NO EVIDENCE FOR CONVERGENCE TO SUB-PHONEMIC F2 SHIFTS IN SHADOWING

Orhun Uluşahin[1], Hans Rutger Bosker[1,2], James M. McQueen[1,2], Antje S. Meyer[1,2]

[1] Max Planck Institute for Psycholinguistics, [2] Donders Institute for Brain, Cognition and Behaviour
orhun.ulusahin@mpi.nl, hansrutger.bosker@donders.ru.nl, james.mcqueen@donders.ru.nl, antje.meyer@mpi.nl

ABSTRACT

Over the course of a conversation, interlocutors sound more and more like each other in a process called convergence. However, the automaticity and grain-size of convergence are not well established. This study therefore examined whether female native Dutch speakers converge to large yet sub-phonemic shifts in the F2 of the vowel /e/. Participants first performed a short reading task to establish baseline F2s for the vowel /e/, then shadowed 120 target words (alongside 360 fillers) which contained one instance of a manipulated vowel /e/ where the F2 had been shifted down to that of the vowel /ø/. Consistent exposure to large (sub-phonemic) downward shifts in F2 did not result in convergence. The results raise issues for theories which view convergence as a product of automatic integration between perception and production.

Keywords: *Phonetic convergence, sub-phonemic variation, alignment, F2, speech shadowing.*

1. INTRODUCTION

During conversation, we constantly switch roles between speaker and listener, shifting between cognitive tasks. Despite this dynamism of natural conversation, language perception and production have historically been studied in isolation, and the resulting theoretical frameworks have largely focused on only one of these primary modes of language processing. Consequently, recent research has called for more comprehensive unified theories of language perception and production which can better explain how perception and production interact [1], [2].

One possible avenue for investigating the interaction between perception and production is afforded by the phenomenon of phonetic convergence, which broadly refers to the tendency of interlocutors to sound more like each other over time. It has been observed both in laboratory settings and in natural conversation [3], and has been operationalized as increasing similarity between interlocutors' speech rates [4], vowel formants [5], and VOTs [6] among other acoustic features. However, the automaticity of the mechanisms that underlie convergence and the sensitivity of these mechanisms to sub-phonemic variation is unclear. One account asserts that alignment between interlocutors (i.e., the gradual establishment of common ground in linguistic representations on the syntactic, lexical, and phonological level) is automatic, and that phonetic convergence may be a natural product of this alignment process [7], [8]. Conversely, an alternative account presents evidence for the influence of various social factors on convergence [3], [9]–[12]. However, much of the previous empirical work on convergence presents evidence based on a combination of acoustic measures for multiple (often covarying) acoustic properties and/or perceptual similarity measures [3]. Furthermore, sub-phonemic convergence to individual formants has only been documented for a restricted set of vowels using strictly acoustic measures [13]. Thus, the questions of whether interlocutors converge to phonemes or to acoustic features, and whether this occurs automatically, remain unanswered.

To investigate the automaticity of convergence to sub-phonemic variation, we created a unidimensional and low-level acoustic metric. Specifically, we measured convergence as a single formant shift (i.e., F2) on a single vowel (i.e., /e/) in trisyllabic words repeated in a shadowing task. We manipulated target words in the shadowing task such that the critical vowel /e/ had a substantially lower F2 than usual, around that of the Dutch vowel /ø/. Other formants and acoustic features, most importantly the F3 which also distinguishes the two vowels, were left untouched during the manipulations to conceal the aim of the experiment and prevent perceptual category shifts (i.e., manipulated vowels would still be perceived as /e/). The unidimensional measure therefore not only provided a strictly acoustic convergence metric, but also afforded extensive experimental control.

We tested whether female native speakers of Northern Standard Dutch converged to a sub-phonemic, unidimensional downward shift in the F2 of the vowel /e/ produced by a female native speaker. Our methodology employed a pre-test/post-test paradigm consisting of 3 tasks: (1) A reading task for obtaining participants' baseline F2 values for the vowels /e/ and /ø/, (2) a shadowing task in which participants repeated trisyllabic Dutch words containing the F2-downshifted critical vowel /e/

(henceforth referred to as /?/), and finally (3) a phonological 2-alternative forced-choice (2AFC) task where participants identified trisyllabic stimuli containing /e/, /ø/, or /?/ (i.e., the manipulated vowel) as containing /e/ or /ø/, to verify that our manipulations were indeed sub-phonemic. In line with theories which argue for automatic convergence, we hypothesized that participants' F2s for the vowel /e/ would decrease (i.e., converge in the direction of /?/) as a function of exposure to /?/ during the shadowing task, and that participants would still categorize manipulated (i.e., F2 downshifted) stimuli as /e/ in the 2AFC task.

2. METHOD

2.1. Participants

40 adult female native speakers of Northern Standard Dutch participated in the experiment. All participants reported normal hearing and normal or corrected-to-normal vision. We recruited only female participants to minimize the acoustic distance between the typical formants of our female speaker and the participants. All participants gave informed consent (Project code: ECSW-2019-019) and received monetary compensation.

2.2. Materials

We selected the vowels /e/ and /ø/ as reference points based on existing reports [14] of their mean formant values[1]. The low F1 difference between the two vowels and the large F2 difference enabled participants' convergence to be measured as a function of observed changes in their F2 relative to their baseline levels. Crucially, the F3 difference, which is essential for distinguishing the two vowels, allowed our manipulations to remain sub-phonemic.

Using CELEX [15], we selected 40 trisyllabic Dutch words as experimental items for the shadowing task. These words all had exactly one instance of /e/ (always in a stressed syllable), and no instances of /ø/ (e.g., *orchidee* [ɔrxiˈde] "orchid"). Moreover, replacing /e/ with /ø/ would render all experimental words non-words. Additionally, a set of 120 trisyllabic filler words was generated. The vowels /e/ or /ø/ did not occur in any of these filler words. Thus, a total of 160 words were selected for the shadowing task[1].

Another set of 20 trisyllabic words was selected from CELEX to be used in the reading task[1]. 10 of these words contained only one instance of the vowel /e/ in a stressed syllable and no vowel /ø/. The other 10 contained only one instance of the vowel /ø/ in a stressed syllable and no vowel /e/.

We recorded a female native speaker of Northern Standard Dutch. For the experimental words in the shadowing task (i.e., words containing /e/), we also recorded "/ø/ nonword versions" in which the vowel /e/ was replaced with the vowel /ø/ (e.g., [ɔrxiˈde] → [ɔrxiˈdø]). In this case, speech was elicited using prompts following Dutch orthography (e.g., *orchidee* → *orchideu*). We recorded these versions to measure a reference F2 value for our speaker's typical /ø/ for each word, thus assigning word-by-word "targets" for our F2 manipulations. The mean F2 and F3 measures (in the middle 50% of the vowels) in our speaker's /e/ and /ø/ recordings were comparable to the reference values in [14] [1].

The 40 experimental words were then manipulated. For each word, the F2 of the vowel /e/ was shifted to match the F2 of its "/ø/ version" based on our recordings (e.g., *orchidee*'s new F2 value was set to our speaker's F2 on *orchideu*). The formant manipulations were carried out using the Burg LPC method implemented in Praat [16]. The F1 and the F3 were left untouched. For each word, a lower-F2 version of the original vowel /e/, labelled /?/, was generated and recombined with the original signal. Measurements confirmed that our manipulations had the desired effect, resulting in an average F2 difference of 490 Hz between /e/ and /?/[1].

2.3. Procedure

After providing consent, participants entered a sound-attenuating room with a monitor, keyboard and microphone. The experiment was run using Presentation® (Version 22.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

The experiment began with the reading task. Participants saw trisyllabic Dutch words on screen and were asked to read them out loud. Each of the 20 words selected for the reading task was presented twice for a total of 40 trials. The words used in the reading task did not appear elsewhere in the experiment. On each trial, a random word was selected and presented on screen. After exactly 1 second, the caption *OPNAME* ("recording") appeared on screen, and participants had 3 seconds to read the word out loud.

After the reading task, participants performed three shadowing blocks. Each block contained a randomized combination of all 40 experimental words and all 120 filler words. In each trial, participants saw a fixation cross while listening to the auditory stimulus for the duration of the whole word (i.e., about 1 second). At word offset, the fixation cross was replaced with the caption *OPNAME*, indicating the 3-second-long response period for each trial. We instructed participants to "repeat [the words]

as quickly and as accurately as possible" without explicitly demanding that they imitate any unusual features they might hear in some of the trials.

Finally, participants performed a 2AFC task in which they categorized stimuli as either containing the vowel /e/ or /ø/. This task used the 40 experimental words from our speaker that were initially recorded for the shadowing task. However, it also incorporated the alternative recordings in which the speaker had replaced the vowel /e/ with the vowel /ø/ (e.g., nonword *orchideu*), as well as F2-downshifted versions of each word. Thus, any item could be heard in any of the three forms. Four evenly distributed pseudo-randomized trial sets each contained a total of 40 trials, and each trial contained only one utterance of one critical vowel (e.g., a trial could feature the word *orchidee, orchideu* or *orchid?* where "?" denotes the manipulated vowel). Of the 40 words in each set, 8 contained the vowel /e/, 16 contained the vowel /ø/, and another 16 contained /?/. We used a higher proportion of /ø/ and manipulated words to gather more data that would reveal whether our manipulations triggered a perceptual category shift (i.e., /?/ perceived as /ø/).

During the task, participants saw two options in written form on screen (i.e., the /e/ spelling as in *orchidee* and the /ø/ spelling as in *orchideu*) upon word onset. They were instructed to indicate which written form represented the word they were hearing using keyboard controls.

2.4. Pre-processing

We used the WebMaus webservice [17] with the "nld-NL" parameters for the automatic word-level and phoneme-level parsing of the tokens recorded during the experiment. Following a data-driven approach for formant analysis (e.g., [18]), participants' audio recordings were analyzed with two sets of parameters: 1) 4 maximum formants at a 5 KHz range, 2) 5 maximum formants at a 5.5 KHz range. For all participants, the reliability of either set of parameters was evaluated as a function of F2 variability across experimental trials for a given participant (the greater the variability, the less accurate the tracking) as well as manual inspection of the estimated formant tracks in the spectrograms of 10 items per participant. All subsequent processing of a given participant's acoustic data was carried out using the best performing set of parameters. All formant values were extracted from the middle 50% of vowels.

3. RESULTS

6400 audio files ((120 experimental shadowing trials + 40 reading trials) * 40 participants) were subjected to the pre-processing steps identified above. We excluded the data from one participant who consistently produced F2s that were higher than 2400 Hz. Thus, formant values from 6240 files were used.

The participants' baseline F2s for the vowels /e/ (i.e., 2271 Hz) and /ø/ (i.e., 1842 Hz) were comparable to reports of typical Northern Standard Dutch [14] and our speaker's recordings.

We used a linear mixed-effects model [19] with the "lmerTest" package [20] in R [21] to test whether exposure to manipulated stimuli significantly predicted F2 convergence. With F2 (Hz) as the dependent variable, we included fixed effects of Task (Reading and Shadowing, the latter mapped onto the intercept) and Trial Number (z-scored) as predictors, and their interaction. Finally, we entered Word and Participant Number as random factors.

We found no effect of Task (β = -16.302, SE = 19.976, t = -0.816, p = 0.36) or Trial Number (β = -0.291, SE = 0.319, t = -0.915, p = 0.42) on F2 frequency. We also did not observe any significant interaction between these two factors (β = 0.3096, SE = 0.3183, t = 0.973, p = 0.33). The lack of an effect of Trial Number suggests that exposure to manipulated stimuli did not significantly predict any gradual change in F2 in either task. The lack of an effect of Task suggests that participants' F2s were not different overall (i.e., compared to the "baseline" reading measurement).

To analyze the output of the 2AFC task, we coded responses where participants categorized /?/ as /ø/ as "critical" given that they might indicate a perceptual category shift. Critical responses constituted only 1.56% of responses to all trials in which /?/ was used (*N* = 640, 16 trials with /?/ * 40 participants). Thus, the results from the 2AFC task suggested that, despite our large F2 shifts, we succeeded in avoiding perceptual category shifts.

4. DISCUSSION

The present study investigated whether consistent exposure to a manipulated (i.e., sub-phonemically downshifted) vowel formant (i.e., F2) influenced participants' speech production and resulted in phonetic convergence in a shadowing task (i.e., lower F2s for the vowel /e/ in participants' speech). We had hypothesized that participants would gradually shift their F2s for the vowel /e/ towards the *targets* represented by our manipulated stimuli, as a function of increased exposure. This hypothesis was in line with a view of convergence as an automatic process which would be sensitive to (consistent) sub-phonemic variation irrespective of participants' awareness of the F2 manipulation. Although a few participants (*n* = 4) followed a pattern of consistent

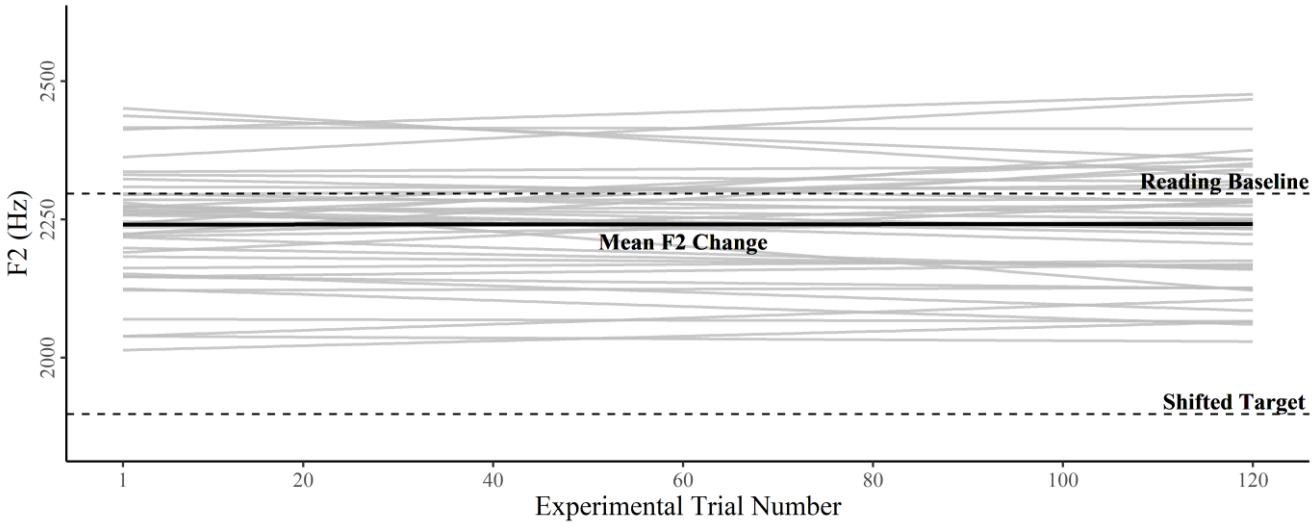

Figure 1: Each solid grey line follows the mean F2 trend of a participant. The solid black line follows the mean across participants. The dashed lines indicate, respectively, participants' baseline F2 values for /e/ and the speaker's average F2 in experimental items containing /?/ ("Shifted Target"). Trial numbers on the x-axis represent the sequence of the experimental trials (i.e., excluding fillers) across all 3 blocks.

downward shifts in F2 across all trials, most participants' F2s remained consistent across trials with some participants even displaying a marginal increase in F2 across all trials (see Figure 1). Although numerically there was a small overall downward shift in F2 from the reading task to the shadowing task across all participants, this difference was not significant.

The results of the phoneme categorization task suggest that we successfully avoided perceptual category shifts. However, our design did not include a task which could directly measure whether any perceptual adjustments occurred in participants' category boundary between /e/ and /ø/. We had theoretical and empirical precedent for assuming that exposure to our manipulated stimuli would indeed result in adjustments in the perceptual stream. Our F2 manipulations were substantially larger than the just noticeable difference for F2 (i.e., our manipulations resulted in a minimum of 20% reduction in F2, with the just noticeable difference reported at 1.5% [22]), and there is evidence that perceptual adjustments to consistent novel stimuli are not only common but also quickly formed and tracked [23]. Future research may introduce intermittent or post-test perceptual measures into a similar design to address the question whether perceptual adjustments to F2 manipulations do occur but fail to translate into production, or whether a sub-phonemic F2 shift alone fails to facilitate perceptual adjustments in the first place.

Although the absence of perceptual category shifts due to the sub-phonemic nature of the acoustic manipulations was an explicit goal of our design, the phonological non-salience of the manipulations (i.e., the manipulations did not introduce comprehension challenges) might also have been a factor behind the lack of evidence for convergence. Nevertheless, it is worth noting that convergence has been reported for other acoustic features of incoming speech that are sub-phonemic such as speech rate [4], and F0 in non-tone languages [24]. Thus, the influence of the non-salience of the manipulation must be evaluated in combination with other factors.

Finally, recent research [25] has highlighted that task engagement can play an important role in facilitating phonetic convergence. Our experimental paradigm, given its non-dialogic setting and the high number of repetitive trials, may have reduced our participants' level of engagement. Given the present study's success in concealing large sub-phonemic F2 shifts (i.e., /?/ was almost always perceived as /e/), future work may increase the experimental word ratio, while retaining sufficient experimental control to be able to apply precise sub-phonemic acoustic manipulations and maintain perceived naturalness.

Given these limitations, the lack of evidence for convergence in F2 despite consistent exposure to large F2 shifts in a consistent target (i.e., the vowel /e/) suggests that convergence in F2 might be dependent on covarying acoustic factors and resulting category shifts. Alternatively, our manipulations may have resulted in perceptual adjustments in our participants without producing observable differences in production. Nevertheless, the lack of evidence for convergence to large sub-phonemic shifts presents a potentially interesting issue for comprehensive theories of language perception and production that claim that there is automatic integration between these two primary modes of language cognition [1], [2], [7].

7. REFERENCES

[1] M. J. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behav. Brain Sci.*, vol. 36, no. 4, pp. 329–347, Aug. 2013, doi: 10.1017/S0140525X12001495.

[2] A. S. Meyer and J. M. McQueen, "Key Issues and Future Directions: Toward a Comprehensive Cognitive Architecture for Language Use," in *Human Language: From Genes and Brains to Behavior*, P. Hagoort, Ed. The MIT Press, 2019. doi: 10.7551/mitpress/10841.001.0001.

[3] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Atten. Percept. Psychophys.*, vol. 79, no. 2, pp. 637–659, Feb. 2017, doi: 10.3758/s13414-016-1226-0.

[4] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evol. Hum. Behav.*, vol. 34, no. 6, pp. 419–426, Nov. 2013, doi: 10.1016/j.evolhumbehav.2013.08.001.

[5] A. Schweitzer and N. Lewandowski, "Social factors in convergence of F1 and F2 in spontaneous speech," in *Proceedings of the 10th International Seminar on Speech Production, ISSP 2014*, 2014, pp. 391–394. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85039164612&partnerID=40&md5=f2bfd3147d28baabd4e03c1b7891ecea

[6] K. Nielsen, "Specificity and abstractness of VOT imitation," *J. Phon.*, vol. 39, no. 2, pp. 132–142, Apr. 2011, doi: 10.1016/j.wocn.2010.12.007.

[7] S. Garrod and M. J. Pickering, "Why is conversation so easy?," *Trends Cogn. Sci.*, vol. 8, no. 1, pp. 8–11, Jan. 2004, doi: 10.1016/j.tics.2003.10.016.

[8] M. J. Pickering and S. Garrod, "Alignment as the Basis for Successful Communication," *Res. Lang. Comput.*, vol. 4, no. 2, pp. 203–228, Oct. 2006, doi: 10.1007/s11168-006-9004-0.

[9] J. S. Pardo, R. Gibbons, A. Suppes, and R. M. Krauss, "Phonetic convergence in college roommates," *J. Phon.*, vol. 40, no. 1, pp. 190–197, Jan. 2012, doi: 10.1016/j.wocn.2011.10.001.

[10] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, 2012, doi: 10.1016/j.wocn.2011.09.001.

[11] M. Babel, "Dialect divergence and convergence in New Zealand English," *Lang. Soc.*, vol. 39, no. 4, pp. 437–456, Sep. 2010, doi: 10.1017/S0047404510000400.

[12] A. Schweitzer and N. Lewandowski, *Social Factors in Convergence of F1 and F2 in Spontaneous Speech*. 2014. doi: 10.13140/2.1.3709.5689.

[13] S. Tilsen, "Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production," *J. Phon.*, vol. 37, no. 3, pp. 276–296, Jul. 2009, doi: 10.1016/j.wocn.2009.03.004.

[14] P. Adank, R. van Hout, and R. Smits, "An acoustic description of the vowels of Northern and Southern Standard Dutch," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1729–1738, Sep. 2004, doi: 10.1121/1.1779271.

[15] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX Lexical Database (CD-ROM)," 1996, Accessed: Jun. 18, 2022. [Online]. Available: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2339741

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Nov. 15, 2022. [Online]. Available: http://www.praat.org/

[17] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Comput. Speech Lang.*, vol. 45, pp. 326–347, Sep. 2017, doi: 10.1016/j.csl.2017.01.005.

[18] P. Escudero, P. Boersma, A. S. Rauber, and R. A. H. Bion, "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1379–1393, Sep. 2009, doi: 10.1121/1.3180321.

[19] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *J. Mem. Lang.*, vol. 59, no. 4, pp. 390–412, Nov. 2008, doi: 10.1016/j.jml.2007.12.005.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest Package: Tests in Linear Mixed Effects Models," *J. Stat. Softw.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[21] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. [Online]. Available: https://www.R-project.org/

[22] D. Kewley-Port and C. S. Watson, "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.*, vol. 95, no. 1, pp. 485–496, Jan. 1994, doi: 10.1121/1.410024.

[23] X. Zhang and L. L. Holt, "Simultaneous tracking of coevolving distributional regularities in speech.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 44, no. 11, p. 1760, 20181001, doi: 10.1037/xhp0000569.

[24] M. Simonet, "Intonational convergence in language contact: Utterance-final F0 contours in Catalan-Spanish early bilinguals," *J. Int. Phon. Assoc.*, vol. 41, no. 2, pp. 157–184, 2011, doi: 10.1017/S0025100311000120.

[25] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Commun.*, vol. 138, pp. 50–66, Mar. 2022, doi: 10.1016/j.specom.2022.02.002.

[1] See supplementary material (formant settings used for each participant, mean reference and speaker formant values, a list of stimuli, and a log of F2 manipulations) at https://osf.io/wb2mn/