A novel paradigm to investigate phonetic convergence in interaction

Martijn Wieling[1,2], Mark Tiede[2], Teja Rebernik[1], Lisanne de Jong[1], Anouck Braggaar[1], Martijn Bartelds[1], Masha Medvedeva[1], Penny Heisterkamp[1], Tom Freire Offrede[3], Hedwig Sekeres[1], Anna Pot[1], Mara van der Ploeg[1], Karin Volkers[4], Gregory Mills[1]

[1]*University of Groningen, the Netherlands*
[2]*Haskins Laboratories, United States of America*
[3]*Humboldt Universität zu Berlin, Germany*
[4]*Philadelphia Care Foundation, the Netherlands*

m.b.wieling@rug.nl, tiede@haskins.yale.edu, t.rebernik@rug.nl, lisannemereldejong@gmail.com, a.r.y.braggaar@student.rug.nl, m.bartelds@rug.nl, m.medvedeva@rug.nl, p.g.heisterkamp@student.rug.nl, tom.offrede@gmail.com, h.g.sekeres@student.rug.nl, a.pot@rug.nl, a.m.van.der.ploeg@rug.nl, k.volkers@philadelphia.nl, g.mills@rug.nl

Abstract

We introduce a phonetic variant of the procedural coordination task. In this task, pairs of participants are presented with the recurrent coordination problem of jointly producing sequences of congruent and complementary vocalizations. Our results, on the basis of analyzing over 75 pairs, show that after only 15 minutes of interaction, task-specific phonetic convergence can be observed. In contrast to, for example, the shadowing task, our task not only offers a high level of control over the stimuli, but also a high level of meaningful interaction. This paper serves as a detailed description of the paradigm, not only focusing on the phonetic convergence aspect, but also illustrating several communicative strategies which emerged when pairs participated in the task.

Keywords: phonetic convergence, formants, procedural coordination task

1. Introduction

Many studies have shown that when two people converse with each other, they progressively adapt their linguistic resources to those of their partner (Pickering & Garrod, 2013). One form of adaptation is phonetic convergence, which is highly context-sensitive, variable and driven by the interactional goals of the participants (see e.g., Pardo *et al.*, 2017). There are several approaches to investigate phonetic convergence. One approach is to use speech shadowing tasks (Pardo *et al.*, 2017) in which a participant provides the pronunciation of the same utterance before and after shadowing (i.e., repeating the utterances of) a model speaker. When the participant's pronunciation after the shadowing task is closer to that of the model speaker than before, phonetic convergence has taken place. Another approach is investigating phonetic convergence in conversational interaction (Pardo, 2007). Here approaches such as using a spot-the-differences task, or a maze game can be used to guide the conversation. Nevertheless, a much lower degree of control is possible in these conversational settings than in approaches such as using a speech shadowing task.

Consequently, experimental approaches to phonetic convergence are faced with a methodological trade-off between experimental control and validity. On the one hand, tasks which use shadowing allow high levels of control but block many interactive mechanisms that underpin convergence. On the other hand, more spontaneous tasks allow high levels of interaction, but remove the tight control over stimuli afforded by shadowing. To side-step this trade-off, we present data from a phonetic variant of the procedural coordination task (Mills, 2011) which presents pairs of participants with a task in which they need to jointly produce collaborative sequences of simple vocalizations.

2. Paradigm

The setup[1] consists of a two-player music game in which in each round one player has the role of director, and the other the role of follower. The director sees the melody which needs to be played from bottom to top (see Figure 1, left; each line represents one distinct note, the color indicates the supposed player: self = red, other = blue), whereas the follower only sees three empty lines (see Figure 1, right).

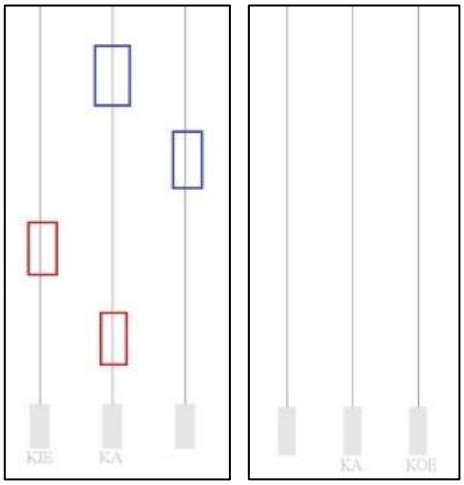

Figure 1: *Game layout for both participants (left: director view; right: follower view).*

The notes are played through pronouncing different consonant-vowel (CV) sequences. The leftmost note is played by pronouncing /ki/, the middle by pronouncing /ka/, and the rightmost by pronouncing /ku/. Each player can only play two of

[1] See https://youtu.be/99PC3a3Pscg for an overview of the data collection procedure and the game.

Figure 2: *Experimental setup of the music game*.

these CV sequences (/ka/ and /ki/, or /ka/ and /ku/). For the example melody shown, the required pronunciation sequence is /ka/ (director), /ki/ (director), /ku/ (follower) and /ka/ (follower). If a mistake is made in this sequence (i.e., a speaker pronounces a wrong sound, or the wrong speaker pronounces a correct sound), the sequence has to be played anew from the start. Correctly played notes can be visually identified by the director (but not the follower) as these will be filled (in red or blue).

Participants are not able to see each other, as they are separated by a large computer monitor (and/or a wall), but can hear each other and are instructed to only communicate with each other via their two assigned CV sequences. (The experimenter fails the present trial if the participants use other sounds or words.) Note that it is very unlikely that a melody is played correctly by chance. This means that each dyad has to develop a communication system if they want to succeed in the task. Figure 2 shows a photo of the experimental setup.

After an initial calibration phase in which a real-time vowel recognizer (implemented in Matlab) is trained to recognize the individual sounds for both speakers, the speakers first finish a few director-only melodies to become familiar with the game. Subsequently, the 15-minute experiment starts. Initially only simple random melodies are shown (for example, pronouncing a single note), but these increase in complexity when participants successfully complete the melody within the time limit (90 seconds). A higher level of complexity is realized through increasing the length of the melody (up to 5 notes), but also through requiring two notes (one by the director and one by the follower) to be played simultaneously. Two types of simultaneous notes were possible, visualized in Figure 3. The simple form (Figure 3, left) simply requires the two notes to overlap at any time, without any restriction regarding exactly when or how long they need to overlap. The difficult form (Figure 3, right) requires one note to be shorter than the other. Specifically, the shorter note has to start after the longer note is initiated, but it has to end before the longer note ends.

To illustrate that our paradigm is more than a simple repetition task and results in real language emergence and communication, we will discuss several communicative strategies which emerged when dyads played the music game. Let's consider an example in which the director was assigned the sounds /ki/ and /ka/, and the follower the sounds /ka/ and /ku/. Sometimes, dyads converged on a system where the director would say the shared CV sequence /ka/ to signal that the follower also had to say /ka/. Similarly, when the director would say /ki/, this would signal that the follower needed to pronounce the unshared CV sequence /ku/.. Using this strategy for the melody shown in Figure 1 (i.e., /ka/, /ki/, */ku/*, */ka/*, with follower-produced sounds in italics), the director could potentially start with saying /ka/, /ki/, /ki/, /ka/, and then after a short pause say /ka/, /ki/ and then wait (and hope) for the follower to say /ku/, /ka/ (i.e., matching with the last /ki/ and /ka/ pronounced by the director in the first utterance). While this strategy works reasonably well for simple melodies such as the one shown in Figure 1, it tends to not function well in situations where there are more switches between director and follower during a single melody. The reason for this is that the follower does not know which part of the instruction provided by the director is for them to play, and which needs to be played by the director.

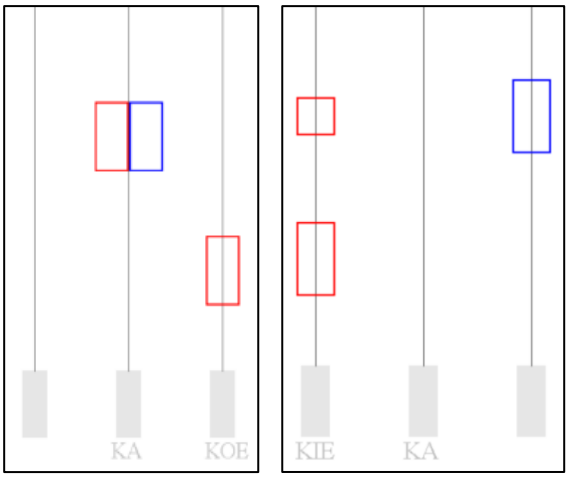

Figure 3: *Simultaneous note configurations. Left: after the director has pronounced /ku/, the director and follower have to both pronounce /ka/, so that there is at least some overlap between the two pronunciations. Right: after the director has pronounced /ki/, the follower has to say /ku/, but before the follower ends the /ku/ pronunciation, the director has to have said /ki/.*

Another example of a more effective strategy several dyads converged on was a communicative system in which the director distinguished director-played notes from those played by the follower. For example, using this strategy for the melody shown in Figure 1, the director would instruct the follower by saying (for example) /ka/, /ki/, /kaka/, /kiki/. The duplication of the notes, which were also sometimes pronounced faster, then indicated that these were notes which needed to be played by the follower (/kaka/ signaled /ka/, and /kiki/ signaled /ku/). The higher difficulty levels, where simultaneous notes were introduced, were usually only reached by dyads who employed a system such as the aforementioned one. Successfully playing such complex melodies (such as those shown in Figure 3) usually involved the director elongating the note, so that some overlap was achieved. For example, the director (who was assigned the sounds /ka/ and /ku/) might instruct the leftmost sequence of Figure 3 by saying /ku/, /kaka/, then after a pause saying /ku/, waiting until the follower said /ka/ and then also quickly saying /ka/ (as a minimal amount of overlap is sufficient). The director might instruct the rightmost sequence by first saying /ki/, /ki:ki:/, then after a pause say /ki/ and wait until the follower pronounces an elongated /ku/, during which the director quickly says /ki/ (which then hopefully ended befor the pronunciation of the elongated /ku/).

3. Data collection

Data was collected at *Lowlands Science 2019*, a public engagement science event hosted at the three-day Dutch music festival Lowlands, with over 50,000 visitors every year. After answering initial assessment questions (including information regarding musical ability and substance use; we also measured blood alcohol concentration using a professional breathalyzer), 77 pairs of (mostly Dutch) speakers played the music game. The pairs generally consisted of friends, partners or family (67 pairs). After the experiment, participants answered a few questions about how they thought the experiment went, how much they liked the person they played the game with, and what their relation was. Right before and directly after the experiment, participants produced a single sentence which included three words for each of the vowels /a/, /i/ and /u/. For one participant the sentence (with relevant vowels marked in boldface) was a question in Dutch: "H**oe** v**aa**k r**ie**p jij K**IE**, K**AA** of K**OE** tijdens dit mooie, m**aa**r n**ie**t ber**oe**rde Lowlands?" (i.e., "How often did you call out /ki/, /ka/ or /ku/ during this beautiful, but not bad Lowlands?"), whereas for the other participant the sentence consisted of a statement to prevent a shadowing effect: "Ik r**ie**p heel v**aa**k K**IE**, K**AA** of K**OE** tijdens dit n**ie**t ber**oe**rde, m**aa**r ber**oe**mde Lowlands." (i.e., "I called out /ki/, /ka/, or /ku/ very often during this not bad, but famous Lowlands."). In addition, during the training phase at the start of the game, as well as at the end of the game, both participants pronounced the sequence of five sounds /ka/, /ka/, /ki/, /ki/, /ka/ (the player who was assigned /ka/ and /ki) or /ka/, /ka/, /ku/, /ku/, /ka/ (the player who was assigned /ka/ and /ku/). To assess phonetic convergence, we analyzed both the sentence pronounced right before and directly after the experiment, and the sequence of five sounds. Specifically, in this study we only analyzed the tokens with the shared vowel /a/. All pronunciations were recorded with headworn microphones (Shure WH20). While the environmental noise was relatively loud (due to concerts playing in the vicinity of the Lowlands Science area), the headworn microphones worked very well in filtering out the background noise.

Note that we collected data in two places at the same time to maximize the amount of data we were able to collect in the three consecutive days (a total of 24 hours). In one of the two places, we collected both acoustic and ultrasound tongue imaging (UTI) data, whereas in the other place (shown in Figure 2), we only collected acoustic data. Especially for the UTI-experiment the setup was relatively elaborate, as one laptop computer was used to run the experiment, and two additional laptops were used to collect the UTI data. For the simpler acoustic-only experiment, we also used two laptops, but one was used as a backup system for the collected acoustic recordings.

4. Results

To assess phonetic convergence in the pre- and post-game sentences and sequences, we calculated F1-F2 (Mel-scaled) based Euclidean distances between the two speakers in a pair for the shared vowel /a/, both at the beginning and at the end of the experiment. Using mixed-effects regression analysis with the optimal random-effects structure, we observed no significant phonetic convergence for the sentences (β = -1.2, t = -0.2, p = .81). However, phonetic convergence was clearly present for the /a/-vowels in the sequence (β = -12.6, t = -2.5, p = .01; see the bean plot in Figure 4 on the basis of all 77 pairs). The effect appeared to be robust, as it remained significant even after excluding 41 pairs where at least one of the speakers had used alcohol or drugs.

To assess whether personal characteristics were affecting the level of convergence, we calculated an individual (rather than a pair-based) measure of convergence for each speaker (S) compared to their interlocutor (I). Our measure was obtained by comparing the pronunciation of the speaker to the interlocutor's pronunciation *at the beginning of the experiment*, both for the speaker's pronunciation at the beginning of the experiment and for the speaker's pronunciation at the end of the experiment (based on the F1-F2 (mel-scaled) Euclidean distances) as shown in Equation (1).

$$S_{conv} = \delta(S_{start}, I_{start}) - \delta(S_{end}, I_{start}) \quad (1)$$

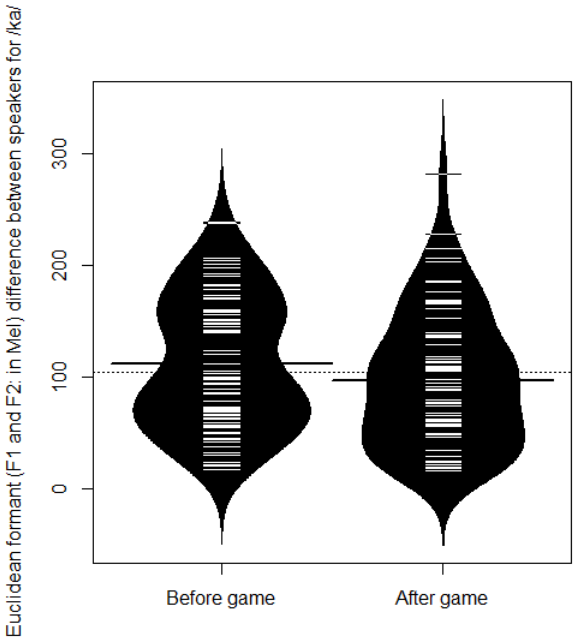

Figure 4: Bean plot visualizing the task-specific phonetic convergence effect.

Using mixed-effects regression analysis with the optimal random-effects structure, no predictors were found to be significant. For both the sentences and the sequence, the (non-significant) predictor which appeared to be most predictive, when focusing on the speakers who had not consumed any alcohol and reported no drug use, was gender. Men tended to show more convergence towards the initial pronunciation of their interlocuter than women (β = 12.6 t = 1.7, p = .10 for the sequence – see Figure 5, and β = 10.9 t = 1.7, p = .10 for the sentence).

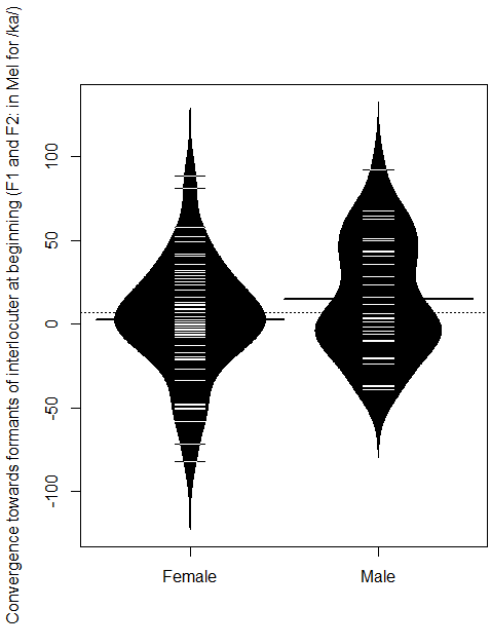

Figure 5: Bean plot visualizing the (non-significant, p = .10) gender effect on convergence towards the initial pronunciation of the interlocuter.

5. Discussion and conclusion

In this study we have illustrated a new experimental paradigm which is both highly controlled, resulting in many repetitions of simple sounds, but also results in the emergence of a simple language and concomitant communicative strategies. Our paradigm was shown to result in task-specific phonetic convergence (i.e., the task-specific sequences converged, but not the normal sentences) after only 15 minutes of interaction. No more general convergence was shown, but this may have been caused by the large majority of the speakers already knowing each other well, but also the difference in type of sentences (i.e. question vs. declarative) the two speakers had to pronounce. We did not find significant personal characteristics that affected convergence. Out of the non-significant predictors, the strongest effect was found for gender, with men showing stronger convergence towards their interlocutor (at least, the initial pronunciation of their interlocutor) than women. While we have no direct explanation for this (also non-significant) pattern, gender-specific phonetic convergence effects have often been observed (e.g., Pardo *et al.*, 2018). We have only analyzed the acoustic characteristics of the /ka/ vowel pronounced at the beginning and at the end of the game. In future work, we aim to investigate whether phonetic convergence can also be observed during the course of the task itself.

6. Acknowledgements

We thank the University of Groningen, the Young Academy Groningen, and the Groningen University Fund for the financial support which has made this research project possible. We further thank the organization of *Lowlands Science 2019* for the opportunity to conduct our study at the festival.

7. References

Mills, G. (2011). The emergence of procedural conventions in dialogue. In Proceedings of the 33rd Annual Conference of the Cognitive Science Society. 471-476.

Pardo, J. S. (2017). Parity and disparity in conversational interaction. E. Fernández, H.S. Cairns (eds.) The Handbook of Psycholinguistics, Wiley-Blackwell, New York. 131-152.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, 79(2), 637-659.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, 69, 1-11.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329-347.