.002.

Pardo, Jennifer, Adelya Urmanche, Sherilyn Wilman, and Jaclyn Wiener (2017). "Phonetic convergence across multiple measures and model talkers". In: *Attention, Perception, & Psychophysics* 79, pp. 637–659. DOI: 10.3758/s13414-016-1226-0.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Trouvain, Jürgen (2003). "Segmenting phonetic units in laughter". In: *Proc. of ICPhS*, pp. 2793–2796.

Truong, Khiet and Jürgen Trouvain (2012a). "Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis". In: *Proceedings of 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)*, pp. 20–24.

— (2012b). "On the acoustics of overlapping laughter in conversational speech". In: *Proc. Interspeech 2012*, pp. 851–854. DOI: 10.21437/Interspeech.2012-192.

Wagner, Mónica, Mirjam Broersma, James McQueen, Sara Dhaene, and Kristin Lemhöfer (2021). "Phonetic convergence to non-native speech: Acoustic and perceptual evidence". In: *Journal of Phonetics* 88, p. 101076. DOI: 10.1016/j.wocn.2021.101076.