Shadowing Synthesized Speech – Segmental Analysis of Phonetic Convergence

Iona Gessinger[1,2], *Eran Raveh*[1,2], *Sébastien Le Maguer*[1,2], *Bernd Möbius*[1], *Ingmar Steiner*[1–3]

[1]Computational Linguistics & Phonetics, Saarland University, Germany
[2]Multimodal Computing and Interaction, Saarland University, Germany
[3]German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany

gessinger@coli.uni-saarland.de

Abstract

To shed light on the question whether humans converge phonetically to synthesized speech, a shadowing experiment was conducted using three different types of stimuli – natural speaker, diphone synthesis, and HMM synthesis. Three segment-level phonetic features of German that are well-known to vary across native speakers were examined. The first feature triggered convergence in roughly one third of the cases for all stimulus types. The second feature showed generally a small amount of convergence, which may be due to the nature of the feature itself. Still the effect was strongest for the natural stimuli, followed by the HMM stimuli and weakest for the diphone stimuli. The effect of the third feature was clearly observable for the natural stimuli and less pronounced in the synthetic stimuli. This is presumably a result of the partly insufficient perceptibility of this target feature in the synthetic stimuli and demonstrates the necessity of gaining fine-grained control over the synthesis output, should it be intended to implement capabilities of phonetic convergence on the segmental level in spoken dialogue systems.

Index Terms: phonetic convergence, synthesized speech, human-computer interaction, shadowing task

1. Introduction

This paper reports the results of an experiment aiming to investigate the degree of phonetic convergence in humans when shadowing synthetic speech. Phonetic convergence, which is defined as an increase in segmental and suprasegmental similarity between two speakers [1], has been found and thoroughly studied in human-human interaction. This includes conversational [1, 2] and non-conversational [3, 4] settings, as well as the segmental [5, 6] and suprasegmental [3, 7] levels. Therefore, phonetic convergence is assumed to be a property of natural dialogue. As spoken dialogue systems are being developed with the goal to eventually emulate natural dialogue situations, the implementation of convergence capabilities in such systems may be one step in the direction of achieving this goal. To gain a better understanding of how this could be implemented, it is crucial to first explore the extent to which phonetic convergence happens when humans interact with synthesized speech. Do they converge to the same, a lesser, or even to a greater extent as when interacting with natural speech? If phonetic convergence is indeed a subconscious process and takes place automatically, as some authors suggest [8, 9], it can be assumed that it will occur in human-computer interaction as well. Given that factors like typicality and attractiveness of voices have been shown to influence the degree of phonetic convergence in humans [4], different types of synthesized speech may have varying influence on their interlocutors, just as different natural voices do.

In the present experiment, the shadowing task paradigm [10] was chosen as an approximation to a dialogue situation. Segment-level phonetic features were manipulated to investigate whether human interlocutors converge to synthesized speech while shadowing it. The first part of the experiment tested natural stimuli to establish a baseline of the convergence degree that can be expected in the given setting. The following two parts used two sets of synthetic stimuli generated with diphone synthesis [11] and hidden Markov model (HMM) based synthesis [12], respectively. Preliminary results were previously reported in [13] for the natural condition and in [14] for the diphone condition. This paper presents the HMM part of the experiment, and summarizes and compares the findings of the full experiment over all three conditions.

2. Stimuli

2.1. Text material and target features

The examined segment-level phonetic features show variation across native speakers of German: realization of the word-medial vowel *-ä-* in stressed syllables as [eː] or [ɛː], realization of the word-final sequence *-ig* as [ɪç] or [ɪk], and elision or epenthesis of [ə] in the word-final sequence *-en*. The former two features vary regionally, occurring roughly in the North and South of the German-speaking region of Europe, respectively. However, they are not assumed to be perceived as strong dialectal markers (cf. [15] for [ɪç] vs. [ɪk]). The third feature varies mainly with speaking style.

The features are embedded in short German sentences (cf. Table 1). Our corpus consists of 15 target sentences and 25 filler sentences. Each of the three target features appears in five target sentences (3 declaratives and 2 interrogatives) and does not appear in other sentences of the corpus. For the purpose of this experiment, all three features – naturally categorical or continuous – are initially described as two-way contrasts. During the experiment, participants' productions are acoustically identified as belonging to one of the two classes. For [ɪç] vs. [ɪk], equivalent productions are counted as well – e.g., [ɪʃ] for the former category, or [ɪɡ̊] for the latter.[1]

2.2. Generation

For the natural stimuli, two native speakers of German (25 year old female and 23 year old male) were digitally recorded (with a sampling rate of 48 kHz) in a sound-attenuated booth using a stationary DPA 2011A cardioid microphone. 15 target sentences and 15 fillers from the corpus were presented on a com-

[1] These variants are equivalent in the sense that they are phonetically different realizations of the same underlying class – fricative vs. plosive.

Table 1: *Examples of sentences containing the target features.*

sentence	target feature
Ich mag die Qualit**ä**t deiner Tasche.	[ɛː] vs. [eː]
Es ist ganz schön staub**ig** im Keller.	[ɪç] vs. [ɪk]
Sind die Aff**en** denn zutraulich?	[n̩] vs. [ən]

puter screen and the speakers were instructed to speak naturally, as if in conversation with someone. Subsequently, the 15 target sentences were presented again, grouped by target feature. The respective feature variations were explained to the speakers and they were asked to produce both versions. The best tokens in terms of target feature production and overall clarity were selected.

The first set of synthetic stimuli was created using diphone synthesis with MBROLA [11]. One female and one male voice were used to match the gender of the natural speakers. For the realization of the target features, phonetic transcriptions of the target sentences were provided to the system in the two pronunciation versions. To control for potential differences in prosody and information structure between the natural and the synthetic stimuli, the f_0 contours and segment durations of the natural stimuli were passed as parameters to the synthesis system.

The second set of synthetic stimuli was created using the state of the art HMM-based Speech Synthesis System (HTS) (version 2.3) [12] with the BITS unit selection corpus.[2] As for the diphone condition, the f_0 contours and segment durations of the natural stimuli were imposed on the synthetic stimuli (one female and one male voice). The process of imposing the f_0 values poses certain difficulties. The consistency between presence and absence of f_0 and the spectral shape, represented by the mel-generalized cepstra (MGC), must be maintained. There are two approaches to achieve this goal while staying close to the standard HTS process: using the voicing decisions obtained from HTS, or predicting voicing directly from the spectrum. In an informal comparison, the second approach showed fewer artifacts in the rendered signal and was therefore applied using a multi-layer perceptron (MLP) architecture. First, the segment durations were taken from the natural stimuli and imposed during the parameter generation stage of HTS. Subsequently, the MGC and band aperiodicity (BAP) coefficients were retained from the output. A neural network was used to predict the voicing property from the MGC coefficients. Then, a voicing mask was applied on the predefined f_0 contour to obtain the final f_0 coefficients. Finally, the MGC, BAP, and f_0 coefficients were used to generate the output signal in a standard synthesis chain with a mel log spectrum approximation filter and the STRAIGHT vocoder [16]. The descriptive features and question sets used to build the decision tree followed the standard English set proposed in [17] with adapted part-of-speech tags and phonemes for German. The hidden layer of the MLP consisted of 128 neurons.

All 270 stimuli (45 stimuli × 3 stimulus types × 2 genders) were stored in a database for later use.

3. Experimental procedure

The procedure of this experiment consists of four phases and varies only with regard to stimulus type in the three conditions (cf. Figure 1). In the baseline production, the participants read 40 sentences from a screen. Depending on their realization of the target features during this phase, the stimuli for the shadowing task were selected from the database so that they contained the opposite of what the participants naturally produced. As there were five occurrences of each feature in the baseline set, a preferred form (at least 3 occurrences) could be found even for participants who varied in their production. In order to weaken the mental representation of their baseline production, the participants performed a visual task after the baseline phase which consisted of playing a game that did not require any linguistic interaction. During the shadowing task, two sets of 30 stimuli were played back to the participants over headphones (male and female voices; 15 targets and 15 fillers per voice; semi-randomized for balanced distribution of targets over the two sets). To avoid priming of convergence, words such as "repeat" and "imitate" were not used in the instructions. Rather, the participants were told to "listen and then speak". Immediately after the shadowing task, the participants read 35 sentences from a screen to record the post production. All productions were recorded in the same manner as the model speakers (cf. Section 2.2).

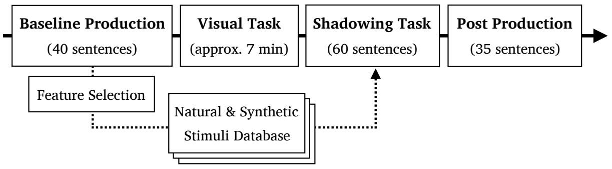

Figure 1: *The four phases of the experiment. The stimuli presented in the shadowing task are selected from the database depending on the feature realization in the baseline production.*

3.1. Participants

The participants were recruited on the Saarland University Campus and paid for taking part in the experiment. All 56 participants were native speakers of German, and 11 of them had more than one native language. All had learned at least one, and the majority more than two, foreign languages. Most of them were students; six participants had non-academic jobs. The participants came from ten different German states and Austria with roughly 70 % from an area where a southern variety of German is spoken and 30 % from a northern region. In a questionnaire completed after the experiment, 80 % of the participants answered affirmative to the question whether they change the way they speak depending on their interlocutor; 50 % believed they would converge to an interlocutor of the same dialectal background; only 15 % claimed they would do the same with an interlocutor of a different dialectal background; 16 % said that they intentionally imitate their interlocutors' pronunciation.

Each participant was presented with only one of the three stimulus types – see Table 2 for details of participants in the three conditions. The participants' preference regarding the examined phonetic features as identified during the baseline phase is given in Table 3.

At the end of the experiment, the participants were asked which version of each feature they believe to produce themselves and what they think of the respective other version. In summary, 70 % to 80 % of the participants reported a positive attitude towards the version they do *not* believe themselves to produce. This includes ratings such as "also ok", "better", and "Standard German". Only a minority of participants showed a

[2] http://www.bas.uni-muenchen.de/forschung/Bas/BasBITSUSeng.html

Table 2: *Participants' gender and age.*

condition	no. of participants		age range	mean age
Natural	17	female	19 to 33	26
	4	male	23 to 34	30
Diphone	14	female	19 to 50	26
	4	male	23 to 34	27
HMM	13	female	18 to 51	28
	4	male	22 to 37	25

Table 3: *Number of participants preferring the respective version of the phonetic features as identified in the baseline phase.*

condition	[ɛː] vs. [eː]		[ɪç] vs. [ɪk]		[n̩] vs. [ən]	
Natural	11	10	12	9	21	0
Diphone	14	4	9	9	17	1
HMM	10	7	6	11	16	1

negative attitude towards the other versions such as "wrong", "weird", and "sounds artificial". It seems plausible that a positive attitude towards the features entails a higher probability of converging to them.

4. Analyses and results

4.1. [ɛː] vs. [eː]

For the analysis of the vowel productions, the participants were split into two groups, namely participants with preference [ɛː] and those with preference [eː]. The first and second formants of the target vowel were measured at the midpoint in all productions as well as in the stimuli using Praat's [18] Burg algorithm. Figure 2 visualizes the productions of the natural condition for preference group [eː] and the stimuli they heard in the F1-F2 space. If participants converged with regard to vowel quality, the productions from the shadowing phase would be located closer to the model speaker productions than those from the baseline phase. The Euclidean distance (EucD) between each participant vowel in the F1-F2 space and the mean stimulus vowel (female and male combined) was calculated. These data were used to fit linear mixed-effects models (LMMs) with experiment phase and subject gender as fixed effects, subject and target word as random effects (intercepts and slopes), and EucD as the dependent variable. Models were fit for both preference groups of each condition and then compared by means of an ANOVA with corresponding null models where experiment phase was removed from the fixed effects structure. These comparisons showed a significant effect of experiment phase on EucD only for the natural condition with

- $\chi^2(2) = 7.2, p < 0.05$ for preference group [ɛː]
- $\chi^2(2) = 8.8, p = 0.01$ for preference group [eː]

A post-hoc Tukey test showed that this effect is significant between baseline and shadowing phase, and not significant between shadowing and post phase for

- preference group [ɛː] with
    - $z = -2.8, p < 0.05$ for base – shadow
      → reduction of EucD by about (34 ± 12) Hz
    - $z = -0.6, p = 0.8$ for shadow – post
- preference group [eː] with
    - $z = -3.7, p < 0.001$ for base – shadow
      → reduction of EucD by about (105 ± 29) Hz
    - $z = -1.9, p = 0.1$ for shadow – post

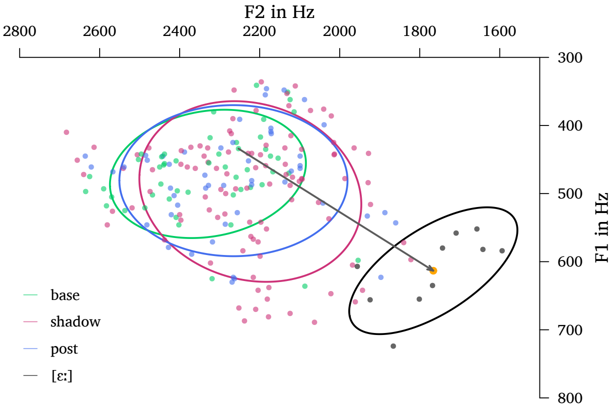

Figure 2: *Vowel productions of preference group [eː] from natural condition in* ***base***, ***shadow***, *and* ***post*** *phase, as well as values of* **[ɛː]** *in the female and male natural stimuli the participants shadowed. The arrow illustrates the Euclidean distance between one participant production and the* ***model mean***. *The ellipses show ±1 standard deviation from the bivariate mean.*

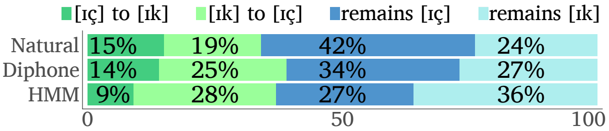

Figure 3: *Percentages of convergence (green) and non-convergence (blue) in the shadowing task for the three stimulus types Natural ($N = 188$), Diphone ($N = 152$) and HMM ($N = 155$).*

By further splitting the groups into sub-groups only containing female or male participants, one more indication of a borderline significant effect of experiment phase on EucD was found: for the HMM condition, the vowel productions of the female participants preferring [ɛː] ($n = 6$) were influenced by experiment phase as shown by

- ANOVA (full LMM vs. null LMM)
    - $\chi^2(2) = 4.5, p = 0.1$
- Post-hoc Tukey test
    - $z = -2.2, p = 0.065$ for base – shadow
      → reduction of EucD by about (75 ± 34) Hz
    - $z = 0.7, p = 0.8$ for shadow – post

These findings suggest that for both preference groups of the natural condition, as well as for the female participants with preference [ɛː] of the HMM condition, vowel productions became closer to the model mean during the shadowing phase and did not return entirely to the baseline level in the post phase.

4.2. [ɪç] vs. [ɪk]

The occurrences of this feature were visually and acoustically analyzed and thereafter classified as either belonging to the fricative category [ɪç] (also including e.g., [ɪʃ]) or the plosive category [ɪk] (also including e.g., [ɪg̊]). In each of the three conditions, at least 50 % of the participants produced only one of the two versions during the baseline phase, viz., 75 % for the HMM condition, 62 % for the natural condition, and 50 % for the diphone condition. The other speakers produced both target forms during the baseline phase. Only cases where baseline production and stimulus version of the feature were not the same, were counted as possible instances of convergence. The amount

Table 4: *Percentage of schwa occurrences and total number of target sentences potentially containing schwa (N) during the three experimental phases.*

condition	**base**	**shadow**	**post**
Natural	1.9 %	10.9 %	3.8 %
	$N = 105$	$N = 210$	$N = 105$
Diphone	2.4 %	4.1 %	1.2 %
	$N = 85$	$N = 170$	$N = 85$
HMM	2.5 %	7.5 %	1.25 %
	$N = 80$	$N = 160$	$N = 80$

of convergence during the shadowing task was determined by counting the cases where [ɪç] became [ɪk] (or vice versa) for the same target word. As shown in Figure 3, this could be observed, for the natural condition in 34 %, for the diphone condition in 39 %, and for the HMM condition in 37 % of the cases. To gain some insight into how persistent the convergence effect was, baseline and post production instances of the same target words were compared. The effect was most persistent for the diphone stimuli with 22 % of the stimuli differing between baseline and post production, followed by 14 % for the natural stimuli, and 11 % for the HMM stimuli.

4.3. [n̩] vs. [ən]

To decide whether schwa was present or absent in the participants' target word productions, the duration of potential segments between the preceding consonant – [d], [t], [ç], [x], or [f] – and the final nasal was measured. A duration of 30 ms was established as a minimum threshold to count the segment as schwa. This decision is supported by the fact that all schwas occurring in the model stimuli, were at least 30 ms long. Furthermore, there were only two participants with a preference of [ən] in their baseline productions (cf. Table 3).[3] This indicates that schwa does not commonly appear in the examined context. Table 4 shows the percentage of schwa occurrences during the three experimental phases. For all stimulus types, there was an increase of schwa occurrences between baseline and shadowing phase with 9 % for the natural condition, 5 % for the HMM condition, and 1.7 % for the diphone condition. In the post production, the number of schwa occurrences decreased to approximately the baseline level for all conditions.

5. Discussion and conclusion

The feature [ɛː] vs. [eː] showed a clearly significant convergence effect only for the natural stimuli. Group effects were not found for the diphone condition and only to a limited extent for the HMM condition. Figure 4 illustrates a possible reason for the overall worse performance of the synthetic stimuli in these two conditions. It shows the areas of the F1-F2 space that are occupied by the target vowels [ɛː] and [eː] for all stimulus types. Two main differences between the natural and the synthetic stimuli are noticeable. First, the vowels of the same category from the male and female diphone voices occupy a much larger area than those from the male and female natural voices. This may result in a less-distinct convergence target. Second, all instances that are supposed to be [ɛː] in the female HMM voice are located in the area of [eː].[4] Therefore, in the HMM condition, all participants with preference [eː] heard their preferred version of the vowel half of the time during the shadowing phase. This leads to the conclusion that the lack of a stronger effect of experiment phase on EucD could be due to the acoustic properties of the target vowels in the diphone and HMM stimuli, and not necessarily to the synthetic nature of the stimuli itself.

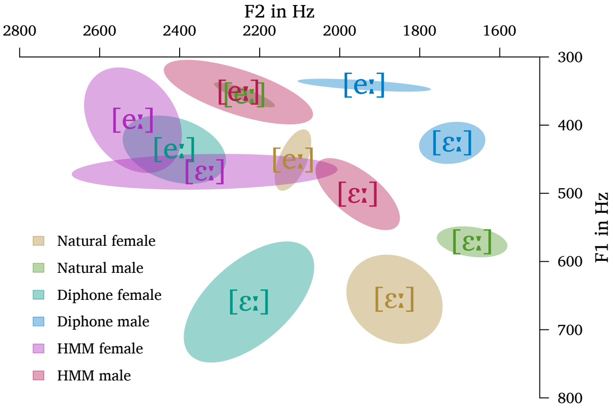

Figure 4: *Areas in the F1-F2 space populated by [ɛː] and [eː] productions from the three stimulus types. The ellipses show ±1 standard deviation from the bivariate mean.*

The feature [ɪç] vs. [ɪk] was found to be a stable trigger of phonetic convergence. For all three stimulus types, the participants produced the opposite of what they preferred during the baseline phase in roughly one third of the possible cases. These cases of convergence do not only stem from participants that already showed both target forms in the baseline production, but also from participants that produced only one of the two forms in the first phase of the experiment.

The feature [n̩] vs. [ən] did lead to a rather small convergence effect. This was expected as schwa is not usually produced in the word-final sequence *-en*. Nevertheless, in all three conditions more instances of schwa were produced during the shadowing phase than in the baseline and post phase. These productions are mostly attributable to one or two participants per condition who responded even to such an unusual segmental feature. This leads us to observe that apart from the identified group differences, the overall degree of convergence varied considerably among the participants, with some being resistant to the manipulations in the speech input and others responding strongly to them. A detailed analysis of converging behavior on the individual participant level is yet to be conducted.

Finally, it can be summarized that humans do indeed converge phonetically when interacting with synthesized speech. However, the degree of convergence depends on the nature of the target feature. Perceptibility of the target feature in the stimuli is proposed as a possible explanation for the fact that one of the examined features did not show the same extent of convergence for the synthetic stimuli as for the natural ones. This indicates that for the implementation of convergence capabilities in spoken dialogue systems, more fine-grained control over the synthesis output must be achieved.

6. Acknowledgements

We are grateful to Florian Schiel for providing access to the BITS unit-selection corpus, and to Johannah O'Mahony for her assistance with the annotation process.

[3] These two participants were excluded from the analysis.

[4] Also for Bark difference normalized formant values.

7. References

[1] J. S. Pardo, "On phonetic convergence during conversational interaction," *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, Stuttgart University, 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. E. 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 nonnative interlocutor," in *16th International Congress of Phonetic Sciences (ICPhS)*, Aug. 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] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[9] S. Dufour and N. Nguyen, "How much imitation is there in a shadowing task?" *Frontiers in Psychology*, vol. 4, p. 346, 2013.

[10] S. D. Goldinger, "Echoes of echoes? An episodic theory of lexical access," *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998.

[11] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *4th International Conference on Spoken Language Processing (ICSLP)*, vol. 3, Oct. 1996, pp. 1393–1396.

[12] H. Zen and T. Toda, "An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005," in *9th European Conference on Speech Communication and Technology (Eurospeech)*, Sep. 2005.

[13] I. Gessinger, E. Raveh, J. O'Mahony, I. Steiner, and B. Möbius, "A shadowing experiment with natural and synthetic stimuli," in *Phonetik & Phonologie 12*, Oct. 2016, pp. 58–61.

[14] E. Raveh, I. Gessinger, S. Le Maguer, B. Möbius, and I. Steiner, "Investigating phonetic convergence in a shadowing experiment with synthetic stimuli," in *28th Conference on Electronic Speech Signal Processing (ESSV)*, Mar. 2017, pp. 254–261.

[15] 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.

[16] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[17] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Workshop on Speech Synthesis*, Sep. 2002, pp. 227–230.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," Version 6.0.25, retrieved 11 February 2017 from http://www.praat.org/.