A Shadowing Experiment with Natural and Synthetic Stimuli

Iona Gessinger[1], Eran Raveh[1], Johannah O'Mahony[1]
Ingmar Steiner[1,2], Bernd Möbius[1]

[1]Saarland University
[2]DFKI GmbH
lastname@coli.uni-saarland.de

Keywords: phonetic convergence, human-computer interaction, shadowing task

1. Introduction

Inter-speaker accommodation is a phenomenon observed in human communication. Phonetic convergence is one way for a speaker to accommodate to an interlocutor. It is defined as an increase in segmental and suprasegmental similarities between two speakers [1]. Phonetic convergence has been found for human-to-human interaction in both spontaneous, conversational speech [1, 2] and non-conversational speech occurring in experimental settings such as the shadowing task [3, 4]. Previous studies on convergence in human-to-human interaction looked at suprasegmental features such as f0 range [5] and speaking rate [6], as well as segmental features such as spectral properties of vowels [3] and voice onset time [7].

Thus far, phonetic convergence has received little to no attention in the field of human-computer interaction (HCI). In the experiment introduced in this paper, we take a first step in investigating whether human speakers also converge to synthesized speech by conducting a shadowing experiment using both natural and computer-generated stimuli, concentrating on selected segmental features (cf. 2.1).

Based on previous findings in human-human interaction, we expect to observe phonetic convergence on the segmental level for the natural stimuli. Since the quality of synthesized speech is improving and HCI is becoming ever more used for various tasks in everyday life, humans are likely to interact in a similar way with computers as they do with humans. Therefore, we expect to observe convergence for the synthetic stimuli as well. However, the degree of convergence might still be influenced by the perceived naturalness of the synthetic stimuli.

2. Experiment

The following experiment consists of two conditions. In the first condition a group of participants is presented with short sentences recorded by natural speakers. In the second condition a different group of participants is presented with a synthesized version of the same sentences. The amount of convergence in the natural speech condition will serve as a baseline for the synthetic speech condition. Only the natural condition will be discussed in this paper.

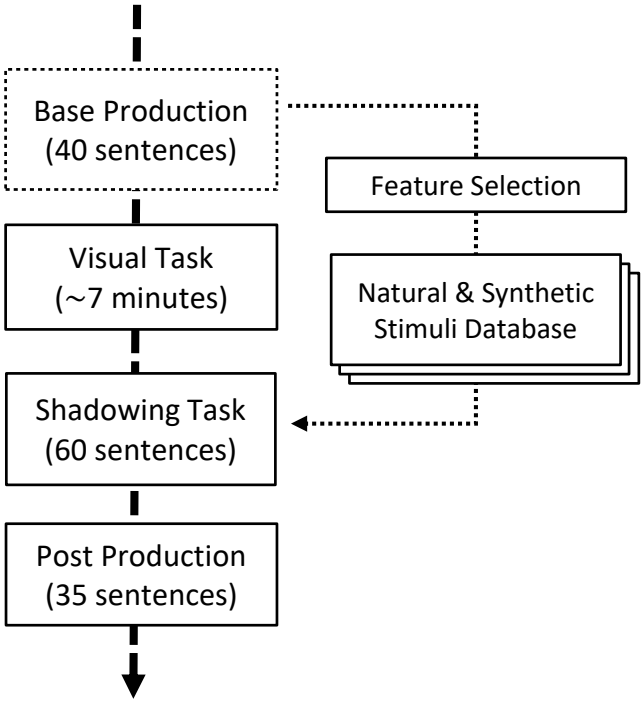

Figure 1: *Workflow of the experiment, showing its four phases. The stimuli presented in the shadowing task are selected based on feature realization in the baseline production.*

2.1. Target phenomena

To investigate convergence at the segmental level, three target features were selected that show variation across native speakers of German: [ɛː] vs. [eː] as a realization of the vowel *-ä-* in stressed position, as in *Gerät*, [ɪç] vs. [ɪk] as a realization of the final syllable *-ig*, as in *König*, and elision or epenthesis of [ə] in the final syllable *-en*, as in *reden*.

The former two features vary mainly regionally, with a preference for [eː] and [ɪç] in Northern Germany and a preference for [ɛː] and [ɪk] in Southern Germany [8, p. 64ff.]. All four forms are part of the phonetic inventory of standard German (*Pfirsich* [ɪç] / *Plastik* [ɪk] / *Säle* [ɛː] / *Seele* [eː]) and are hence expected to be known to all speakers. Despite the regional distribution of the two features, they are not strong dialectal markers [9, p. 560 for [ɪç]/[ɪk]] and often persist in otherwise standard German

Table 1: *Examples of target and filler sentences with corresponding target features.*

target sentence	target feature
Die Best**ä**tigung ist für Tanja.	[ɛː] vs. [eː]
Ich bin süchti**g** nach Schokolade.	[ɪç] vs. [ɪk]
Wir begleit**en** dich zur Taufe.	[ən] vs. [n̩]
filler sentence	
Der Kaffee war ja schon kalt.	—

productions of native speakers. Nevertheless, [ɛː] and [ɪç] are the official standard German forms of the respective features in the given contexts.

The elision of [ə] in the final syllable *-en* after plosives and fricatives is a highly expected phenomenon in standard German speech. In this position, [ə] is only produced when speaking particularly slowly and clearly [8, p. 39].

2.2. Stimuli

Thirty short German sentences (15 targets and 15 fillers) serve as stimuli for the shadowing task. Each target sentence contains one target feature only (i.e. five sentences per feature). The filler sentences do not contain any of the target features (cf. Table 1).

10 additional filler sentences are included in the baseline production, five of which are shown at the beginning of the recording to familiarize the participants with the task. The other five additional fillers contain tokens such as *Pfirsich* and *Plastik* to verify that participants were able to produce [ɪç] and [ɪk] respectively in a context other than *-ig*.

The first set of stimuli was recorded by two native speakers of German (1 female, 25 years old and 1 male, 23 years old). All 30 sentences were presented on a computer screen in random order. The speakers were instructed to speak naturally. Then the 15 target sentences were presented again, grouped by target feature. Possible feature variations as presented above were pointed out, and speakers were instructed to produce both variations. The best tokens regarding target feature production and overall clarity were selected for presentation in the shadowing task.

The second set of stimuli was generated using the text to speech system MaryTTS [10] with HMM synthesis.[1] One female and one male synthetic voice were used to match the gender of the natural speakers. In order to control for potential differences in information structure between the natural and the synthetic stimuli, prosodic characteristics of the synthetic stimuli were manipulated to match the natural stimuli.

2.3. Participants

21 native speakers of German (17 females, 19-33 years old, mean = 25.8, and 4 males, 23-34 years old, mean = 29.5) with no speech, language, or hearing impairments were recruited as participants for the first condition of the experiment. Another group of participants will be recruited for the second condition.

[1] http://mary.dfki.de/

2.4. Procedure

The experimental procedure consists of four tasks: baseline production, visual task, shadowing task and post production (see Figure 1). For the baseline production, 40 short sentences (15 targets, 15 fillers, and 10 additional fillers) were presented to the participants on a computer screen in random order. There were no instructions with respect to speaking style. Productions were recorded under the same conditions as model speaker productions. The realizations of the target features were noted by the experimenters during the baseline production. In order to weaken the mental representation of their first production, the participants were asked to perform a visual task after the baseline production.

In the shadowing task, the participants were presented with the productions of the two model speakers (15 targets and 15 fillers per model speaker; grouped by model speaker; semi-randomized for balanced distribution of targets over the two sets; alternating order of model speaker presentation). The target sentences played back to the participants always contained the opposite target feature realization of that observed in the participants' baseline productions (for instance, a participant who predominantly produced [ɪk], [ɛː] and elided [ə] in the baseline condition was exposed to [ɪç], [eː] and [ən] in the shadowing condition).

Words such as "repeat" and "imitate" were avoided in the instructions, so that converging behavior was not encouraged by the choice of words. Immediately after the shadowing task, participants were again presented with the written form of the stimuli to record the post production.

The second group of participants will undergo the same experimental procedure, but with synthetic instead of natural stimuli in the shadowing task.

3. Results

For a preliminary analysis of the target feature [ɛː] vs. [eː] the participants were divided into two groups based on their baseline production. The first two formants of the vowels were measured at mid-point and plotted along with the productions of the models the respective group heard in the shadowing task (cf. Figure 2).

For the group of participants which prefer [eː], the productions in the shadowing condition tend to move toward the model speakers' productions compared to the baseline condition. The same, slightly smaller effect can be observed for the post condition.

For the group of participants which prefer [ɛː], the productions in the shadowing condition also tend to move toward the model speakers' productions, but to a lesser extent than in the first group. The post condition, however, does not differ remarkably from the baseline condition for the second group.

These tendencies also become apparent when calculating the mean Euclidean distance between each of the participants' productions and the mean production of the models, using the formula

$$d(p,m) = \sqrt{(p_{F1} - m_{F1})^2 + (p_{F2} - m_{F2})^2} \qquad (1)$$

where p and m are points in the two-dimensional space

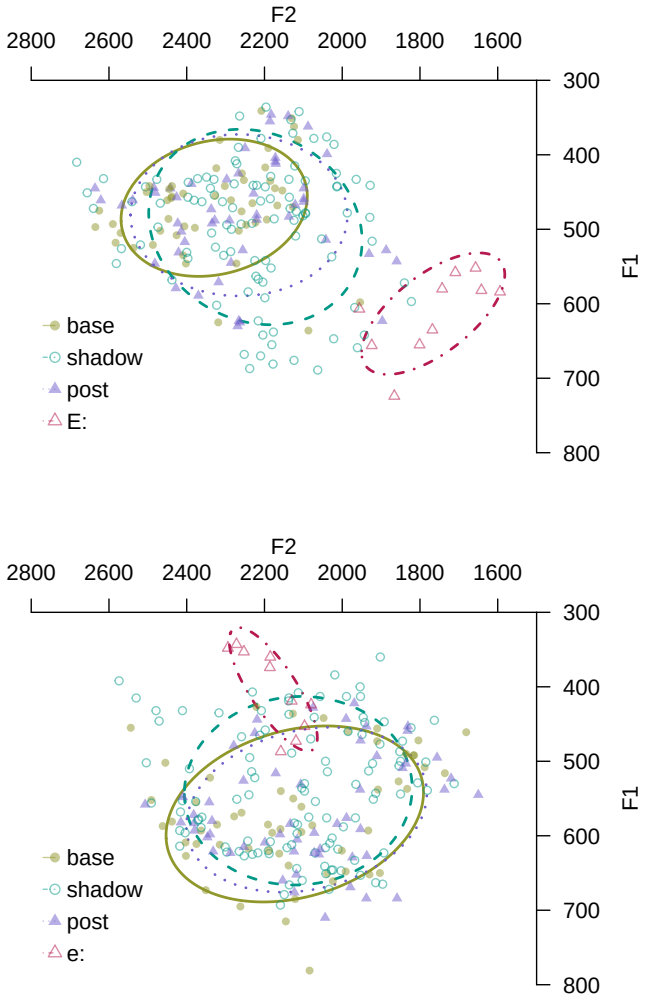

Figure 2: *Visualization[2] of target feature [ɛː] vs. [eː] produced by participants with baseline preference [eː] (upper figure; n = 11) and [ɛː] (lower figure; n = 10) in the production tasks **base**, **shadow** and **post**, as well as the human models (n = 2) they heard in the shadowing task, producing [ɛː] (upper figure) and [eː] (lower figure) (**red**). The ellipses visualize the confidence level of the estimated true mean (here: ±1 standard deviation).*

representing the production of the first two formants of vowels (see also Figure 2), p_{F1} and m_{F1} are the values of the respective first formants of the productions in Hertz, and p_{F2} and $mF2$ are the respective values of the second formants of the productions in Hertz. As this measure of distance is derived from frequency in Hertz, its unit will be called Hertz distance (HzD) in the following.

For the group of participants which prefer [ɛː] the Euclidean distance is 587 HzD (sd = 149) for the baseline condition, 482 HzD (sd = 175) for the shadowing condition, and 524 HzD (sd = 176) for the post condition. For the group of participants which prefer [eː] the Euclidean distance is 276 HzD (sd = 90) for the baseline condition, 241 HzD (sd = 92) for the shadowing condition, and 266 HzD (sd = 96) for the post condition.

The target feature [ɪç] vs. [ɪk] was categorically evaluated. Each segment was categorized either as a fricative (accounting for [ɪç]) or as a plosive (accounting for [ɪk]). The fricative category also includes the variations [ʃ] or [j̊], which were produced in a small number of cases. Figure 3 shows the distribution of changes between productions in baseline condition and shadowing condition. It comprises productions of speakers with plosive preference producing a fricative ([k] → [ç] convergence), productions of speakers with fricative preference producing a plosive ([ç] → [k] convergence), and productions that were the same category as in the baseline production of the speaker (no convergence).

In total, convergence was observed in 39% of the productions (20% [k] → [ç] and 19% [ç] → [k]). Participants' productions showed the following three patterns: consistent production in the baseline condition and consistent opposite production in the shadowing condition (i.e. complete convergence), same production in baseline and shadowing conditions (i.e. no convergence) or non-consistent productions in one or both of the conditions (i.e. partial convergence).

Figure 3: *Visualization of target feature [ɪç] vs. [ɪk] in the shadowing condition, showing cases where speakers changed their production compared to the baseline condition from fricative to plosive, from plosive to fricative, or didn't change their production.*

Regarding the target feature [ən] vs. [n̩], no participant showed a natural preference for schwa epenthesis in the baseline production. Hence, all participants were presented with the unreduced productions of the model speakers. For the preliminary analysis, each final syllable *-en* was acoustically and visually checked for epenthesis of schwa. A segment was only counted as schwa if it was longer than 30 ms.

Under this condition, the following number of schwa epentheses was observed: 2 out of 105 trials in the baseline condition (1.9%), 22 out of 210 trials in the shadowing condition (10.5%) and 4 out of 105 in the post condition (3.8%).

As in the case of [ɪç] vs. [ɪk], different individual production patterns were observed across the participants, ranging from no convergence to complete convergence.

4. Conclusion

We presented first results from an ongoing shadowing experiment with natural and synthetic stimuli. In this experiment, phonetic convergence on the segmental level is examined in the context of short sentences. Preliminary analysis of the natural condition shows convergence for all three target phenomena. The degree of convergence varied across the participants.

[2]Plots were generated using *phonR*, http://drammock.github.io/phonR/

5. Acknowledgments

We would like to thank Dr. Les Sikos for helpful suggestions concerning the experimental design and Dr. Sébastien Le Maguer for technical assistance in generating the synthetic stimuli.

6. References

[1] J. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[2] N. Lewandowski, "Talent in nonnative phonetic convergence," Ph.D. dissertation, Universität Stuttgart, 2012.

[3] K. Shockley, L. Sabadini, and C. Fowler, "Imitation in shadowing words," *Perception & Psychophysics*, vol. 66, no. 3, pp. 422–429, 2004.

[4] M. Babel, G. McGuire, S. Walters, and A. Nicholls, "Novelty and social preference in phonetic accommodation," *Laboratory Phonology*, vol. 5, no. 1, pp. 123–150, 2014.

[5] C. Smith, "Prosodic accommodation by french speakers to a non-native interlocutor," in *Proceedings of the XVIth International Congress of Phonetic Sciences*, 2007, pp. 313–348.

[6] J. Pardo, I. Jay, and R. Krauss, "Conversational role influences speech imitation," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2254–2264, 2010.

[7] A. Walker and K. Campbell-Kibler, "Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task," *Frontiers in Psychology*, vol. 6, no. 546, 2015.

[8] S. Kleiner and R. Knöbel, Eds., *Der Duden in zwölf Bänden*, 7th ed. Berlin: Dudenverlag, 2015, vol. 6: Duden - Das Aussprachewörterbuch.

[9] H. Mitterer and J. Müsseler, "Regional accent variation in the shadowing task: evidence for a loose perception-action coupling in speech," *Attention, Perception & Psychophysics*, vol. 75, no. 3, pp. 557–575, 2013.

[10] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: a tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003.