INVESTIGATING PHONETIC CONVERGENCE IN A SHADOWING EXPERIMENT WITH SYNTHETIC STIMULI

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

[1] *Universität des Saarlandes,* [2] *DFKI GmbH*
raveh@coli.uni-saarland.de

Abstract: This paper presents a shadowing experiment with synthetic stimuli, whose goal is to investigate phonetic convergence in a human-computer interaction paradigm. Comparisons to the results of a previous experiment with natural stimuli are made. The process of generating the synthetic stimuli, which are based on the natural ones, is described as well.

1 Introduction

Phonetic convergence as one form of inter-speaker accommodation is defined as an increase in segmental and suprasegmental similarities between two speakers [1]. So far, phonetic convergence has been observed in human-human interaction (HHI) [2, 3], but has received little attention in the field of human-computer interaction (HCI). The existence of phonetic convergence in HCI may play a decisive role in the further development of spoken dialogue systems. For example, it could help improve the adaptive capabilities of such systems and contribute to the overall fluency and naturalness of the dialogue.

In a previous shadowing experiment [4], natural stimuli were used to investigate phonetic convergence in HHI. The analysis focused on the following three segmental features which show variation across native speakers of German: realization of the vowel *-ä-* in stressed syllables as [ɛː] or [eː], realization of final syllables ending with *-ig* as [ɪç] or [ɪk], and elision or epenthesis of [ə] in final syllables ending with *-en*. Participants showed a higher degree of convergence for [ɛː] vs. [eː] and [ɪç] vs. [ɪk] than for elision or epenthesis of [ə]. However, the overall degree of convergence varied considerably among the participants.

This paper presents a second shadowing experiment attempting to replicate these findings using synthetic stimuli in the same setting. The results of the experiment will be compared with the results of the previous experiment. This will shed light on the question whether the convergence effect found in HHI can occur in human-computer interaction (HCI) as well. The synthetic stimuli were generated using diphone synthesis [5]. In order to prevent prosodic characteristics like stress and intonation from influencing the perception of the sentences, the f_0 contours and segment durations of the natural stimuli were provided to the system as input for the synthesis. By doing so, the cognitive load does not increase due to differences in prosodic characteristics. Instead, the listener's perception should be influenced only by the natural or synthetic source of the speech stimuli.

2 Experiment

The experimental procedure was similar to the experiment described in [4], but with the difference of using computer-synthesized stimuli in the shadowing task instead of natural ones.

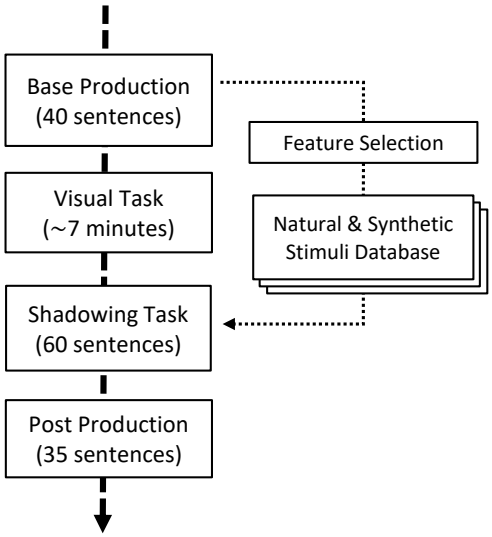


Figure 1 – Workflow of the experiment showing its four phases. The stimuli presented in the shadowing task are selected based on feature realizations in the baseline production.

Table 1 presents the examined target features. These sentences were presented to the participants in three tasks: In a written form (baseline production), as audio stimuli (shadowing task), and again in a written form (post production). Figure 1 summarizes the flow of the experiment. The stimuli used in the shadowing task were selected from the stimuli database so that they contained feature realizations of the opposite category (cf. Table 1) than the one produced by the participant in the baseline production. 14 female participants (19-50 years old, mean = 26) and 4 male participants (23-34 years old, mean = 27) took part in the experiment.

Table 1 – Examples of sentences containing the target features. Five sentences for each target feature were used in the experiment (filler sentences were added).

sentence	**target feature**
Die Best**ä**tigung ist für Tanja.	[ɛː] vs. [eː]
Ich bin süchti**g** nach Schokolade.	[ɪç] vs. [ɪk]
Wir begleite**n** dich zur Taufe.	[ən] vs. [n̩]

3 Synthetic Stimuli

For the current experiment, synthetic stimuli were created. There are various methods to synthesize speech: formant synthesis [6], diphone synthesis [7], unit selection [8], and hidden Markov model (HMM)-based synthesis [9], to name some. Diphone synthesis was chosen for generating the stimuli of this experiment, mainly for its direct control over the crucial synthesis parameters needed for this experiment. The stimuli were generated using MBROLA[1] [5]. The voice *de6* was used for the male stimuli, and the voice *de7* was used for the female stimuli.

The experiment examines convergence of specific segment-level phonetic features. It is preferable to keep other speech characteristics like intonation and stress unchanged, in order to prevent them from influencing the listeners' perception of the sentences. To achieve this, the f_0

[1] `http://tcts.fpms.ac.be/synthesis/mbrola.html`

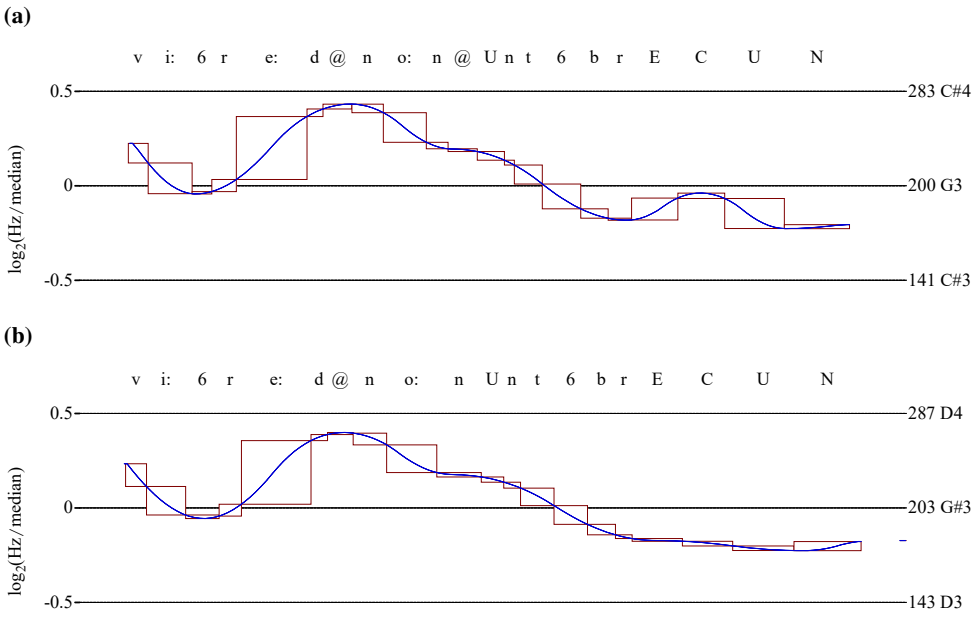


Figure 2 – The MOdeling MELody (MOMEL)-INternational Transcription System for INTonation (INTSINT) contours[2] of the natural stimulus (2a) and its corresponding MBROLA synthetic stimulus (2b) for the sentence *"Wir reden ohne Unterbrechung"*. The numeric and digit-numeric values to the right are the absolute pitch frequencies and their corresponding musical note; the scale to the left always displays one octave around the median pitch of the signal. The segments are labeled using the SAMPA annotation.

contours and segment durations of the natural stimuli were extracted from the natural stimuli (male and female, respectively) and imposed on the synthetic stimuli. The segment durations were taken directly from the annotations; the f_0 contours were acquired by measuring the f_0 at the beginning and in the middle of each segment. These three values per segment were then used as input for the synthesis models. However, because of the synthesis process, the generated segment durations and contours were not *completely* identical to those of the corresponding natural stimuli. Nonetheless, no substantial differences in overall sentence intonation or stress were introduced. An example for such a comparison is shown in Figure 2 (the f_0 contour was calculated using MOMEL and INTSINT [10]). The similarity between the all and synthetic contours was evaluated using objective methods using windows of length 5 ms. The root mean square error (RMSE) as well of the voicing error rate (VER) for the all averaged male and female contour comparisons are presented in Table 2.

Table 2 – Contour comparison between male and female voices of the synthetic and natural.

	RMSE (Hz)	**VER** (%)
Female	11.5	10.7
Male	6.2	11.0

To check whether the target features are salient in the synthetic stimuli and comparable to the natural stimuli, the signals of all three features were examined visually and acoustically. For

[2]Figure generated using ProZed plugin for Praat by Daniel Hirst.

the non-categorical target features [ɛː] vs. [eː] and [ən] vs. [n̩], objective comparisons between the natural and synthetic instances were also carried out. The formant areas occupied by [ɛː] and [eː] productions do not overlap and are linearly separable. This is true for both the natural and the synthetic condition (see Figure 6). However, the male and female productions of the same vowel are considerably closer to one another for the natural than for the synthetic condition. This means that the natural productions occupy a smaller area, forming a more stable target for convergence.

The natural and synthetic [ə] segment durations were also compared, to make sure that the synthesis process didn't introduce any significant differences. Since these segments' durations were imposed based on the duration of the corresponding segments of the natural stimuli, such differences were not expected. Indeed, only a very small difference was found (see section 4.3).

4 Results

The three target features were analyzed as described in [4]. To put the results of the synthetic stimuli into context, they are compared to the results of the natural stimuli from the previous experiment.

4.1 [ɛː] vs. [eː]

Each target vowel's first and second formants were measured at their temporal mid-point in the subjects' productions and the synthetic stimuli. These values were then plotted separately in four sub-groups which divide the values both by participants' preference with respect to [ɛː] and [eː] and gender of the shadowed stimuli (see Figure 3). The instances were grouped based on the condition in which they were produced (base, shadow, post), along with the instances of the synthetic stimuli. Euclidean distance in formant space was used for measuring how close the participants' vowel productions are to the mean production of the respective synthetic voice in each condition (see Table 3).

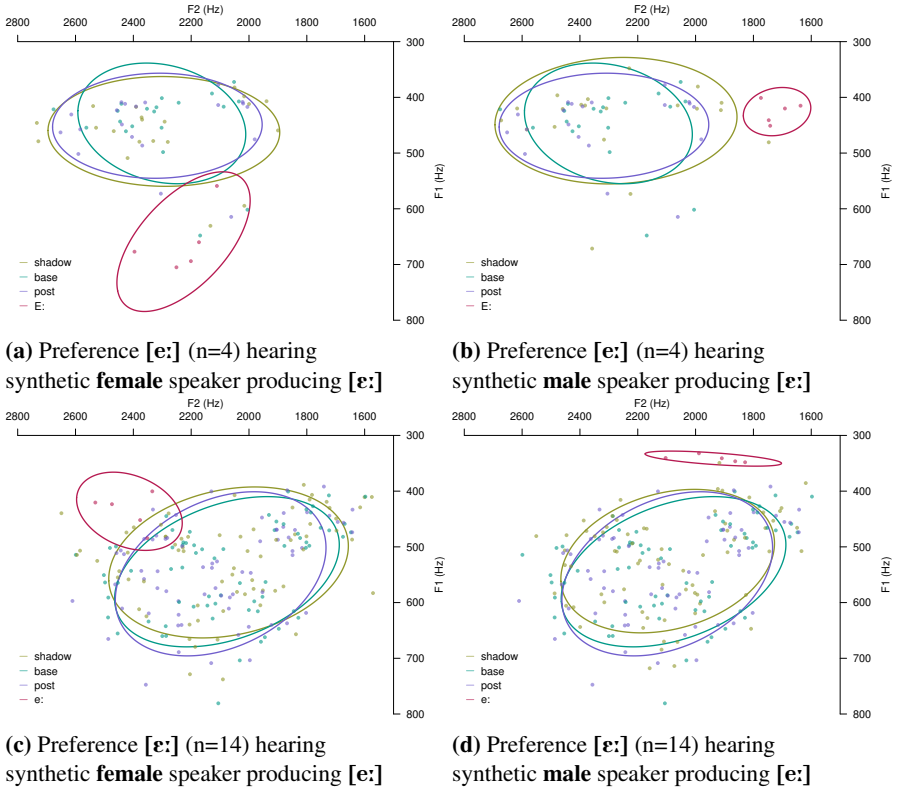
Table 3 – Euclidean distance (in Hz) between participant productions and mean production of the respective synthetic voice. Shown are the mean and standard deviation for the groups in each condition.

Group	**Base**		**Shadow**		**Post**	
	mean	sd	mean	sd	mean	sd
pref. [ɛː] / syn. female	396	199	397	220	372	197
pref. [ɛː] / syn. male	333	142	323	135	335	142
pref. [eː] / syn. female	281	87	308	110	311	84
pref. [eː] / syn. male	589	175	562	257	603	223

Statistical analysis of the four sub-groups with a Linear Mixed-Effects Model (LMM) including random intercepts and random slopes, did not show significant differences between any of the conditions. For the natural stimuli, an equivalent analysis of two sub-groups yielded significant differences between baseline and shadowing condition with $p < 0.02$ for the group of participants with preference [ɛː] and $p < 0.002$ for the group of participants with preference [eː].

4.2 [ɪç] vs. [ɪk]

The number of target [ɪç] and [ɪk] occurrences in the participants' shadowing productions were counted. To keep the feature categorical, variations of these sounds were counted as well, such as [ɪʃ] for the former or [ɪg] for the latter. Cases in which no convergence could have taken



(a) Preference **[eː]** (n=4) hearing synthetic **female** speaker producing **[ɛː]**

(b) Preference **[eː]** (n=4) hearing synthetic **male** speaker producing **[ɛː]**

(c) Preference **[ɛː]** (n=14) hearing synthetic **female** speaker producing **[eː]**

(d) Preference **[ɛː]** (n=14) hearing synthetic **male** speaker producing **[eː]**

Figure 3 – Comparison across male and female participants with **[ɛː]** or **[eː]** preference. The graphs show all **baseline** and **post** productions of the respective group. Only **shadow** production opposite to the plotted model are shown, as well as the productions of the **model** itself.

place were excluded. Thus, if a participant produced at most two instances of one form during the baseline phase (but still showed overall tendency toward the other form), these productions were excluded from the analysis. Following this principle, 28 out of 180 productions in the shadowing phase were excluded (16 %).

In total, convergence took place in 39 % of the possible cases (see Figure 4). For comparison, the result for this feature under the natural stimuli condition was 34 %.

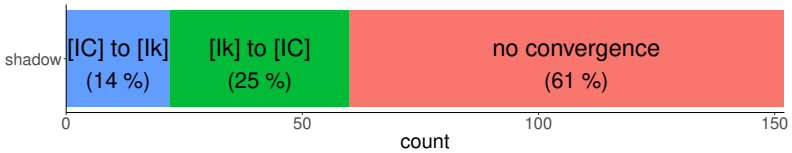


Figure 4 – Target feature [ɪç] vs. [ɪk]: Productions of all participants (n = 18) divided into three groups, indicating whether and how they converged to the synthetic stimuli productions of the feature.

4.3 [n̩] vs. [ən]

Potential schwa segments were divided into three groups: first, productions where the segment's duration was 30 ms or longer (clearly perceptible); subsequently, vowel-transition segments shorter than 30 ms (hardly perceptible); and last, immediate transition (not audible), which

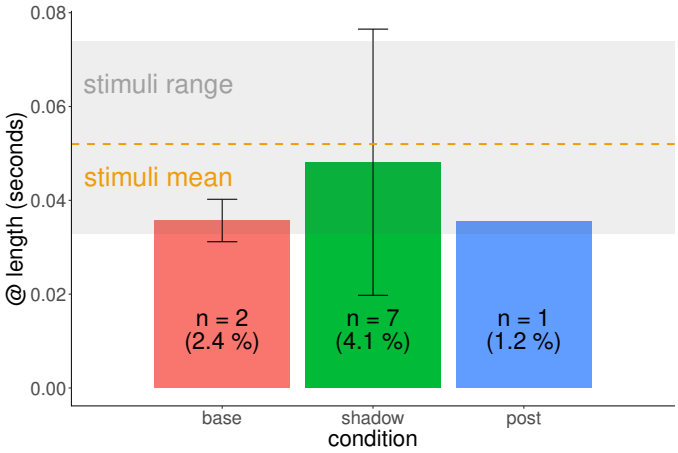


Figure 5 – Target feature [ŋ] vs. [ən]: Occurrences of [ə] in all participants (n = 17) and their respective mean duration per condition. The duration of [ə] in the synthetic stimuli ranged from 33 ms to 74 ms (mean 52 ms).

counted as zero-duration segments. Only the productions of the first group were counted as occurrences of schwa. As mentioned above, the segment durations of the synthetic stimuli were taken from those of the natural stimuli.

As expected, very rarely did the participants pronounce the schwa in a final syllable ending with *-en* in the baseline phase. However, as in the case of the natural stimuli, more instances of schwa were produced in the shadowing phase (see Figure 5). The duration range of schwa in the synthetic stimuli was 33 ms to 74 ms (average 52 ms), compared to 30 ms to 69 ms (average 48 ms) in the natural stimuli.

5 Discussion and Conclusion

The results of the experiment show different degrees of convergence for the target features. Regarding the feature [ɪç] vs. [ɪk], the number of cases in which convergence occurred is 5 % higher for the synthetic stimuli than in the preceding experiment using natural stimuli. Regarding the feature [ɛː] vs. [eː], the conclusion is less decisive. It is important to note that the distance between the male and female productions of [ɛː] as well as [eː] were larger for the synthetic stimuli than for the natural ones (see Figure 6). To quantify these distances, the deltas between the complete link (i.e. two most distant inter-group instances) and the single link (i.e. two closest inter-group instances) of each of these area pairs were calculated. These deltas are 431 and 468 for the [ɛː] and [eː] productions in the synthetic stimuli, and 303 and 155 for the productions in the natural stimuli. This results in a much larger overall target area for convergence, which might lead to a more spread effect – even in cases where convergence does take place. While keeping that in mind, an effect of convergence could not be statistically shown for the synthetic stimuli. Finally, [ə] elision was the most persistent feature, with a difference between baseline and shadowing condition of 8.6 % in the previous experiment compared to only 1.7 % in this experiment.

HCI via speech is becoming increasingly common in everyday life, with an ever-growing number of systems capable of interacting that way in different areas and for various purposes.

All in all, the results of the experiment are comparable to those of the preceding experiment with the natural stimuli. Additional experiments using different synthesis methods may shed light on the question whether the degree of convergence depends on specific characteristics and

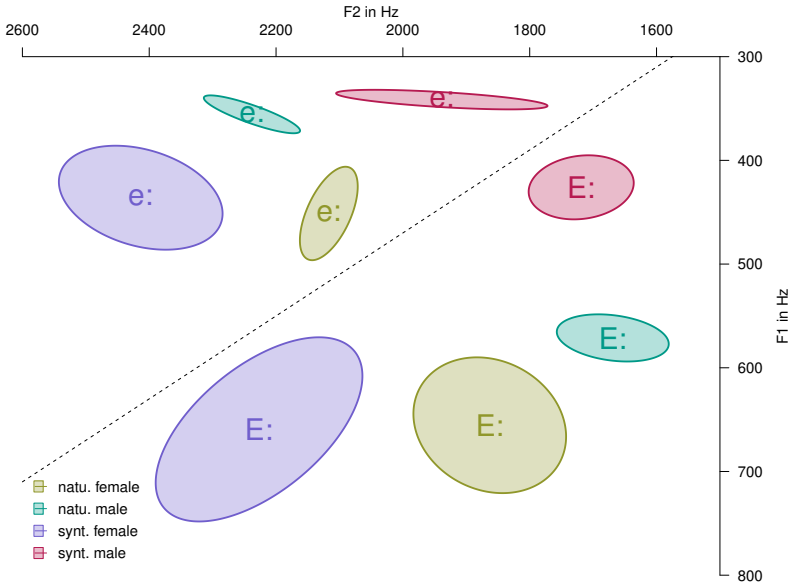


Figure 6 – Linearly separated formant areas occupied by [ɛː] and [eː] productions, respectively, for the natural model speakers (*natu. female* and *natu. male*) and the synthetic voices (*synt. female* and *synt. male*). The male and female productions of the same vowel are considerably closer to one another for the natural than for the synthetic condition. The vowels are labeled using the SAMPA annotation.

overall naturalness of the stimuli. Still, the participants of this experiment apparently did not judge the synthetic stimuli as entirely acceptable. They rated the voices as rather unnatural with an average of 3.5 on a 8-point Likert-scale for the male voice and only 2.5 for the female voice (with 1 being "very unnatural" and 8 being "very natural").

Overall, the results of the experiment are comparable to those of the preceding experiment with the natural stimuli. Additional experiments using different synthesis methods may shed light on the question whether the degree of convergence depends on specific characteristics and overall naturalness of the stimuli. It might also be the case that humans generally show less tendency to converge when they know they are talking to a computer. Therefore, synthesized speech would have to sound very natural for humans to show phonetic convergence like they do with human interlocutors. In any case, the results presented here show that there is evidence for convergence by humans while listening to synthetic voices.

6 Future Work

To investigate the influence of the characteristics of synthetic stimuli on the degree of convergence, another variant of the experiment is planned, this time using synthetic stimuli generated using another method, namely HMM synthesis.

As it can be assumed that introducing convergence capabilities on the computer's side could contribute to more fluent and natural communication, we also plan to examine this phenomenon in a more interactive environment, such as a spoken dialogue system (SDS). Widening the setting to an interactive dialogue rather than shadowing stimuli could also trigger additional effects aside from the segmental changes found in this experiment, like lexical convergence [11] or more variation in prosody.

References

[1] PARDO, J. S.: *On phonetic convergence during conversational interaction*. *Journal of the Acoustical Society of America*, 119(4), pp. 2382–2393, 2006. doi:10.1121/1.2178720.

[2] LEWANDOWSKI, N.: *Talent in nonnative phonetic convergence*. Ph.D. thesis, Universität Stuttgart, 2012.

[3] BABEL, M., G. MCGUIRE, S. WALTERS, and A. NICHOLLS: *Novelty and social preference in phonetic accommodation*. *Laboratory Phonology*, 5(1), pp. 123–150, 2014.

[4] GESSINGER, I., E. RAVEH, J. O'MAHONY, I. STEINER, and B. MÖBIUS: *A shadowing experiment with natural and synthetic stimuli*. In *Phonetik & Phonologie 12*, pp. 58–61. Munich, Germany, 2016.

[5] DUTOIT, T., 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 *International Conference on Spoken Language Processing (ICSLP)*, vol. 3, pp. 1393–1396. 1996. doi:10.1109/ICSLP.1996.607874.

[6] HÖGBERG, J.: *Data driven formant synthesis*. In *5th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 565–568. 1997. URL `http://www.isca-speech.org/archive/eurospeech_1997/e97_0565.html`.

[7] HAMON, C., E. MOULINE, and F. CHARPENTIER: *A diphone synthesis system based on time-domain prosodic modifications of speech*. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 238–241. 1989. doi:10.1109/ICASSP.1989.266409.

[8] HUNT, A. J. and A. W. BLACK: *Unit selection in a concatenative speech synthesis system using a large speech database*. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 373–376. 1996. doi:10.1109/ICASSP.1996.541110.

[9] ZEN, H., K. TOKUDA, and A. W. BLACK: *Statistical parametric speech synthesis*. *Speech Communication*, 51(11), pp. 1039–1064, 2009. doi:10.1016/j.specom.2009.04.004.

[10] HIRST, D.: *A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation*. In *International Conference of Phonetic Sciences (ICPhS)*, pp. 1233–1236. 2007. URL `http://icphs2007.de/conference/Papers/1443/`.

[11] LOPES, J., M. ESKENAZI, and I. TRANCOSO: *Towards choosing better primes for spoken dialog systems*. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 306–311. 2011. doi:10.1109/ASRU.2011.6163949.