Speech Communication 127 (2021) 43–63

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing

Iona Gessinger [a,*], Eran Raveh [a], Ingmar Steiner [b], Bernd Möbius [a]

[a] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany
[b] audEERING GmbH, Gilching, Germany

ARTICLE INFO

Keywords:
Phonetic accommodation
Speech shadowing
Human-human interaction
Human-computer interaction
Synthetic speech
German

ABSTRACT

The present study investigates whether native speakers of German phonetically accommodate to natural and synthetic voices in a shadowing experiment. We aim to determine whether this phenomenon, which is frequently found in HHI, also occurs in HCI involving synthetic speech. The examined features pertain to different phonetic domains: allophonic variation, schwa epenthesis, realization of pitch accents, word-based temporal structure and distribution of spectral energy. On the individual level, we found that the participants converged to varying subsets of the examined features, while they maintained their baseline behavior in other cases or, in rare instances, even diverged from the model voices. This shows that accommodation with respect to one particular feature may not predict the behavior with respect to another feature. On the group level, the participants of the natural condition converged to all features under examination, however very subtly so for schwa epenthesis. The synthetic voices, while partly reducing the strength of effects found for the natural voices, triggered accommodating behavior as well. The predominant pattern for all voice types was convergence during the interaction followed by divergence after the interaction.

1. Introduction

In this paper we discuss *phonetic accommodation* in spoken interaction. The term *accommodation* is used in the sense of an adjustment to the circumstances. In the case of spoken interaction, this refers to adjustments made by a speaker as a reaction to being exposed to another speaker. As a consequence, the speech of the interlocutors may become more or less similar to each other. The former behavior is called *convergence* and the latter *divergence*.

Phonetic accommodation has been found in human spoken interaction (e.g., Pardo, 2006; Levitan and Hirschberg, 2011; Lewandowski, 2012) and has been linked to communicative success and dialog quality (e.g., Lee et al., 2010; Manson et al., 2013; Borrie et al., 2015). The present work contributes to the growing body of studies on phonetic accommodation which consider the fact that we are no longer communicating exclusively with human interlocutors (e.g., Levitan et al., 2016; Lubold et al., 2016; Beňuš et al., 2018) by including natural as well as synthetic voices into the experimentation. We intend to determine whether phonetic accommodation occurs to a similar extent in human-computer interaction (HCI) involving synthetic speech as it does in human-human interaction (HHI).

We present an analysis of accommodation with respect to a set of phonetic phenomena in a *speech shadowing* experiment. In this experiment, native speakers of German repeat short German sentences immediately after hearing them from a female or a male model speaker. The voices are either natural or synthetic. The overall assumption is that the participants accommodate their speech to that of the respective model speaker. More specifically, we expect to find convergence towards the model speakers. This expectation is motivated below (Section 1.4).

The phonetic phenomena under examination include allophonic variation and schwa epenthesis as segment-level phenomena, the realization of pitch accents as a phenomenon of local prosody, as well as the word-based temporal structure and distribution of spectral energy as measures of global similarity. Conducting an analysis of accommodation on such a diverse set of features pertaining to different phonetic domains allows for an extensive assessment of the participants' behavior. To the best of our knowledge, this is the first study that investigates the accommodation of such phenomena when shadowing short utterances and includes both natural and synthetic speech stimuli as accommodation targets.

The results are discussed at the group level with the aim to compare the accommodating behavior of humans towards natural and synthetic

* Corresponding author.
E-mail address: gessinger@coli.uni-saarland.de (I. Gessinger).

https://doi.org/10.1016/j.specom.2020.12.004
Received 15 December 2019; Received in revised form 13 November 2020; Accepted 8 December 2020
Available online 29 December 2020

speech. Since idiosyncratic variation in accommodation is often observed (e.g., Pardo et al., 2018), the results are further discussed for individual speakers.

This article summarizes and improves upon previous analyses of the allophonic variation and schwa epenthesis (Gessinger et al., 2017), as well as the pitch accent realization (Gessinger et al., 2018). In particular, it expands on them by including an additional phase of the experiment, namely the post production after the actual speech shadowing task, which allows to assess whether the accommodation effect is sustained. The analyses of the word-based temporal structure and distribution of spectral energy, as well as the individual behavior of the participants have not previously been published.

1.1. Theoretical frameworks

The phenomenon of inter-speaker accommodation is often assessed in the framework of the *Communication Accommodation Theory (CAT)*, a generalized model of communicative interaction (Giles, 1973; Giles et al., 1991; Shepard et al., 2001). The theory assumes interpersonal conversation to be a dynamic adaptive exchange of verbal and nonverbal behavior. During this exchange, the listener-speaker directs their attention to the speech of the interlocutor and adjusts their own speech as a way of reducing or increasing social distance to the interlocutor. According to CAT, convergence will therefore occur when social distance should be decreased — as opposed to divergence, which will occur when social distance should be increased. This suggests that accommodating behavior is socially motivated and, to some extent, consciously controlled by the speaker.

The *Interactive Alignment Model (IAM)* (Pickering and Garrod, 2004, 2013) represents another point of view, which is reflected in the use of the term *alignment* instead of *accommodation*. Where *accommodation* allows for both converging and diverging behavior, *alignment* necessarily leads to an increased similarity, hence convergence. The model postulates that it is a priming mechanism that leads to alignment between interlocutors during conversation. This suggests that converging behavior is an automatic process which is triggered subconsciously.

Both theories, as opposing as they might appear at first sight, concurrently agree that convergence is deeply rooted in human communicative behavior. They are not mutually exclusive, as both the social motivation and the automatic process can coexist and vary in dominance between individuals, which could partially explain the fact that different speakers exhibit different degrees of accommodation.

A model of phonetic accommodation combining the two accounts is likely to be a better approximation of the actual phenomenon than restricting oneself to either one of the theories (e.g., Krauss and Pardo, 2004; Babel, 2010; Coles-Harris, 2017).

The two accounts might further be weighted differently in different types of communicative interaction. Section 1.4 discusses their respective relevance for the present shadowing experiment.

1.2. Influence of the interlocutor

It has been shown that the interlocutor significantly influences the degree of phonetic accommodation exhibited by a speaker. A selection of influential factors that have been raised in this context is presented in this section.

First of all, the attitude towards the interlocutor in terms of their perceived attractiveness and likability can have an influence on accommodation. Schweitzer and Lewandowski (2014), for example, found a strengthened convergence effect for first and second vowel formants with increasing perceived likability of the interlocutor. In an analysis of measures related to fundamental frequency by Michalsky and Schoormann (2017), likability was predictive of convergence as well, but increasing perceived attractiveness was a stronger predictor. In Schweitzer et al. (2017), however, effects of convergence and divergence with respect to pitch accent realization were more pronounced when speakers disliked their respective interlocutor. These three studies treated conversational interaction. In a shadowing experiment evaluating similarity perceived by listeners, Babel et al. (2014) found that the effect of attractiveness on convergence only applied to female speakers.

Furthermore, the hierarchy between speaker and interlocutor can be relevant for phonetic accommodation. Results by Gregory and Webster (1996) analyzing long-term average spectra (LTAS) in conversational interaction suggest that speakers on the lower end of the hierarchy or in a less dominant role, converge to the hierarchically higher or more dominant interlocutor.

The aspect of social dominance may also play a role for the following factor, namely whether the speaker's sex matches that of the interlocutor (cf. Bilous and Krauss, 1988). Analyses of this factor have yielded varied outcomes. Levitan et al. (2012), for example, found more convergence of acoustic-prosodic features such as fundamental frequency, intensity, voice quality, and speaking rate in the conversational interaction of mixed-sex dyads as opposed to same-sex dyads. Bailly and Martin (2014), on the other hand, observed stronger convergence for same-sex dyads in an analysis of vowel spectra and global convergence as assessed by means of a speaker recognition technique.

A final factor, examined by Babel et al. (2014) as well, concerns the typicality of the interlocutor's voice. Typicality was quantified here by means of a speeded identification task to determine the ease of speaker sex classification as female or male. In the evaluation of similarity perceived by listeners, men showed convergence only towards the more atypical voices, whereas women converged to the typical voices as well.

Apart from the obvious differences between settings and features, it should be mentioned that these studies examined various languages situated in different societies whose influence, especially with respect to social factors, needs to be taken into account (German/Germany in Schweitzer and Lewandowski (2014), Michalsky and Schoormann (2017), Schweitzer et al. (2017), French/France in Bailly and Martin (2014), and English/USA in Gregory and Webster, 1996; Levitan et al., 2012; Babel et al., 2014).

All of these factors potentially influence the outcome of the present study.

For instance, the voice typicality factor can be quantified in various other ways than as the ease of female-male-classification. In the present study, a more prominent source of atypicality is the use of synthetic voices in the experiment. Synthetic voices emulate human voices, but we can assume that they are to some extent not typical of human voices. If there is an effect of atypicality promoting convergence for certain groups of speakers, the synthetic voices may be more successful here — unless there is a threshold of how atypical a voice can be before such an effect is inhibited or even reversed to divergence.

The hierarchical situation in a shadowing task is not clear. Does the model speaker who is shadowed by the participant play a dominant role? This may vary according to the perception of the participant. In any case, only the participant can accommodate to the model speakers in the present study, since the stimuli are prefabricated and merely played back to the participant during the shadowing task.

We collected simple scores for perceived naturalness and likability of the model voices from the participants. Information about the sex of participant and interlocutor is taken into consideration in the analysis of the data, where applicable.

When talking about the interlocutor in spoken interaction, we typically picture this to be another human. However, the interaction with computers via spoken language is becoming more and more integrated into our everyday life. Therefore, it is of valid interest to include this interlocutor type into considerations about phonetic accommodation.

It is certainly a non-trivial question whether the social aspects discussed so far hold for a virtual interlocutor as well. Starting with the social motivation advocated by CAT all the way to the factors mentioned in this section, it seems crucial for the interlocutor to be perceived as a social actor. Nass et al. (1994) claimed that the latter is indeed the case for computers, too. This concept has been established

as the *Computers are Social Actors (CASA)* paradigm (Reeves and Nass, 1996; Nass and Moon, 2000).

We can therefore assume that a virtual interlocutor should generally be able to trigger phonetic accommodation in a human speaker. This has been demonstrated to be the case for global prosodic features such as speaking rate and intensity (e.g., Bell et al., 2003; Oviatt et al., 2004; Suzuki and Katagiri, 2007). It has even been proposed that accommodation in the form of convergence may be stronger when communicating with a computer as compared to a fellow human, if it is the speaker's belief that this leads to greater communicative success (Branigan et al., 2010).

One component of HCI is the use of synthetic speech. Just as it is the case with human voices, there may be synthetic voices that are considered to be more or less attractive, likable, typical, natural, or dominant, leading to varying reactions on the side of the human interlocutor. In the present study, we use natural speech and two types of synthetic speech (diphone- and hidden Markov model (HMM)-based) to assess whether participants accommodate to them to the same or different degrees.

1.3. Experimental settings and phonetic features

The experimental settings in which phonetic accommodation has been observed include dynamic, conversational approaches (e.g., Pardo, 2006; Levitan and Hirschberg, 2011; Lewandowski, 2012; Schweitzer et al., 2017; Michalsky and Schoormann, 2017), and also less interactive tasks, such as consecutive speech shadowing, in which a participant repeats an utterance immediately after hearing it from a model speaker[1] (e.g., Shockley et al., 2004; Babel et al., 2014; Walker and Campbell-Kibler, 2015; Dias and Rosenblum, 2016; Pardo et al., 2017). The former qualify as fully social scenarios and therefore likely trigger the social motivation for accommodation. A shadowing task, on the other hand, generates a socially impoverished interaction during which the automatic motivation to accommodate might be predominant.

The present study uses a shadowing paradigm. However, unlike most shadowing experiments investigating accommodation, participants shadow full sentences instead of single, often mono- or bisyllabic (non-)words.

Babel (2012), for example, asked participants to repeat low frequency monosyllabic English words (e.g., *br**ee**ze*, *sm**a**sh*) "as clearly and naturally as possible" (p. 180) after a male model speaker and measured the difference in distance in the F1–F2 formant space. While testing various vowels, [æ] and [ɑ] showed the strongest convergence effect. A possible explanation offered by Babel (2012) is the fact that these low vowels vary regionally in North American English. This might have resulted in a greater a priori phonetic distance between participants and model speaker, hence more space for the participants to converge to the latter.

Dufour and Nguyen (2013) used bisyllabic French words ending in /e/ (e.g., *beaut**é***, *soir**ée***) or /ɛ/ (e.g., *proj**et***, *jam**ais***) and measured F1 to test whether speakers of Southern French, who usually produce both endings as [e], converge to a Standard French model speaker, who differentiates [e] and [ɛ]. After hearing a word from the female model speaker, participants were either asked to shadow it ("repeat it as naturally and as clearly as possible", p. 3) or to imitate it ("repeat it by imitating the speaker's specific pronunciation", p. 3). A convergence effect was found for both groups; however, it was stronger in the imitation group. For the shadowing group, it only occurred for words that had not been used in a pre-test, i.e., words participants heard for the first time during the shadowing task.

[1] To be distinguished from *close speech shadowing* where speech input is repeated while it is still ongoing.

In Mitterer and Müsseler (2013) participants repeated bisyllabic German words (e.g., ***sp**ielen*, ***St**unde*, *fert**ig***, *Käf**ig***) and non-words (e.g., ***sp**ümen*, ***str**eipen*, *ons**ig***, *wüss**ig***) "as quickly as possible" (p. 561) after a female model speaker to test the influence of being confronted with different phonetic implementations of the fricative-stop clusters, namely [ʃp]/[ʃt] vs. [sp]/[st], and the word ending ⟨-ig⟩, namely [ɪç] vs. [ɪk]. [ʃp]/[ʃt] and [ɪç] are the Standard German forms. [sp]/[st] are Northern German realizations of the fricative-stop clusters, while [ɪk] is a Southern German realization of the word ending ⟨-ig⟩. However, Mitterer and Müsseler (2013) state that the two variations "differ clearly in their markedness" (p. 560), with the fricative-stop cluster variation being undisputedly dialectal and the ⟨-ig⟩ variation having a rather unclear status. Both variations were imitated by the participants, with the more salient fricative-stop clusters showing a stronger effect. Most corrections occurred for the word ending ⟨-ig⟩ from stimulus [ɪç] to participant production [ɪk].

The fact that participants in the present study shadow full sentences moves the task from mere repetition slightly in the direction of conversational interaction. Shadowing short words entails a narrow focus and facilitates attention to phonetic detail, while shadowing longer utterances requires a broader focus and leads to higher cognitive load, as it is the case in fully conversational interaction.

While the experimental settings are arranged along a continuum from mere repetition to fully conversational interaction whose stages can be fairly straightforwardly examined and compared, the choice of which actual phonetic features to investigate seems a much more open question.

One approach is to evaluate accommodation holistically by measuring perceptual similarity (e.g., Goldinger, 1998; Namy et al., 2002; Miller et al., 2013; Babel et al., 2014; Dias and Rosenblum, 2016). Apart from the perceptual approach, there are global acoustic measures which have been applied to estimate overall accommodation, such as the long-term average spectrum (LTAS) (Gregory and Webster, 1996), mel-frequency cepstral coefficients (MFCCs) (Delvaux and Soquet, 2007), and amplitude envelopes (Lewandowski, 2012; Lewandowski and Jilka, 2019). A next step in substantiating these holistic findings is to examine the global acoustic-prosodic level, e.g., by measuring accommodation in overall or turn-based fundamental frequency, intensity, or speaking rate (e.g., Coulston et al., 2002; Bell et al., 2003; Levitan and Hirschberg, 2011; Michalsky and Schoormann, 2017). Eventually, more local phenomena are targeted, such as vowel quality (e.g., Babel, 2010, 2012; Nguyen et al., 2012; Dufour and Nguyen, 2013), voice onset time (VOT) (e.g., Fowler et al., 2003; Abrego-Collier et al., 2011; Nielsen, 2011; Yu et al., 2013), pitch accents (Schweitzer et al., 2017), or allophonic variation (e.g., Mitterer and Ernestus, 2008; Honorof et al., 2011; Mitterer and Müsseler, 2013).

The present study examines features pertaining to both the global and the local level. On the global level, we use amplitude envelopes to characterize the distribution of spectral energy of individual target words within the short sentences uttered in the shadowing task (see Section 2.2.3). On the level of local prosody, we compare pitch accent realization in these sentences by parameterizing their shapes with the PaIntE model (see Section 2.2.2). Further, we examine the variation of the German allophones [ɛː] vs. [eː] as a realization of the long vowel ⟨-ä-⟩ in stressed syllables, e.g., ⟨Best**ä**tigung⟩ (*engl. confirmation*), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩, e.g., ⟨Ess**ig**⟩ (*engl. vinegar*), as well as the epenthesis of schwa in a context where schwa is usually elided, namely in the word ending ⟨-en⟩ when preceded by a plosive or a fricative, e.g., ⟨begleit**en**⟩ (*engl. accompany*) (see Section 2.2.1).

Amplitude envelopes have been demonstrated to be useful in accounting for phonetic convergence. Lewandowski (2012), for example, showed for a HHI corpus of quasi-spontaneous dialogs between non-native and native speakers of English that the amplitude envelopes of tokens of the same word uttered by the interlocutors become more similar over time. We expect this to happen during the present shadowing task as well (see Section 3.4).

The analysis of pitch accent realization as parameterized by the PaIntE model is motivated by Schweitzer et al. (2017), who showed that native speakers of German accommodate their realization of pitch accents during spontaneous HHI dialogs in the GErman COnversations (GECO) corpus (Schweitzer et al., 2014, 2015). They diverged when they could see each other and converged when they could not see each other. Based on the design of our experiment, in which the participants do not see the model speakers they are shadowing, we therefore expect convergence to occur with respect to pitch accent realization (see Section 3.3).

The segmental phenomena for which accommodation in the form of convergence has been demonstrated to occur include vowel quality, motivating our choice of the vowel contrast [ɛː] vs. [eː] (e.g., Babel, 2012; Dufour and Nguyen, 2013), as well as the German allophone pair [ɪç] vs. [ɪk] (Mitterer and Müsseler, 2013) in the present study. The pronunciation variation [n̩] vs. [ən] has, to the best of our knowledge, not been studied in a shadowing experiment so far. Although the present study uses longer utterances as stimuli to investigate segment-level phenomena than previous shadowing experiments, we expect similar accommodation effects to occur as long as the variations are clearly perceptible (see Section 3.2).

We acknowledge the fact that the accommodating behavior of a speaker can vary between experimental settings, as has been shown by Pardo et al. (2018) comparing perceptual similarity between speakers and model talkers in conversational interaction and speech shadowing. However, since the pronunciation variations are embedded in full sentences in the present study, they are less salient and thus less obvious targets for accommodation. Under these circumstances, although staying within the rather static shadowing paradigm, occurrence of accommodation may be more readily transferable to actual dialog.

1.4. Hypotheses and predictions

We assume that the participants of the present study generally accommodate their speech to the stimuli during the shadowing task. Specifically, we predict that the participants converge to the stimuli, since convergence has been proposed as the default behavior under the assumption that accommodation is triggered automatically, and because the automatic motivation is presumably dominant in the socially impoverished environment of a shadowing task.

It is unclear to what extent social motivation for accommodation applies under the given circumstances. However, if it does apply — for example, due to mere exposure to a human voice — we have no reason to believe that participants would feel the need to increase social distance to the shadowed speakers and hence phonetically diverge from them, since there is no further interaction between the participants and the model speaker voices beyond the shadowing task itself and the text material used in the latter is uncontroversial, i.e., does not inspire resentment.

The focus of our study lies on the question whether participants behave similarly when confronted with either natural or synthetic stimuli. Again, under the assumption that automaticity is the main driving factor of accommodation in a shadowing task, we expect participants to converge to the synthetic stimuli as well. With respect to the social motivation, what was said above holds for both stimulus types: there is no a priori reason to increase social distance to the shadowed voices. However, the fact that the synthetic stimuli are probably recognized as non-human (cf. typicality) by the participants may trigger a feeling of social separation, which may lead to a reduction of the convergence effect and potentially even to divergence.

Although we expect overall accommodation, we predict that there is substantial variation between the participants of the experiment, presumably due to factors mentioned above such as their perception of the interlocutor.

This individual variation may, on the one hand, surface as different degrees of accommodation in one phonetic feature; on the other hand, participants may accommodate to different subsets of the phonetic features examined in this study, rather than to either all or none of them.

Coming back to the distinction between natural and synthetic stimuli in this context, it is possible that certain phonetic features are difficult to perceive in synthetic speech, and therefore do not lead to accommodating behavior. This may, for example, be the case if a phenomenon is not present in the underlying database or the model built for synthesis. If so, this may concern different features for the two different synthesis methods used in our study.

2. Material and methods

2.1. Corpus

The present analyses are carried out on a corpus of shadowed speech.[2] The corpus contains 6720 instances of short German sentences (both declaratives and interrogatives) which were uttered by 56 native speakers of German in a shadowing experiment. The experiment included a shadowing task, in which the participants repeated the sentences after hearing them from a female or a male voice, which were either natural or synthetic. The natural stimuli were recorded by a female and a male native speaker of German; two sets of synthetic stimuli were created using diphone and HMM synthesis, both with a female and a male voice (see Section 2.1.1). Each participant shadowed only one stimulus type, but in both the female and the male version. The participants were not told whether the stimuli they heard were natural or synthetic. The shadowing task was preceded by a baseline production phase and followed by a post production phase in which the participants read the same text material from a screen. Between the baseline production (ca. 4 min) and the shadowing task (ca. 6 min), the participants played a game on a tablet that involved no linguistic input or output, which we refer to as the visual task (ca. 7 min). The post production (ca. 4 min) immediately followed the shadowing task. The entire experimental procedure including informed consent, instructions, a final questionnaire (see Section 2.1.2), and the remuneration took about 45 min.

While the baseline production serves to determine the participants' preference with respect to the pronunciation variants — or the baseline values of the other examined features — and the shadowing task tests the accommodation towards the model voices, the post production allows to evaluate whether the accommodation effect is fully or partially sustained after the shadowing task or the participants return to the baseline level of the respective feature. The visual task was incorporated to weaken the participants' mental representation of their own baseline productions before continuing with the shadowing task.

The text material presented to the participants consists of 15 target and 15 filler sentences (see Appendix). Every target sentence contains one of three segments for which two prototypical pronunciation variants are expected to occur in native speakers of German (see Section 2.2.1):

(1) ⟨-ä-⟩ as [ɛː] or [eː] – e.g., in: Die Bestätigung ist für Tanja.
the confirmation is for Tanja

(2) ⟨-ig⟩ as [ɪç] or [ɪk] – e.g., in: Kommt Essig in den Salat?
does go vinegar into the salad

(3) ⟨-en⟩ as [n̩] or [ən] – e.g., in: Sie begleiten dich zur Taufe.
they accompany you to the baptism

All natural and synthetic target stimuli exist in both versions, with the exception of the female HMM [ɛː] targets, which were indeed more open than the [eː] targets, but not undisputably distinguishable due to technical reasons (see Section 2.1.1).

[2] The corpus was annotated using the WebMAUS services (Kisler et al., 2017). Manual corrections were carried out where necessary for the analyses.

For the purpose of this study, the three variations are initially regarded as binary contrasts: [ɛː] vs. [eː], [ɪç] vs. [ɪk], and [n̩] vs. [ən]. During the baseline phase, the participants' productions are auditorily identified by the experimenters as belonging to one of the two categories. Each of the three variations appears in five target sentences. Since some speakers use both variants of a pair interchangeably, forms with a minimum of three out of five possible occurrences are considered to be the preferred variant of a speaker.

During the shadowing task, the stimuli were selected so that the participants heard the opposite of the pronunciation variant they had uttered predominantly during the baseline production, hence their dispreferred variant, for most of the items. This provided them with the opportunity to accommodate phonetically. For some of the items, namely the ones the participants produced with their dispreferred variant during the baseline production, they would hear their preferred version during the shadowing task.

The filler sentences are comparable in length, but do not contain any of the target features listed above, for example:

- Die Glühbirne ist leider kaputt.
  the lightbulb is unfortunately broken
- Habt ihr das rote Auto erkannt?
  did you the red car recognize

Apart from the explicit manipulation on the segmental level, it can be assumed that all stimuli — targets as well as fillers — naturally differed from the versions the participants uttered in the baseline production on various levels (e.g., speaking rate, intonation pattern, rhythm, segmental pronunciation), giving additional opportunities to accommodate.

The participants of the experiment (see Section 2.1.2) were recorded in a sound-attenuated booth using a stationary cardioid microphone. The instructions for the experiment were given in written form on a screen in front of the participants. To avoid priming of convergence through the instructions (cf. Dufour and Nguyen, 2013), words such as "repeat" and "imitate" were not used. For the baseline and post productions the pertinent part of the instructions read as follows (English translation): "We will now record 30 short sentences with you. Please speak completely normally."; for the shadowing task the instructions were: "We will now record another 60 short sentences with you. This time, you will not read the sentences, but hear them. Please speak completely normally again." We cannot exclude the possibility that some participants might still have interpreted the task as an imitation task. Given the length of the stimuli, which is unusual for a shadowing experiment in accommodation research, the target for imitation would still be rather broad and it would not be obvious which specific features were to be imitated.

To allow the participants to become familiar with the task, a small number of test sentences were provided at the beginning of the experiment, which were not included in the later analysis.

During the shadowing task, 60 stimuli in two blocks of 30 stimuli each from a female and a male voice were played back to the participants over headphones, with half of the participants hearing the female voice first, the other half hearing the male voice first. Within the blocks, the stimuli were semi-randomized for balanced distribution of the targets over the two sets.

Fig. 1 illustrates the flow of the data collection process.

2.1.1. Stimuli

For the natural stimuli, two native speakers of German (female, 25 years old; male, 23 years old) were recorded in a sound-attenuated booth using a stationary cardioid microphone. The 30 target and filler sentences were presented on a computer screen and the speakers were instructed to speak naturally, as if in conversation with someone. Subsequently, the 15 target sentences were presented again. The three pronunciation variations were explained to the speakers and they were asked to distinctly produce the two corresponding variants for every target sentence. The best tokens in terms of target feature production and overall clarity were selected.

The first set of synthetic stimuli was created using diphone-based synthesis with MBROLA (Dutoit et al., 1996). One female and one male voice were used to match the sex of the natural speakers. For the realization of the segmental variations, different phonetic transcriptions of the target sentences were provided to the system, one for each of the pronunciation variants. To control for potential differences in prosody and information structure between the natural and synthetic stimuli, the F_0 contours and segment durations of the natural stimuli were specified as parameters to the synthesis system. This resulted in diphone-based stimuli with the same F_0 contours and segment durations as the natural stimuli.

The second set of synthetic stimuli was created using the HMM-based Speech Synthesis System (HTS, version 2.3) (Zen and Toda, 2005) with the BITS unit selection corpus (Ellbogen et al., 2004). Again, one female and one male voice were used and the F_0 contours and segment durations of the natural stimuli were imposed on the synthetic stimuli.

The resulting 270 stimuli (45 stimuli × 3 types × 2 sexes) were stored in a database for use in the experiment.

To assess the perceived quality of the stimuli, scores of naturalness and likability were collected from the participants directly after the experiment. They rated only those female and male voices they had heard during the experiment on 8-point scales from *1 – very unnatural* to *8 – very natural* and from *1 – not likable* to *8 – likable*. We used 8-point scales to provide enough room for differentiation between the six voices. Since we can assume that the participants interpreted the unlabeled steps between the endpoints as equidistant intervals, we can consider this an approximation of an interval scale and calculate the mean as a measure of the central tendency.

The naturalness of the natural stimuli was judged with a mean score of 6.2 (SD = 1.2) for the female voice and 5.5 (SD = 1.6) for the male voice. Thus, even the natural stimuli were not evaluated as perfectly natural. This may be partly due to the central tendency bias, which disfavors extreme responses on such rating scales. But it also suggests that the participants' concept of a *very natural* sounding voice is not necessarily fulfilled by a natural voice. The diphone stimuli received mean naturalness scores of 2.6 (SD = 1.8) for the female voice and 3.5 (SD = 1.9) for the male voice and were thus perceived as least natural. The HMM stimuli, finally, were rated with a mean naturalness of 4.0 (SD = 2.4) for the female voice and 4.3 (SD = 2.8) for the male voice and thus showed the greatest variance in ratings, indicating that the participants were less in agreement about their degree of naturalness. We can conclude that the synthetic voices were perceived as less natural than the natural voices. This was the case although they were evaluated separately by different listeners. A direct comparison in a joint evaluation would likely reinforce this difference. However, for the present study only the assessment of the stimuli that the participants actually heard is of relevance.

The likability of the six voices was rated somewhat more uniformly by the participants. The natural stimuli received mean likability scores of 5.5 (SD = 1.4) for the female voice and 5.1 (SD = 2) for the male voice; the diphone stimuli were rated with a mean likability of 3.8 (SD = 1.3) for the female voice and 4.6 (SD = 1.6) for the male voice; the HMM stimuli scored a mean likability of 5.1 (SD = 2.1) for the female voice and 5.6 (SD = 1.9) for the male voice. Thus while the likability of the natural and HMM voices was rated almost the same, slightly on the positive side of the scale, the female diphone voice was rated both as the most unnatural and least likable, followed by the male diphone voice.

We selected two synthesis methods which made it possible to control the output on the level of individual segments. This was done directly, by changing the desired target diphone, in the case of diphone synthesis, and indirectly, by training the voices with the pronunciation variants, in the case of HMM synthesis.

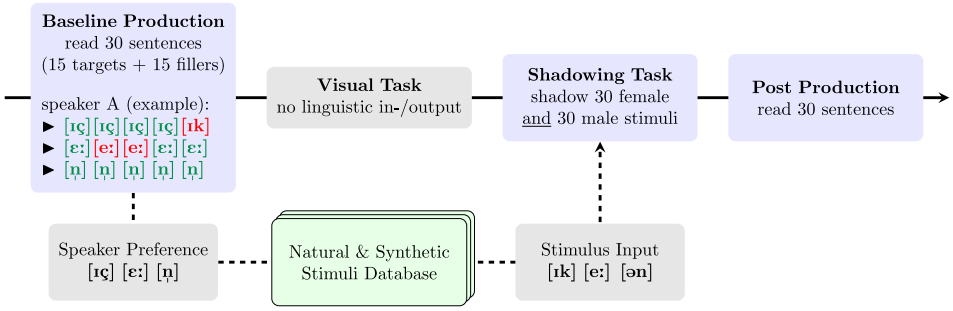

Fig. 1. Overview of the data collection process. The stimuli presented during the shadowing task are selected from the database depending on the *speaker preference*, i.e. the participant's preferred pronunciation variants in the baseline production. The stimuli containing the dispreferred variants are passed to the participants as *stimulus input*. Note that for **some items**, the participants still hear their preferred variant during the shadowing task; for **most items**, however, they hear their dispreferred variant.

Since HMM synthesis uses machine learning techniques, the degree of flexibility depends on the corpus used for training. Due to an imbalanced number of occurrences of the target sounds [ɛː] (n = 282) and [eː] (n = 1457) in the corpus underlying the female HMM voice, it was not possible, with the synthesis process applied here, to produce female HMM stimuli containing the target allophone [ɛː] that were clearly distinguishable from those containing the target allophone [eː]. Therefore, we decided to let the participants of the HMM group shadow both male and female stimuli to keep the experimental flow identical to the natural and diphone group, but only included their productions shadowing the male HMM [ɛː] stimuli in the present analysis. Note that this merely concerns the participants with a baseline preference for the allophone [eː] (n = 7, see Table 1).

One long-standing point of criticism towards diphone synthesis is the large number of concatenation points, which is detrimental to the perceived naturalness (Olive et al., 1998; Taylor, 2009). Discontinuities at concatenation points result in discontinuities of spectral trajectories that may be audible. Diphone systems use several techniques, including a careful construction of the diphone inventory, to reduce the discontinuities. Our diphone stimuli are generally rather smooth, but it is still possible that audible glitches affect the ability of the stimuli to trigger accommodation.

For these reasons, the present experiment used stimuli generated by HMM synthesis, which are generally smooth but can sound buzzy, and by diphone synthesis, which can be directly controlled but may contain discontinuities.

2.1.2. Participants

The participants were recruited on the Saarland University campus and paid for taking part in the experiment. 50 participants were students and six had non-academic jobs. All 56 participants were native speakers of German and 11 spoke more than one native language (e.g., Turkish, French, Vietnamese, Dutch). All had learned at least one, and the majority more than two, foreign languages. The most frequent foreign languages were English (n = 55), French (n = 44), and Spanish (n = 30). Multilingualism may be a favorable basis for phonetic accommodation, as it entails a certain amount of experience in switching between different pronunciation settings. It is likely that this basis would be more pronounced in native bilinguals and speakers who have achieved native-like pronunciation in a foreign language, as they either have more experience with different pronunciation settings or have been more successful in switching between them than speakers who have maintained a strong accent of their native language in acquired foreign languages. However, the participants in the present study were not selected according to these criteria and we are therefore not in a position to systematically investigate the effect of multilingualism on phonetic accommodation.

Table 1
Number of participants preferring the respective pronunciation variant as identified during the baseline phase. Participants in parentheses were excluded from the analysis of the corresponding feature.

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

The participants came from ten different German states and Austria with roughly 60 % from central regions and 20 % from northern and southern regions, respectively. The regional origin of the speakers will mainly be reflected in the baseline productions of the allophonic contrasts [ɛː]/[eː] and [ɪç]/[ɪk], as these are regionally distributed (see Section 2.2.1). We do not expect the regional origin to influence the accommodating behavior and do not investigate this further.

In a questionnaire completed after the experiment, which asked the participants to assess their general communicative behavior, 80 % answered affirmatively to the question whether they change the way they speak depending on their respective 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 the pronunciation of interlocutors.

These numbers, although they may not agree with the actual behavior of the participants, show that there is a certain awareness of the phenomenon of accommodation to an interlocutor in spoken communication. The readiness to accommodate seems to be higher when the accommodation target is more familiar (e.g., own vs. different dialect). A small number of participants perceives convergence to an interlocutor even as an intentional, active process. We will assess whether participants who indicated that they would converge to dialects of other regions or intentionally imitate interlocutors exhibit particular patterns of accommodation.

Each participant of the present study was presented with only one of the three stimulus types — natural, diphone, or HMM. This resulted in the following three experimental groups: the natural group with 21 participants (17 female/4 male; mean age 26.6 years; age range 19 to 34 years), the diphone group with 18 participants (14 female/4 male; mean age 26.2; age range 19 to 50 years), and the HMM group with 17 participants (13 female/4 male; mean age 26.8 years; 18 to 51 years). The between-subjects design was chosen to avoid learning and transfer effects over conditions that might have occurred if the same participants had been exposed to all three stimulus types.

2.2. Analyzed features

2.2.1. Allophones and schwa epenthesis

We examine whether participants accommodate on the level of segmental pronunciation with respect to the three types of pronunciation variation that were explicitly manipulated during the shadowing task. These pronunciation variations are commonly found among native speakers of German.

The realization of the long vowel ⟨-ä-⟩ in stressed syllables as [eː] or [ɛː],[3] and the realization of the word ending ⟨-ig⟩ as [ɪç] or [ɪk], vary regionally, occurring roughly in the North and South of the German-speaking region of Europe, respectively (Kleiner, 2011).[4] The Standard German variants of each pair are [ɛː] (predominant in the South) and [ɪç] (predominant in the North) (cf. Dudenredaktion, 2015). However, it has been shown that the respective non-standard forms, [eː] and [ɪk], are not perceived as strong dialectal markers by native listeners of German. According to Kiesewalter (2019), the realization of ⟨-ä-⟩ as [eː] subjectively corresponds to the standard, and the realization of ⟨-ig⟩ as [ɪk] is perceived as only slightly dialectal.

Elision or epenthesis of [ə] in the word ending ⟨-en⟩ when preceded by a plosive or a fricative varies mainly based on speaking style. In Standard German, schwa is elided in this position. An epenthetic schwa in this context, despite occurring in certain German dialects, is primarily produced when speaking particularly slowly and clearly. It is often perceived as hyperarticulation, especially when the quality is additionally shifted towards [e] or [ɛ]. Since humans have been shown to apply hyperarticulation when conversing with computers (e.g., Burnham et al., 2010), it may be the case that participants are more likely to pick up this trait from a synthetic voice than from a natural one.

Although speakers have their preferred variants in the contexts given in this study, [ɛː], [eː], [ɪk], [ŋ], and [ən] are all part of the basic phonetic inventory of native speakers of German and used by all speakers in other contexts. Only [ɪç] is an exception here, since many speakers realize [ç] as [ʃ] or [ɕ]. In our analysis, the latter are evaluated as phonetically different members of the underlying fricative class and included in the [ɪç] category. Ultimately, every participant has the necessary means to accommodate with respect to the pronunciation variations examined in the present study.

The degree of accommodation for the three pronunciation variations was quantified as follows:

The vowel quality [ɛː] vs. [eː] was evaluated as a continuum in the F1–F2 formant space. Automatic annotations (WebMAUS, Kisler et al., 2017) of all target vowel segments were manually corrected by a trained phonetician. The first and second formants of each target vowel were measured at the temporal midpoint in all productions as well as in the stimuli using Praat's (Boersma and Weenink, 2017) Burg algorithm. In contrast to a preliminary analysis in Gessinger et al. (2017), where the mean of all model speaker vowels (female and male combined) was defined as the overall convergence target, we now took a more fine-grained approach by calculating the Euclidean distance between each of the speakers' productions and the corresponding vowel of the model speaker they were shadowing in the respective instance, as

$$dist = \sqrt{(F1_{participant} - F1_{model})^2 + (F2_{participant} - F2_{model})^2}$$

Fig. 2 illustrates the utterance pairings for which the Euclidean distance was calculated. For the baseline and post productions, the Euclidean distance was calculated twice per speaker production: once in comparison to the female model speaker vowels and once in comparison to the male model speaker vowels. For the shadowing productions, only the Euclidean distance to the stimulus shadowed in the respective instance was calculated. This resulted in six comparisons per speaker and item.

A decrease of Euclidean distance in the F1–F2 formant space indicates convergence of vowel quality to the model speakers; conversely, an increase indicates divergence.

For further analysis, the DID_{vowel} was calculated between baseline and shadowing (bs), baseline and post (bp), and shadowing and post (sp) productions. DID_{vowel} is positive in the case of convergence and negative in the case of divergence.

The variation [ɪç] vs. [ɪk] was evaluated as a binary contrast. All target segments were manually annotated by a trained phonetician as belonging to the fricative or plosive class of the contrast by correcting automatic annotations. As mentioned above, some speakers produced instances of both categories in the baseline phase. In those cases, participants heard their preferred variant for some of the items in the shadowing phase (see Fig. 1). The present analysis of the data accounts for this fact by comparing each participant production to the variant they heard from the model speakers and determining whether these are the same or different variants of the binary contrast. A significant increase of *same*-cases indicates convergence of the pronunciation variant to the model speakers, and a decrease indicates divergence.[5]

The presence or absence of [ə] in the word ending ⟨-en⟩ was determined by measuring the duration of potential schwa segments between the preceding consonant (here [d], [x], [t], [ç], or [f]; see Appendix) and the final nasal, which were determined by manual correction of automatic annotations as performed by a trained phonetician. A duration of 30 ms was established as a minimum threshold to count the segment in question as a schwa. This decision is supported by the fact that all unambiguous schwas occurring in the stimuli were at least 30 ms long. As in the case of [ɪç] vs. [ɪk], we were taking all speaker productions into account and counted *same* (as model) vs. *different* (from model) cases. A significant increase of *same*-cases indicates convergence of the pronunciation variant to the model speakers, while a decrease indicates divergence.

2.2.2. Pitch accent comparison with PaIntE

In German, post-lexical accentuation is achieved by increasing intensity and length, as well as producing full instead of reduced vowel qualities. If such stressed units are further accompanied by pitch movement, they are called pitch accents (Möbius, 1993). A nuclear pitch accent is the last pitch accent in a prosodic phrase and may, in the text material of the present study, coincide with the last syllable of an utterance or occur in non-final position. Prenuclear pitch accents are all pitch accents occurring before the nuclear pitch accent in a prosodic phrase. To characterize and compare the pitch accents phonetically in the present study, we use the PaIntE model (Möhler, 1998; Möhler and Conkie, 1998; Schweitzer et al., 0000).

The Parametric Intonation Event (PaIntE) model approximates the F_0 contour of intonation events with the sum of a rising and a falling sigmoid as shown in Fig. 3. Each parameterization takes the syllable carrying the intonation event σ^*, as well as one preceding and one following syllable σ as the basis for the analysis. The length of each syllable is normalized to 1; the three syllables thus fit into the range of −1 to 2.

The model function is characterized by six parameters: *c1* and *a1* represent the height and slope of the rising sigmoid, respectively; *c2* and *a2* provide the same information for the falling sigmoid. The parameters *d* and *b* describe the absolute height and the relative syllable alignment of the F_0 peak, respectively.

[3] This contrast also occurs word-initially, but we only take word-medial occurrences into account in this study.

[4] Note that for Austria [eː] is more common in the East, whereas [ɛː] is typically encountered in the West (Dudenredaktion, 2015; Kleiner, 2011).

[5] Note that a preliminary analysis of [ɪç] vs. [ɪk] in this corpus only counted cases of convergence vs. cases of non-convergence from baseline to shadowing phase and excluded those instances where the same variant was already produced in the baseline phase (Gessinger et al., 2017).

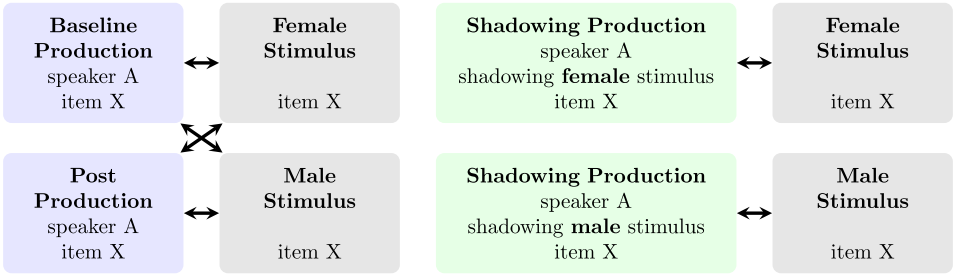

Fig. 2. Utterance pairings for the analyses of vowel quality, pitch accent realization, and amplitude envelopes. Baseline and post productions are compared twice, i.e., to the corresponding female and male stimuli. Shadowing productions are compared to the stimulus shadowed in the respective instance. This results in six comparisons per speaker and item.

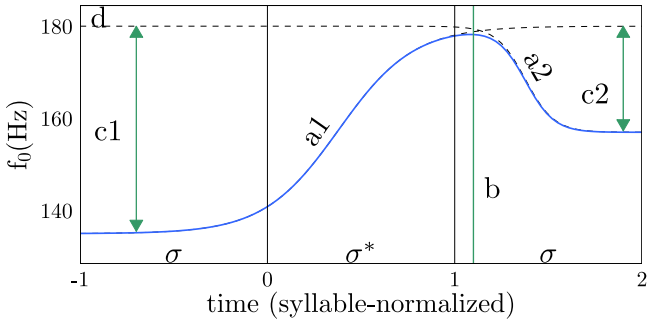

Fig. 3. Parameterization of an intonation event on the syllable σ^* by the PaIntE model. The F_0 contour is approximated with the sum of a rising and a falling sigmoid function. The approximation is characterized by six parameters: *a1*, *a2*, *b* (peak alignment), *c1*, *c2*, and *d* (peak height).
Source: Figure adapted from Möhler and Conkie (1998) and Schweitzer et al. (2017).

If the F_0 contour cannot be fitted with two sigmoids, only a single sigmoid is applied (either rising or falling, see the dashed lines in Fig. 3), leaving one set of *c* and *a* parameters unspecified. If a single sigmoid is not a good fit either, PaIntE only provides the mean F_0 value as the *d* parameter, leaving all other parameters unspecified.

To extract the PaIntE parameters, F_0 is tracked using the `get_f0` function of the Entropic Signal Processing System (ESPS) (Talkin, 1995). The resulting raw contour is smoothed by the `smooth_f0` algorithm authored by Gregor Möhler (March 2000). It uses the `smooth_phrase` algorithm from the Edinburgh Speech Tools Library[6] (King et al., 1999).

To determine the target syllables for the present study, prenuclear and nuclear pitch accents of all stimuli used in the shadowing task were manually annotated by a trained phonetician. Since the F_0 contours and segment durations from the natural stimuli were imposed on the two types of synthetic stimuli during their generation, it is expected that the same pitch accent locations are found in all three stimulus sets — natural, diphone, and HMM. This was true for the vast majority of the utterances. In very few cases, additional pitch accents occurred in the synthetic stimuli. These were taken into account in the analysis. Overall, the distribution of accent types found in the stimuli was 59 % prenuclear and 41 % nuclear pitch accents. Of the nuclear pitch accents, 34 % coincided with the last syllable of an utterance and 66 % occurred in non-final position.

The PaIntE parameters were extracted for every syllable of all utterances after manual correction of the automatically determined syllable boundaries by a trained phonetician. Then, the data was cleaned by executing the following steps: Cases in which the pitch accent shape was estimated by the mean F_0 alone, leaving all other parameters unspecified, were excluded from the analysis (approx. 6 % of the data). Furthermore, cases in which one of the six (two sigmoids fitted) or four (one sigmoid fitted) parameter values fell into the 1th or 99th percentile for that parameter in one speaker, were excluded as well (approx. 10 % of the data) to remove potential measurement errors while keeping atypical yet plausible values in the data. Such atypical values are expected when a speaker accommodates to an interlocutor.

To subsequently calculate the Euclidean distance between 6-dimensional PaIntE parameter vectors, the *c* (height) and *a* (slope) parameters were set to 0 wherever they were unspecified. Remember that this is the case when only a single sigmoid was fitted. Then, all PaIntE parameters were standardized to speaker specific z-scores to eliminate differences linked to the speaker sex and to give all parameters the same weight in the distance analysis.

Finally, the Euclidean distance between the 6-dimensional PaIntE parameter vectors $\vec{p}$ of a participant and $\vec{m}$ of a model voice was calculated for the same syllable as

$$d(\vec{p},\vec{m}) = \sqrt{\sum_{i=1}^{6}(participant_i - model_i)^2}; \text{where } i = \text{vector dimension}$$

This was done for the pairings detailed in Fig. 2, as described above in detail for the vowel quality analysis. A decrease of Euclidean distance indicates convergence of pitch accent realization to the model speakers; conversely, an increase indicates divergence.

For further analysis, we reduced the data set to the target syllables defined above, i.e., the syllables carrying a prenuclear or nuclear pitch accent in the stimuli, and calculated the DID_{PaIntE} between baseline and shadowing (bs), baseline and post (bp), as well as shadowing and post (sp) productions. DID_{PaIntE} is positive in the case of convergence and negative in the case of divergence.

2.2.3. Word-level amplitude envelope analysis

Contrary to the features discussed so far, amplitude envelopes represent the speech signal globally by the distribution of spectral energy across time and do not single out specific areas of interest from the signal (Wade et al., 2010).

In the present study, the amplitude envelope analysis is carried out on one word per utterance. In the target utterances, this is the word containing the segmental manipulation, whereas in the filler utterances, a regular content word was selected. For the target utterances, the analysis of word-level spectral composition is therefore related to the assessment of segmental pronunciation. See Appendix for an overview of the words in question and their location in the original target and filler sentences. It is possible that the spectral composition assimilates to a greater degree in utterances for which the stimulus explicitly encourages, or makes room for, accommodation, hence the target utterances.

The word boundaries were manually corrected in automatic annotations by a trained phonetician. For the analysis, the acoustic signal of a word was separated into four logarithmically spaced frequency bands between 80 Hz and 7800 Hz in MATLAB (version R2017a). An

[6] http://festvox.org/docs/speech_tools-2.4.0.

amplitude envelope was calculated for each resulting band using the linear Hilbert transform. The band-separated amplitude envelopes were then compared to their corresponding counterpart as detailed in Fig. 2.

Subsequently, each pairing of amplitude envelopes was transformed to have equal length while taking spectral characteristics into account, by performing DTW with the Speech Signal Processing Toolkit (SPTK)[7] (version 3.7). This resulted in the first similarity measure, i.e., the cost of the DTW operation, which is lower for more similar signals.

The resulting time-warped amplitude envelopes were then compared by cross-correlation. This resulted in the second similarity measure, i.e., the match value, which is the maximum value of the cross-correlation transformed onto a scale from zero to one with 1 indicating maximal similarity, i.e., identity.

As it was done for the PaIntE analysis, the DTW cost and match value data sets were cleaned by excluding values that fell into the 1th or 99th percentile for the respective parameter in one speaker (approx. 6 % of the data in both data sets).

For further analysis, we calculated the difference in distance for both similarity measures, DID_{DTW} and DID_{match}, between baseline and shadowing (bs), baseline and post (bp), and shadowing and post (sp) productions. DID_{DTW} is negative in the case of convergence and positive in the case of divergence, whereas DID_{match} is positive in the case of convergence and negative in the case of divergence.

2.3. Further factors

Apart from the influence of the experimental phase itself, namely baseline production, shadowing task or post production, there are further factors which might influence the measured variables and need to be accounted for in the analyses. These factors, discussed below, are either given by the design of the experiment or motivated by theoretical considerations (see Section 1).

Speaker preference. For each variation of segmental pronunciation examined in this study, the participants' preferred variant was identified during the baseline phase. Refer to Table 1 for an overview of the preference groups. Since only two out of 56 participants had a preference to produce [ən] in the baseline phase, we excluded these participants from the analysis of the schwa epenthesis. For the other two variations of segmental pronunciation, there are two preference groups: [ɛː] or [eː] and [ɪç] or [ɪk]. It is possible that the readiness to produce the respective other variant depends on the speaker's preference group. Especially since one of the variants is considered Standard German for the respective variation in the given context (see Section 2.2.1), there might be a bias in favor of producing this more prestigious variant. The factor PREFERENCE is included in the analysis of the allophonic contrasts (see Section 3.2).

Speaker attitude. At the end of the experiment, the participants were asked which variant of each pronunciation variation they believe to produce themselves and what they think of the respective other variant.[8] The majority of the participants reported a positive attitude towards the variants they do *not* believe to produce themselves — 80 % for [ɛː]/[eː], 70 % for [ɪç]/[ɪk], and 72 % for [n]/[ən]. This includes ratings such as "also ok", "better", and "Standard German". Only a minority of participants showed a negative attitude towards the other versions such as "wrong", "weird", and "sounds artificial". It seems plausible that a positive attitude towards a pronunciation variant might entail a higher probability of converging to it, whereas the production of variants carrying a negative connotation might be inhibited. The factor ATTITUDE with the two levels *positive* and *negative* is included in the analyses of all three pronunciation variations (see Section 3.2).

Pairing: same-sex vs. mixed-sex. In the present study, each speaker shadowed a female and a male model voice. As discussed in Section 1.2, this factor has yielded different outcomes in prior analyses, some suggesting that more accommodation occurs in same-sex, others in mixed-sex pairings. The analysis includes the factor PAIRING with the two levels *same-sex* and *mixed-sex*, where applicable, namely for *vowel quality* (see Section 3.2), *pitch accent realization* (see Section 3.3), as well as *DTW cost* and *match value* (see Section 3.4).

Sentence type. The analyses of pitch accent realization and word-level spectral composition are performed on both target and filler sentences. While the analysis of pitch accent realization has no particular link to the pronunciation variations in the target sentences, the analysis of spectral composition is based on the words containing these segmental variants. For the measures associated with spectral composition, i.e., the *DTW cost* and the *match value*, it can therefore be assumed that the distance between participant and model speaker baseline productions is greater for the target sentences than for the filler sentences. This additional space may enhance the accommodation effect. The factor SENTENCE with the two levels *filler* and *target* is therefore included in the analyses of the *DTW cost* and the *match value* (see Section 3.4).

Accent type. In the analysis of pitch accent realization, an additional factor comes into play that is motivated by prosodic theory, namely the accent type. We distinguish between *prenuclear* and *nuclear* pitch accents as the two levels of the factor ACCENT. The latter are known to be perceptually more salient (Jagdfeld and Baumann, 2011). We therefore expect a stronger accommodation effect for nuclear than for prenuclear pitch accents (see Section 3.3).

3. Analysis and results

3.1. Modeling

The dependent variables (see Section 2.2) are analyzed using linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) formulated with the lme4 package (1.1-18-1) (Bates et al., 2015) and evaluated with the lmerTest package (3.0-1) (Kuznetsova et al., 2017) in R (3.5.1) (R Core Team, 2018).

To strike a compromise between accuracy and complexity, model selection is carried out bottom-up, starting with a model which only includes the random factor intercepts for SUBJECT and ITEM. Then, theoretically relevant fixed factors (sum coded) and interactions as given by the design of the experiment or as motivated by the predictions made in Section 2.3 are added to the model. Random slopes for SUBJECT and/or ITEM are added for every effect where there is more than one observation for each unique combination of SUBJECT/ITEM and treatment level. Random slopes are only removed to simplify the model in cases of convergence errors or to allow a non-singular fit. The influence on the model fit is assessed by means of the Akaike information criterion (AIC), which estimates the relative quality of a statistical model for a given data set by taking into account the likelihood function and the number of estimated parameters (Akaike, 1973). A factor is kept in the model if the AIC value decreases by at least two points as compared to the model without the factor in question. Modeling is concluded by visual inspection of the residuals' normality and homoscedasticity. Factors kept in the model are being considered significant predictors of the respective dependent variable at $\alpha = 0.05$.

For the analyses taking a difference in distance (DID) measure as dependent variable (DID_{vowel}, DID_{PaIntE}, DID_{DTW}, DID_{match}), the information about the experimental phase is included in the dependent measure, since the DID measures are calculated as comparisons of the experimental phases: baseline and shadowing (bs), baseline and post

[7] http://sp-tk.sourceforge.net

[8] Note that approximately 30 % of the participants for each of the three features misjudged which variant they (predominantly) produce themselves. This is in line with the assumption stated in Mitterer and Müsseler (2013) with regard to [ɪç] vs. [ɪk] that speakers are often not consciously aware of which variant they use.

(bp), as well as shadowing and post (sp). It is therefore the model intercept that provides insight about accommodating behavior. The intercept is considered to significantly differ from 0 at $\alpha = 0.05$.

In comparison, the analyses of the binary contrasts [ɪç] vs. [ɪk] and [n̩] vs. [ən] take PHASE as a fixed factor into the model to assess accommodation. As for the DID measures, all experimental phases are compared to each other.

Comparing all experimental phases to each other allows to assess whether participants accommodate to the model speakers during the shadowing task (baseline vs. shadowing), whether the respective effect is sustained or reverted in the post phase (shadowing vs. post), and whether participants reach their baseline level again in the post phase (baseline vs. post).

3.2. Segmental pronunciation

Variation [ɛː] *vs.* [eː]. The distributions of DID_{vowel} measured for the vowel realizations are shown in Fig. 4. A positive DID_{vowel} indicates convergence to the model speakers, a negative DID_{vowel} divergence, and a DID_{vowel} close to zero maintenance of the vowel quality.

Recall that the analysis of the seven participants constituting the HMM group with a baseline preference for [eː] only includes their productions shadowing the male HMM [ɛː] stimuli (see Section 2.1.1).

Note also that the baseline productions of the two preference groups [ɛː] and [eː] were located at opposite ends of the F1–F2 space and their shadowing productions were expected to move towards each other, i.e., towards the model speaker vowels of the other variant. However, this difference in direction is canceled out in the calculation of the Euclidean distance. The two preference groups can therefore be jointly analyzed.

LMMs with DID_{vowel} as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. The factors PREFERENCE, PAIRING, and ATTITUDE were tested following the method in Section 3.1. Including the random factor intercepts for ITEM resulted in a singular fit for eight out of the nine models. Therefore, we only included SUBJECT as a random factor in all models. Table 2 shows the parameter estimates for the nine final models.

In the natural data set, mean DID_{vowel} is significantly positive for the base-shadow comparison, indicating convergence to the model speakers during the shadowing task, significantly negative for the shadow-post comparison, indicating divergence from the model speakers after the shadowing task, and not significantly different from zero for the base-post comparison, indicating that the participants reached their baseline level again in the post phase. Additionally, the convergence effect in the shadowing task is stronger for participants with baseline preference [eː], as indicated by the significant effect of PREFERENCE.

No effect was found for the diphone data set; participants do not seem to have accommodated to the diphone model speaker vowels.

In the HMM data set, we found a significant convergence effect in the shadowing task, but no significant divergence effect in the post phase. The diverging movement from shadowing task to post phase is, however, so substantial that participants ended up close to their baseline level again, as shown by the non-significant base-post phase comparison.

The factors PAIRING and ATTITUDE did not account for variance in the data.

Variation [iç] *vs.* [ik]. The percentages of cases in which participant and model speakers realized the *same* or a *different* variant of the segmental pronunciation variation [ɪç] vs. [ɪk] are shown in Fig. 5. In all three data sets, the number of same variants increases by about 30 % from the baseline phase to the shadowing phase, and decreases again in the post phase, yet to different degrees.[9]

Table 2
Results for variation [ɛː] vs. [eː] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{vowel} as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	69.94*** (14.89)	32.88 (17.53)	−33.67** (11.44)
PREFERENCE	33.79* (14.89)		
Observations	210	210	209
Diphone			
Intercept	−1.68 (8.72)	−1.49 (9.93)	0.44 (7.24)
Observations	177	179	178
HMM			
Intercept	32.64* (13.61)	−2.12 (12.38)	−31.23 (15.86)
Observations	134	135	133

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

GLMMs with IDENTITY (*same* or *different*) as the dependent variable were fitted for each stimulus type data set separately, always comparing two experimental phases at a time, resulting in nine models. The outcome *same* is coded as success in the models.

The factors PHASE, PREFERENCE, and ATTITUDE were tested following the method described in Section 3.1. Table 3 shows the parameter estimates for the nine final models. Note that these are binomial models and the coefficients are hence in logit-space. If a logit-coefficient is positive, the effect of the corresponding predictor on the response variable is positive as well, and vice versa.

The increase of same variants in the shadowing task is significant for both the natural data set (10 % to 39 %) and the diphone data set (16 % to 48 %). Moreover, in both data sets the number of same variants decreases again in the post phase, although, not all the way to the baseline level (natural: 39 % to 23 %; diphone: 48 % to 36 %).

For the HMM data set, the increase of same variants in the shadowing task (8 % to 40 %) does not reach significance in the statistical model. However, the decrease of same variants in the post phase is significant and reaches the baseline level (40 % to 12 %). The latter is shown by the fact that PHASE did not account for variance in the data set and was therefore not included in the HMM base-post model.

The factors PREFERENCE and ATTITUDE did not show any significant effect on IDENTITY, although the former factor did improve overall fit in various models.

Variation [n̩] *vs.* [ən]. The percentages of cases in which participant and model speakers realized the *same* or a *different* variant of the segmental pronunciation variation [n̩] vs. [ən] are shown in Fig. 6. In 85 % to 95 % of the cases over all experimental phases of all three data sets, participants produced a different variant than the model speakers. The statistical analysis was carried out as described for the [ɪç] vs. [ɪk] variation above, without testing the factor PREFERENCE, however, since all analyzed speakers preferred [n̩] in the baseline phase. Table 4 shows the parameter estimates for the nine final models.

Only in the case of the natural data set did participants produce significantly more [ən] (i.e., same variants) during the shadowing task, compared to the baseline phase (5 % to 15 %). The amount of same variants does not decrease significantly from the shadowing task to the

[9] Note that the numbers given in the text are descriptive values, whereas the GLMM result tables contain model estimates.

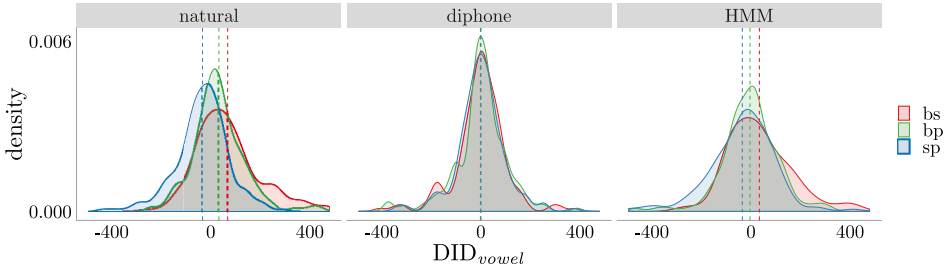

Fig. 4. Distributions of DID$_{vowel}$ for the pronunciation variation [ɛː] vs. [eː] in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means. (For a color version of the figure, the reader is referred to the web version of this article).

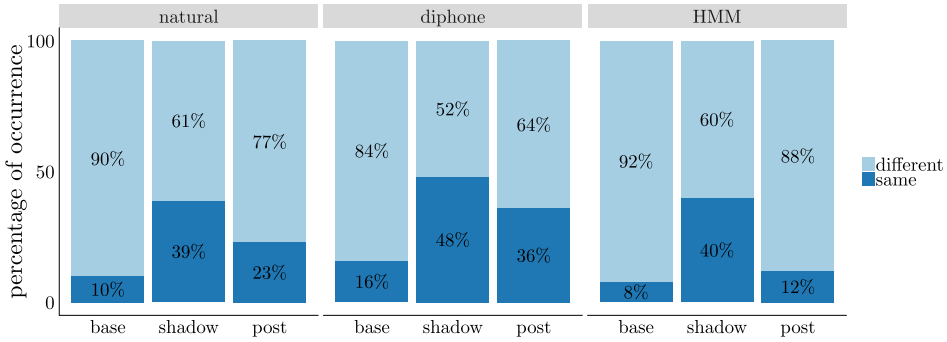

Fig. 5. Results for variation [ɪç] vs. [ɪk]. Cases where speaker and model realize the same variant are indicated in dark blue; cases where they realize a different variant are indicated in light blue. (For a color version of the figure, the reader is referred to the web version of this article).

Table 3
Results for variation [ɪç] vs. [ɪk] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on identity as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	−1.75***	−3.01***	−1.47*
	(0.47)	(0.79)	(0.6)
PHASE	−0.93**	−0.7**	0.75*
	(0.3)	(0.25)	(0.3)
PREFERENCE	0.05	−0.07	
	(0.57)	(0.64)	
Observations	315	210	315
Diphone			
Intercept	−1.43**	−1.54***	−0.44
	(0.54)	(0.42)	(0.57)
PHASE	−1.33***	−0.7***	0.49**
	(0.24)	(0.21)	(0.18)
PREFERENCE	−0.37	0.01	−0.52
	(0.62)	(0.39)	(0.67)
Observations	270	180	270
HMM			
Intercept	−1.9**	−3.64***	−2.87**
	(0.62)	(1.03)	(0.98)
PHASE	−0.84		1.34***
	(0.49)		(0.3)
PREFERENCE			−0.52
			(0.88)
PHASE:PREF			−0.5
			(0.29)
Observations	254	170	254

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

post phase (15 % to 10 %) and there is no significant difference between the baseline and the post phase.

For both synthetic data sets, the factor PHASE did not remain in the final models; participants do not seem to have accommodated to the synthetic model speakers with respect to this feature.

Table 4
Results for variation [n] vs. [ən] — parameter estimates (coefficients with standard errors in parenthesis) of the effects on identity as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	−3.27***	−10.11	−2.76***
	(0.63)	(5.46)	(0.63)
PHASE	−0.79**	−1.04	−0.01
	(0.28)	(0.56)	(0.27)
Observations	315	210	315
Diphone			
Intercept	−4.79***	−10.98*	−4.79***
	(1.28)	(4.71)	(1.28)
Observations	255	170	255
HMM			
Intercept	−3.13***	−2.74***	−3.89***
	(0.73)	(0.6)	(1.11)
Observations	239	159	238

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

The factor ATTITUDE did not influence IDENTITY.

3.3. Pitch accent realization

The distributions of DID$_{PaIntE}$ measured for the pitch accent realizations is shown in Fig. 7. A positive DID$_{PaIntE}$ indicates convergence to the model speakers, a negative DID$_{PaIntE}$ divergence, and a DID$_{PaIntE}$ close to zero maintenance of the pitch accent realization. As mentioned in Section 2.3, we distinguish prenuclear and nuclear pitch accents.

LMMs with DID$_{PaIntE}$ as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. The factors PAIRING and ACCENT were tested following the method in Section 3.1. Table 5 shows the parameter estimates for the nine final models.

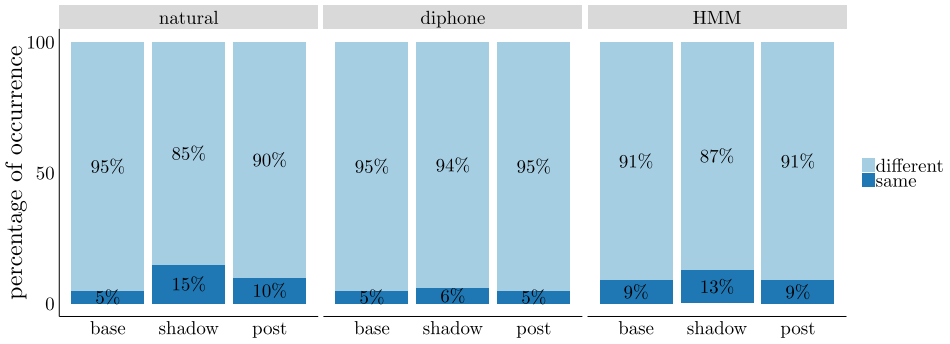

Fig. 6. Results for variation [n̩] vs. [ən]. Cases where speaker and model realize the same variant are indicated in dark blue; cases where they realize a different variant are indicated in light blue. Since all of the participants heard the model variant [ən], the percentage indicating *same*-cases coincides with the percentage of [ən] occurrences. (For a color version of the figure, the reader is referred to the web version of this article).

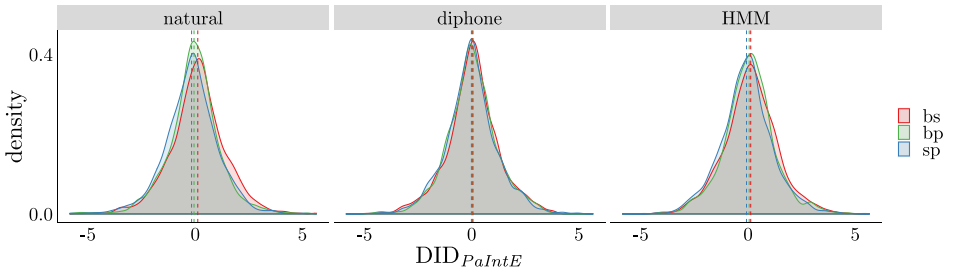

Fig. 7. Distributions of DID_{PaIntE} for the comparison of pitch accent realizations in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means. (For a color version of the figure, the reader is referred to the web version of this article).

Table 5
Results for the PaIntE analysis of the pitch accent realization — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{PaIntE} as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	0.11*	−0.04	−0.17***
	(0.04)	(0.06)	(0.04)
ACCENT		−0.04	
		(0.05)	
Observations	2136	2065	2031
Diphone			
Intercept	0.06	0.03	−0.01
	(0.04)	(0.05)	(0.04)
ACCENT	0.01		
	(0.03)		
PAIRING	0.02		
	(0.03)		
Observations	1687	1653	1643
HMM			
Intercept	0.12*	0.05	−0.8
	(0.05)	(0.06)	(0.05)
Observations	1637	1599	1580

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

In the natural data set the participants converged to the model speakers during the shadowing task and diverged again in the post phase, reaching the baseline level.

In the diphone group, no accommodation is observed.

For the HMM group participants converged towards the model speakers during the shadowing task, and, although there is no significant divergence effect in the post phase, became indistinguishably close to the baseline level in the post phase.

The factors PAIRING and ACCENT did not show any significant effect on the intercept, although they improved the overall fit in two models.

3.4. Word-level spectral composition

DTW cost. The distributions of DID_{DTW} resulting from the DTW cost analysis is shown in Fig. 8. A positive DID_{DTW} indicates convergence to the model speakers, a negative DID_{DTW} divergence, and a DID_{DTW} close to zero maintenance of the temporal structure of the target words.

LMMs with DID_{DTW} as the dependent variable were fitted for each stimulus type and phase comparison data set separately, resulting in nine models. As expected, the effects are very small, since we are comparing the same word spoken by different speakers and the room for variation, on the temporal as well as the spectral level, is therefore quite limited. The factors PAIRING and SENTENCE were tested following the method in Section 3.1. Table 6 shows the parameter estimates for the nine final models.

In the natural and diphone data sets, participants converged to the model speakers in the shadowing task and diverged during the post phase, reaching the baseline level.

For the HMM data set, convergence during the shadowing task is not significant; however, there is substantial movement away from the model speakers during the post phase and, eventually, no difference between baseline and post phase. Additionally, the sentence type accounts for variability in the case of the HMM data set: the diverging movement from shadowing task to post phase is stronger for the target sentences than for the filler sentences. Furthermore, the HMM base-post model suggests that — although there is no significant difference between baseline and post phase for the entire data set — the filler sentences are relatively closer to the model speakers in the post phase, compared to the baseline phase, while the target sentences are relatively farther away from the model speakers.

The factor PAIRING did not account for variance in the data.

Match value. The distribution of DID_{match} resulting from cross-correlating the time-warped amplitude envelopes is shown in Fig. 9. Contrary to the other DID measures, a **negative** DID_{match} indicates convergence to the model speakers and a **positive** DID_{match} divergence from the model speakers with respect to the spectral composition of the

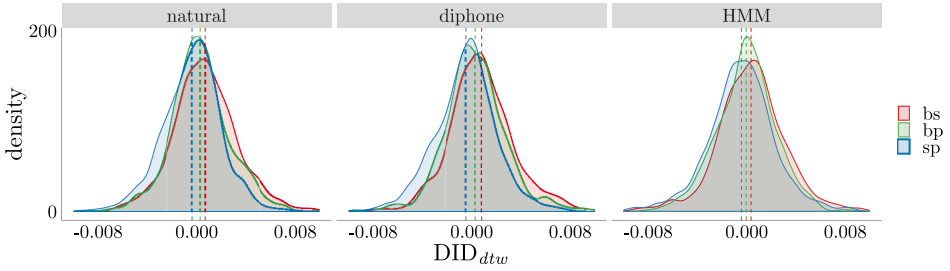

Fig. 8. Distributions of DID_{DTW} for the DTW cost analysis in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means. (For a color version of the figure, the reader is referred to the web version of this article).

Table 6
Results for the DTW analysis of the amplitude envelopes — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{DTW} as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	6.95×10^{-4}***	2.81×10^{-4}	-3.78×10^{-4}**
	(1.59×10^{-4})	(1.42×10^{-4})	(1.36×10^{-4})
SENTENCE	-0.01×10^{-4}		
	(1.5×10^{-4})		
Observations	1136	1139	1137
Diphone			
Intercept	8.28×10^{-4}***	2.9×10^{-4}	-4.86×10^{-4}*
	(1.95×10^{-4})	(2.05×10^{-4})	(1.9×10^{-4})
Observations	966	965	960
HMM			
Intercept	3.40×10^{-4}	-0.18×10^{-4}	-4.06×10^{-4}**
	(2.27×10^{-4})	(1.68×10^{-4})	(1.32×10^{-4})
SENTENCE	-2.0×10^{-4}	2.03×10^{-4}*	3.93×10^{-4}**
	(1.39×10^{-4})	(0.92×10^{-4})	(1.24×10^{-4})
Observations	908	909	909

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Table 7
Results for the match value analysis of the amplitude envelopes — parameter estimates (coefficients with standard errors in parenthesis) of the effects on DID_{match} as estimated in separate models for the three different stimulus types and the three phase comparisons.

Natural	Base-shadow	Base-post	Shadow-post
Intercept	−0.017***	−0.007*	0.010*
	(0.003)	(0.003)	(0.004)
PAIRING	0.003		
	(0.002)		
Observations	1138	1142	1139
Diphone			
Intercept	−0.019***	−0.002	0.016**
	(0.005)	(0.006)	(0.006)
SENTENCE	0.007		
	(0.005)		
PAIRING	0.005*		
	(0.002)		
Observations	964	971	965
HMM			
Intercept	−0.011	0.001	0.011**
	(0.006)	(0.005)	(0.004)
SENTENCE	0.004	−0.006*	−0.010*
	(0.005)	(0.003)	(0.004)
Observations	907	910	911

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

target words. As before, a DID_{match} close to zero indicates maintenance of the baseline behavior.

Recall that the match value itself is bounded between 0 and 1 and can therefore be interpreted as probability, with 1 indicating maximal similarity, i.e., identity. The distribution of the match value is skewed towards 1, since we are comparing the same word spoken by different speakers. Using DID_{match} as a dependent variable resolved these issues and we could still fit LMMs for each stimulus type and phase comparison data set separately, which resulted in nine models. The factors PAIRING and SENTENCE were tested following the method in Section 3.1. Table 7 shows the parameter estimates for the nine final models.

As for the DTW analysis, participants shadowing natural and diphone stimuli converged to the model speakers during the shadowing task and diverged again in the post phase. However, the natural group did not reach the baseline level in the post phase, but stayed in between baseline and shadowing levels. The diphone group reached the baseline level in the post phase. Additionally, for the diphone group, the convergence effect in the shadowing task was influenced by the pairing of participants: the effect is stronger in mixed-sex than in same-sex pairings.

For the HMM group, the accommodating effect from baseline to shadowing phase again does not reach significance. There is, however, a significant movement away from the model speakers in the post phase, reaching the baseline level. As for DID_{DTW}, the sentence type accounts for variability in the HMM data set, with the target sentences showing a stronger divergence effect from shadowing task to post phase and reaching values farther from the model speakers in the post phase, compared to the baseline phase.

3.5. Individual results

To go beyond the analysis of accommodation on the group level, we assessed the performance of the individual participants with respect to the six features discussed above, focusing on the comparison of the baseline phase to the shadowing task.

For the DID measures (DID_{vowel}, DID_{PaIntE}, DID_{DTW}, DID_{match}), we conducted Wilcoxon signed-rank tests to determine whether each individual participant converged to or diverged from the model speakers (i.e., significant difference of their individual DID measure distribution from 0 at $\alpha = 0.05$), or whether they maintained the distance to the model speakers (i.e., no significant difference of their individual DID measure distribution from 0).

The degree of accommodation for the two binary contrasts [ɪç] vs. [ɪk] and [n] vs. [ən] was assessed as the percentage of possible category changes. When determining the number of possible instances of accommodation, cases in which a participant already produced the same variant as the model speakers during the baseline phase were taken into consideration. The degree of accommodation for the binary contrasts was classified at the following thresholds so that single occurrences of convergence or divergence were still considered as maintaining behavior

- **convergence**: increase of same variants ≥ 20 %
- **maintenance**: increase of same or different variants < 20 %
- **divergence**: increase of different variants ≥ 20 %

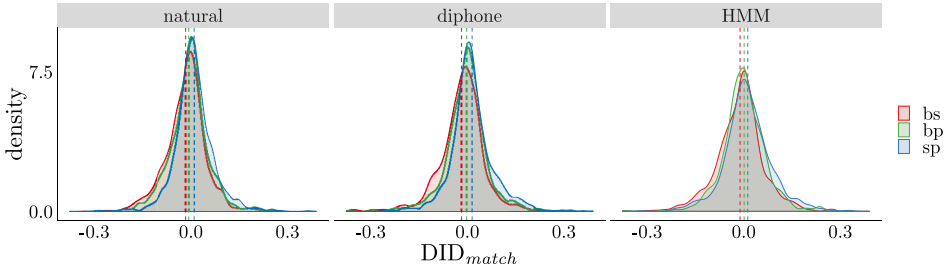

Fig. 9. Distributions of DID_{match} for the amplitude envelope analysis in the three experimental groups. Comparisons are made between **base-shadow**, **base-post**, as well as **shadow-post** phases. The dashed lines indicate the distribution means. (For a color version of the figure, the reader is referred to the web version of this article).

Table 8
Percentage of participants converging to the model speakers with respect to the respective feature in the three experimental groups, as well as in the entire participant group.

Feature	Natural % $n = 21$ (for [ŋ̩]/[ən]: $n = 21$	Diphone % $n = 18$ $n = 17$	HMM % $n = 17$ $n = 16$	Total % $n = 56$ $n = 54$)
[ɪç]/[ɪk]	61.9	72.2	64.7	66.1
DTW cost	52.4	61.1	41.2	51.8
match value	47.6	50.0	23.5	41.1
PaIntE	23.8	5.6	17.6	16.1
[ɛː]/[eː]	28.6	11.1	5.9	16.1
[ŋ̩]/[ən]	19.0	5.9	18.8	14.8

Table 9
Distribution of participants converging to the model speakers with respect to a different number of features (0 to 6) in the three experimental groups, as well as over all 56 participants. Dominant groups are highlighted in gray.

No. of features	Natural % $n = 21$	Diphone % $n = 18$	HMM % $n = 17$	Total % $n = 56$
0	9.5	11.1	17.6	12.5
1	28.7	33.3	35.3	32.1
2	9.5	5.6	11.8	8.9
3	33.3	38.9	29.4	33.9
4	9.5	11.1	5.9	8.9
5	9.5	–	–	3.6
6	–	–	–	–

Fig. 10 shows a summary of the individual results. The six features under examination are ordered by decreasing number of individual participants converging significantly to them. Most participants converge with respect to the binary contrast [ɪç] vs. [ɪk] ($n = 37$, 66.1 %), followed by the two measures related to the amplitude envelopes, DTW cost ($n = 29$, 51.8 %) and match value ($n = 23$, 44.1 %). The pitch accent realization assessed by the PaIntE model as well as the binary contrast [ɛː] vs. [eː] trigger convergence in 9 participants (16.1 %), respectively, and the binary contrast [ŋ̩] vs. [ən] in 8 participants (14.8 %).

Table 8 breaks these numbers down for the three experimental groups: natural, diphone, and HMM. Conducting 2 × 3 two-tailed Fisher's exact tests for the distribution of participants converging or non-converging (i.e., maintaining their behavior or diverging) over these three experimental groups, did not yield a significant result for any of the features. This suggests that, in every experimental group and for every feature, a similar proportion of participants converged to the model speakers.

Fig. 10 further illustrates the number of features with respect to which each individual participant converges out of the possible 6 under examination. No participant actually converged to all 6 features and only two participants converged to 5 features. Five participants converged to 4 and 2 features, respectively. The majority of the participants accumulated at 3 features ($n = 19$) and 1 feature ($n = 18$). A total of seven participants did not converge at all.

Table 9 details how these different degrees of convergence are distributed over the three experimental groups: natural, diphone, HMM. Conducting a 6 × 3 two-tailed Fisher's exact test for the distribution of participants converging to the model speakers for 0 to 5 features over these three experimental groups, did not yield a significant result. This suggests that, in every experimental group, a similar proportion of participants showed convergence to the model speakers with respect to the same number of features.

Some cases of individual divergence from the model voices were found as well, i.e., 4 cases for [ŋ̩]/[ən], 2 cases for [ɪç]/[ɪk], DTW cost and match value, respectively, and one case for [ɛː]/[eː]. No individual divergence was found for the pitch accent comparison with PaIntE.

We identified the participants who stated in the questionnaire after the experiment that they converge to dialects of other regions ($n = 8$) or intentionally imitate the pronunciation of interlocutors ($n = 9$) (see Section 2.1.2). Only two participants appeared in both groups.

For the first group, we could assume that they would specifically pick up the two regionally distributed features, [ɪç]/[ɪk] and [ɛː]/[eː]. However, only four of the eight speakers converged with respect to [ɪç]/[ɪk] and none with respect to [ɛː]/[eː], while one speaker from this group even diverged with respect to [ɪç]/[ɪk]. This does not indicate a particular inclination for convergence to regional features. In terms of overall convergence, the members of this group were not particularly successful either: they converged to a maximum of 3 features.

The second group, namely the speakers who claimed to intentionally imitate the pronunciation of interlocutors, also did not include any of the speakers converging to more than 3 features. With respect to [ɪç]/[ɪk] and DTW cost, five of the nine speakers converged, respectively; two speakers each picked up the schwa from the model voices and converged with respect to the amplitude envelope match; only one speaker converged to the pitch accent realization and none with respect to [ɛː]/[eː]. Divergence was not found in this group. These results do not notably reflect a possible effect of intentional imitation.

Of the two speakers who claimed to converge to dialects of other regions and to intentionally imitate interlocutors, one converged to 3 features and the other to none.

4. Discussion

The goal of the present study was to investigate phonetic accommodation of human interlocutors in a shadowing task with a specific focus on the accommodation effect evoked by synthetic stimuli. Diphone- and HMM-based synthetic stimuli, as well as natural stimuli, were used in the process. The language under investigation in this study is German. The shadowing task was carried out by native speakers of German.

To get a broader picture of phonetic accommodation in the experimental data, we examined features pertaining to different phonetic domains, i.e., variation of segment-level phenomena as well as variation with respect to pitch accent realization (local prosody) and word-based global similarity (temporal structure and distribution of spectral energy). The segment-level phenomena under investigation are allophonic variation of [ɪç]/[ɪk] and [ɛː]/[eː], as well as schwa

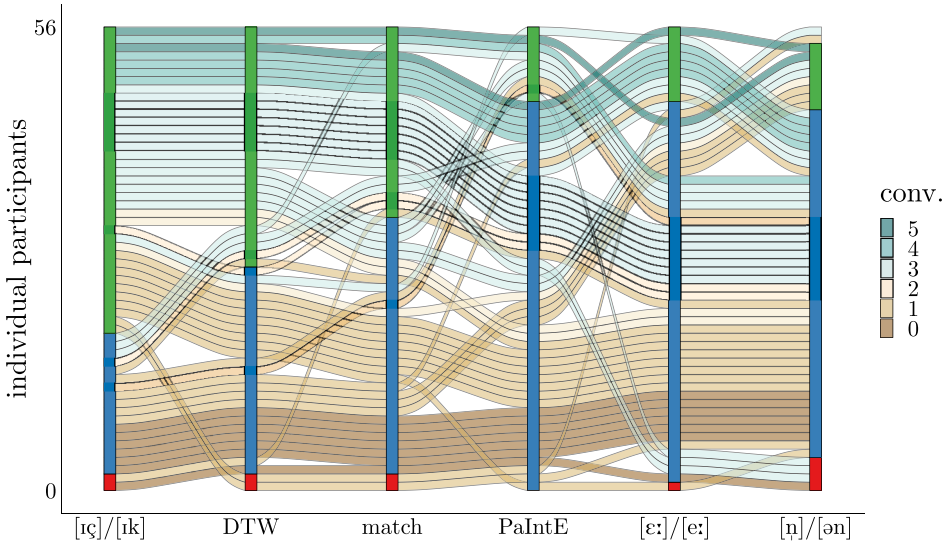

Fig. 10. Accommodating behavior of the 56 participants for the comparison of baseline phase and shadowing task. Each vertical bar stands for one examined feature; the colors of the sections indicate whether the corresponding participant shows **convergence**, **maintenance**, or **divergence** for the respective feature. Two participants were excluded from the analysis of schwa epenthesis as they were the only participants producing schwa as a baseline preference. Hence they are not included in the rightmost vertical bar. Each horizontal line stands for one individual participant; the colors of the lines indicate with respect to how many features each participant converged to the model speakers (see legend). (For a color version of the figure, the reader is referred to the web version of this article).

epenthesis. To make the systematic investigation of accommodation with respect to these features possible, the stimuli for the shadowing task were chosen depending on the participants' baseline productions: the participants were presented with the opposite of their preferred variants.

Analyses were carried out at the group level and for individual participants. Since the experimental procedure comprised three phases — baseline production, shadowing task, and post production — we drew three comparisons for each group data set, namely baseline vs. shadowing, shadowing vs. post, and baseline vs. post. For the individual behavior, we focused on the comparison of baseline phase and shadowing task. Combining the results of these analyses provides an overview of the phonetic accommodation in the present shadowing corpus.

The allophonic variation [ɪç] vs. [ɪk] was most successful in triggering convergence when looking at the individual results, with two thirds of all participants converging to the model speakers during the shadowing task. This may be due to the relative salience of this feature, even though it was embedded in a larger utterance, and to the fact that participants can presumably access the binary variation between fricative and plosive more easily than other, more gradual, changes.

At the group level, we found the same pattern for the natural and the diphone group: convergence to the model speakers during shadowing, divergence in the post phase, although not entirely falling back to the baseline level, but rather sustaining the convergence effect in attenuated form.

In the HMM group, although the relative increase of *same* forms is equal to the diphone group, the convergence effect is not significant in the statistical model. The group does, however, show a divergence effect from shadowing task to post phase and reaches the baseline level in the latter.

These results, in combination with the fact that a similar proportion of participants converges to the model speakers in all three groups, shows that [ɪç] vs. [ɪk] is a rather successful target for convergence in native speakers of German for natural as well as synthetic stimuli.

Neither the preference for one or the other variant in the baseline phase, nor the attitude towards the dispreferred variant being positive or negative, had an impact on the accommodation for this feature. The standard variant [ɪç] does not seem to be an easier target for convergence. This may have to do with the fact that the participants were in many cases not certain which of the variants is the standard, or at least did not have a negative attitude towards the variant they believed not to produce.

The second and third most frequent cases of convergence for individual speakers were found in the measures pertaining to the word-based global similarity.

The first measure, i.e., the cost of the dynamic time warping (DTW) process, emphasizes changes in the temporal domain while taking the spectral domain into account. It shows that more than half of the participants converged to the model speakers with respect to word-based timing.

The second measure, i.e., the match value, resulted from cross-correlating the time-warped amplitude envelopes. The analysis of the match value shows that almost half of the participants converged to the model speakers with respect to word-based distribution of spectral energy alone, i.e., excluding timing.

Fig. 10 shows that these two groups widely overlap, with 21 participants converging with respect to both features, eight participants converging only with respect to DTW cost, and two only with respect to the match value. This was expected, since these measures are closely related. However, taking both measures into account disentangles the contributions of timing and energy distribution across spectral bands to the accommodation effect.

On the group level, both measures behave similarly. There is a pattern of convergence to the model speakers in the shadowing task and divergence in the post phase reaching the baseline level, which occurred in the natural and diphone group for both measures. The only exception to this pattern is that the match value does not reach the baseline level in the natural group, meaning that the convergence effect was partially sustained in this case.

As before, the HMM group behaved differently. There is no significant convergence effect in the shadowing task for either of the two measures. However, as for the other two groups, we found a significant divergence effect in the post phase and the HMM group did reach the baseline level in the post phase.

It may have been the case that the target sentences, in particular, drive the accommodation for DTW cost and match value, since they specifically offer room for convergence in the form of the dispreferred segmental variants. This presupposes, of course, that participants accommodate with respect to the offered variants. Overall, such an influence of the sentence type did not manifest itself, especially not

in the actual shadowing phase. Only in the HMM group, the sentence type emerged as a significant predictor: the divergence effect in the post phase was stronger for the target sentences resulting in post productions which were relatively farther from the model speakers, compared to the baseline phase. Whether this behavior is indeed causally related to the dispreferred segmental variants in the target stimuli remains unclear.

The participant-model pairing only surfaced as a significant predictor in the match value analysis of the diphone group: convergence during the shadowing task was slightly stronger in mixed-sex pairings. Remember that prior studies on phonetic accommodation found both cases of more convergence in mixed-sex and same-sex pairings. Our results, although showing one incident of increased convergence in mixed-sex pairings, do not make a strong case in favor of speakers converging more to a model talker of the opposite sex.

The number of individual participants converging drops drastically for the allophonic variation [ɛː] vs. [eː] and the pitch accent realization, to only 16 %, respectively. Fig. 10 shows that among these, there is only one participant who converged with respect to both features.

On the group level, the observed patterns for both measures are again very similar to each other and also distinct from the patterns observed for other measures. The natural group converged to the model speakers, with respect to both vowel production and pitch accent realization, and diverged again in the post phase reaching the baseline level.

The HMM group showed convergence in the shadowing task, too. However, the divergence effect in the post phase was not significant and the group still reached the baseline level.

The diphone group, finally, did not show any accommodation with respect to vowel quality and pitch accents.

The speaker preference had a significant influence on the vowel production in the natural group: the convergence effect in the shadowing task was stronger for those participants whose baseline preference was [eː]. One possible explanation may be that [ɛː] is an easier target for convergence, since it is the standard German form and therefore more prestigious. However, although still being considered the prescriptive norm, [ɛː] is generally used less frequently by native speakers of German in Germany. Therefore, it may also be the case that [ɛː] is more salient to hearers and therefore picked up from the speech input more easily. Remember that atypicality has been shown to promote accommodation for some speakers (Babel et al., 2014). As for the allophonic contrast [ɪç] vs. [ɪk], the attitude towards the variant participants did not believe to produce themselves did not influence accommodation for the vowel quality. Recall that this attitude was predominantly positive, only 20 % of participants had a negative attitude towards the dispreferred variant in the case of [ɛː] vs. [eː].

For DID_{PaIntE} as a measure of similarity in pitch accent realization, the accent type was tested as an additional factor. The expected effect of higher perceptual salience of nuclear as opposed to prenuclear pitch accents did not appear.

Eventually, the participant-model pairing did not surface as a significant predictor in the analysis of vowel quality and pitch accent realization. This means that it did not make a difference for the accommodating behavior whether participant and model were of the same or different sex.

Note that the Euclidean distance between the 6-dimensional PaIntE vectors that underlies the DID_{PaIntE} measure is a rather coarse estimation of similarity in pitch accent realization. The relative contribution of the individual PaIntE parameters to the accommodation effect is subject to further analysis.

Epenthesis of schwa in the word ending ⟨-en⟩ was least successful in triggering accommodating behavior.

On the individual level, there are still about 15 % of participants who converge to the model speakers with respect to schwa epenthesis. However, taking the entire group into account, the only significant convergence effect emerged in the shadowing task of the natural group. Contrary to every other feature, even in the natural group there was no significant divergence effect in the post phase and the baseline level was still reached, which suggests a rather weak effect.

For both synthetic groups, no accommodation was observed, nor did the attitude towards schwa epenthesis play a role in the statistical models. As stated initially, producing a schwa in the word ending ⟨-en⟩ is rather unusual. This statement was confirmed by the fact that the vast majority of the participants (54 of 56) preferred schwa elision in the baseline production and hence shadowed [ən] stimuli. Recall that only these 54 participants were subsequently analyzed. It was claimed above that such atypicality might promote accommodation. We can assume, however, that there are limits to how atypical such a variant may be to still be considered a target for accommodation. It may have been the case that an unusual variant such as [ən] would be more likely picked up from a synthetic than a natural voice, since hyperarticulation occurs in human-computer interaction (HCI) (Burnham et al., 2010). However, synthetic voices alone do not make HCI. The present shadowing scenario lacks the need to be understood by the interlocutor, which is an important layer to HCI and spoken interaction in general. Therefore, even schwa might be picked up in a more conversational scenario, which would presumably also trigger the speaker's belief that converging to the computer leads to greater communicative success (cf. Branigan et al., 2010).

Concerning the individual accommodation behavior of the participants in the shadowing task, we found mainly convergence and maintenance, as well as some cases of individual divergence. Note that we are taking a categorical approach and do not further distinguish degrees of convergence or divergence here. The participants varied regarding the number of tested features they accommodated to. This supports our initial assumption that we would find considerable variation between the participants, which manifests itself in the form of accommodation to different subsets of the features. The two top convergers — both from the natural condition — accommodated to five out of the six examined features; for both of them it was the schwa epenthesis which they did not pick up from the model speakers.

The self-assessment of a few participants stating that they converge to dialects of other regions or consciously imitate the pronunciation of their interlocutors was not confirmed by the data.

Recall that the regionally distributed features were deliberately chosen not to be strong dialectal markers. In order to trigger the convergence to a dialect that the participants were referring to, more salient dialectal features may be required.

For a speaker to be able to intentionally imitate their conversational partners, the salience of the features in question plays a role, as does their selective realizability. In the present study, the allophonic contrasts and the schwa epenthesis lend themselves as targets for such intentional imitation. The other features, namely the pitch accent realization, the temporal structure and the distribution of spectral energy, seem to be less easily imitated intentionally, but rather a result of a more holistic high-level adjustment. This should be examined in a further study, in which participants are explicitly asked to imitate the stimuli.

It is not unexpected that the participants' self-assessment of phonetic accommodation is often inaccurate. An adaptation at the phonetic level is certainly more difficult for speakers to evaluate and quantify than, for example, an adaptation at the lexical level, where the use of certain words is easier to capture.

Another factor that may influence individual differences in accommodating behavior is the general speaker disposition, which includes aspects such as innate phonetic talent, personality traits, and cognitive abilities. Yu et al. (2013) observed, for example, that openness and a strong attention focus were positively correlated with the degree of word-initial VOT convergence during a non-conversational phonetic imitation task in English.

Lewandowski and Jilka (2019) examined accommodation of word-based amplitude envelope match in dialogs between non-native and native speakers of English. They found a higher degree of convergence among phonetically talented, more neurotic and more open speakers, as well as among speakers with higher attention scores. Convergence was found to be negatively correlated with behavioral inhibition.

This factor was not included in the present study and deserves further investigation.

Cohen Priva and Sanker (2019) have recently pointed out potential limitations of the DID measure to account for convergence in corpora of spoken interaction and, particularly, for the attempt to establish individual differences with respect to accommodating behavior. Their three main concerns are: firstly, in an extreme case of over-convergence, the DID measure might not reflect the convergence that has taken place, but suggests maintaining behavior; secondly, convergence might be underestimated for small initial distances between participant and model speaker; and lastly, the baseline measures might not be representative of the speaker's usual behavior and therefore convergence might partly be an effect of becoming closer to the latter independent of the interlocutor's influence.

Although Cohen Priva and Sanker (2019) examined a very different set of features from the one used in the present study, namely median and range of fundamental frequency, speaking rate, as well as the ratio of two types of filled pauses, and mention that their findings may be less problematic for other features, their concerns should be discussed with respect to their implications for the present study.

For the DTW cost (DID_{DTW}) and the match value (DID_{match}), the concern regarding over-convergence does not hold, since identity is an upper boundary to similarity inherent to these measures. This is not the case for the vowel quality measure (DID_{vowel}). It needs to be considered that, contrary to the one-dimensional features examined in Cohen Priva and Sanker (2019), vowel quality is a two-dimensional feature here, which makes the definition of over-convergence difficult. However, the space to move is somewhat bounded by neighboring vowel categories. Given that we systematically maximize the baseline difference between speaker and model and minimize contextual variability (see discussion below), we assume that cases of over-convergence to the extent that they will be mistaken for maintenance are unlikely to occur. For the comparison of pitch accent realizations within the six dimensions of the PaIntE model (DID_{PaIntE}), the definition of over-convergence becomes even more difficult and would have to be established for individual dimensions. The dimensions themselves differ with respect to their linguistic interpretability and presumably their relative contribution to the perception of pitch accents. Specifically, this relative contribution would have to be examined further to establish what over-convergence really means in the realm of pitch accent realization. A certain limitation for over-convergence seems to be given by the plausible and well-formed pitch accent shapes.

Regarding the concerns about variance in initial distance to the model speakers, the features examined in the present study are very different from each other. While participants are expected to exhibit small initial distances to the model speakers for the DTW cost and the match value, since we compare the same lexical items, the design of the study maximizes initial distances with respect to the vowel quality for all participants by presenting them with instances of their dispreferred variant. In the case of the allophonic variation, maximizing this distance is possible without leaving the range of normal human performance, and therefore without jeopardizing the ecological validity of the findings. The initial distances in PaIntE parameters are mainly guided by the sentence structure and an assumed default placement of pitch accents. If the initial distances vary mainly by feature and are rather balanced between speakers for the same feature, the concern of potential underestimation of convergence would be less of a problem for the analysis of the individual behavior of different participants, but more so for the different features as a whole. However, the small initial distances for the DTW cost and the match value do not exhibit the same problem as small initial distances in speaking rate, for example, since there is very little expected variability of these features as opposed to a feature like speaking rate.

In accommodation research, it is always a point of concern whether the selected baseline is representative of the speaker's usual behavior. The shadowing paradigm entails a switch of elicitation technique — in the present case from reading text to repeating speech, which is a certain limitation. In the specific shadowing experiment at hand, there may be a further effect of first exposure — in the baseline phase — versus repetition — in the shadowing task and the post phase. However, this repetition, or in other words the stability of the linguistic context throughout the experiment, also enhances the relative representativeness of the baseline productions: Although a lot of variation is possible within a vowel category, the variation occurring in our data is limited due to the comparison of identical vowel contexts (i.e., lexical items) in all three phases of the experiment; the same is true for the word-based measures and pitch accent realizations, which are themselves embedded and tested in the same sentences throughout the experiment. Moreover, allophonic variation, pitch accent realization, and word-based intensity distribution of targets embedded in short utterances are less likely affected by extreme baseline values than measures stemming from targets read and shadowed in isolation. These features also seem less prone to task-induced variation as opposed to features such as the range of fundamental frequency or speaking rate, which are likely to change over the course of an interaction as a result of familiarization with the task at hand.

While we certainly need to keep these potential limitations in mind, we hope to have shown that for certain features they do not or only partially apply. It is safe to say that the concerns have to be evaluated separately for each feature used to examine accommodation.

Coming back to the focus of the present study, namely the question whether participants behave similarly when confronted with either natural or synthetic stimuli, we can summarize that the participants of the natural condition have accommodated during the shadowing task in the expected direction, i.e., towards the model speakers, on all tested features. Remember, however, that the effect was weak for schwa epenthesis, which supports the assumption that speakers accommodate less to unusual features. Furthermore, with the exception of schwa epenthesis, the participants of the natural condition always diverged significantly from the model speakers in the post phase, partly reaching the baseline level (vowel quality, pitch accent realization, and DTW cost), partly showing a sustained convergence effect (allophonic variation [ɪç] vs. [ɪk] and match value).

The participants of the two synthetic conditions did not show an accommodation effect for schwa epenthesis. The two other cases for which no accommodation was found, are the vowel quality and pitch accent realization measures for the participants of the diphone condition. However, for the remaining features — allophonic variation [ɪç] vs. [ɪk], DTW cost, and match value — the participants of the diphone condition behaved similarly to those of the natural condition.

The participants of the HMM condition, finally, never showed the complete pattern of significant convergence in the shadowing task, complemented by significant divergence in the post phase reaching the baseline level. However, they always showed substantial movement within the overall constellation of the three phase comparisons carried out in the present study, which suggests that this general pattern — even if in a weaker form — is underlying the HMM data as well. That is, we either found convergence in the shadowing task and no significant divergence in the post phase while still reaching the baseline level (vowel quality and pitch accent realization), or no significant convergence in the shadowing task, yet divergence in the post phase, again reaching the baseline level (allophonic variation [ɪç] vs. [ɪk], DTW cost, and match value).

For the HMM voices, our initial assumption that certain phonetic features might not be clearly distinct in the synthetic stimuli proved true: with the synthesis process applied here, it was not possible to produce female HMM stimuli with a clearly distinguishable target allophone [ɛː]. The seven participants of the HMM condition with a baseline preference for [eː] therefore heard a lower total number of clear [ɛː] target allophones, namely only from the male model voice, which could be a disadvantage for the emergence of an accommodation effect. Nevertheless, we found overall convergence of vowel quality for the entire HMM group, in contrast to the diphone group, in which all participants heard clear target allophones from both model voices, but still no overall convergence occurred.

In summary, we observe the same behavior in the diphone group as in the natural group with respect to several features and no accommodation for other features. For the HMM group, we observe a similar underlying pattern as for the natural group, but in some individual phase comparisons the effect is not up to par with that of the latter. Technical differences between the synthesis methods may have contributed to the differences in performance. However, neither of the two synthesis qualities made accommodation impossible.

One aspect which needs to be taken into consideration is that the six model voices employed in the present study differ with respect to stimulus type (natural, diphone, and HMM) and sex (female and male), but of course exhibit a variety of other characteristics that may affect the degree of accommodation to them, for example their perceived naturalness and likability (see Section 2.1.1).

The participants of the natural condition gave higher ratings of naturalness to the voices they shadowed than the participants of the HMM condition. The diphone voices were rated as sounding least natural by the participants of the respective condition. This supports our initial assumption that the participants would recognize the synthetic voices as non-human. We had further speculated that this could trigger a feeling of social separation in the participants, which may lead to a reduction of the convergence effect or even to divergence. It may be the case that this factor indeed contributed to the overall weaker effects of the synthetic stimuli. However, the diphone stimuli that were rated as most unnatural sounding showed effects of similar strength as the natural stimuli for some of the examined features and it is unclear why the social component should only influence such a subset.

In terms of likability, the natural voices were rated on a par with the HMM voices, while the diphone voices again received the lowest ratings. Thus the diphone voices, on the one hand, set themselves apart from the two other voices by their lower naturalness and likability, but still triggered considerable accommodation effects for a subset of the examined features. The HMM voices, on the other hand, although being as likable as the natural voices, did not trigger the same strength of accommodation for most examined features.

Such differences need to be explored further by testing various voices of each stimulus type. However, the present experiment showed that synthetic voices, while partly reducing the strength of effects, do trigger accommodating behavior. As for the natural voices, convergence during interaction followed by divergence after the interaction is the predominant pattern.

5. Conclusion

The present shadowing experiment used natural and two types of synthetic voices (diphone- and HMM-based) to test whether native speakers of German accommodate to these voices when repeating short German sentences after them. The use of short sentences as target utterances provided a controlled context while still keeping a broad focus. The examined features pertain to different phonetic domains allowing for an extensive assessment of the participants' behavior: allophonic variation ([ɛː] vs. [eː], [ɪç] vs. [ɪk]), schwa epenthesis, realization of pitch accents (PaIntE parameters), as well as word-based temporal structure (DTW cost) and distribution of spectral energy (match value). We predicted accommodation in the form of convergence to occur with respect to these features.

The results of the individual accommodation behavior analysis need to be interpreted with caution due to potential limitations of the difference in distance (DID) measures. Concerning the predicted individual variation, we found that the participants converged to varying subsets of 0 to 5 out of the six examined features, with the most individual convergers for [ɪç] vs. [ɪk], followed by DTW cost and match value, and least for the PaIntE parameters, [ɛː] vs. [eː], and the schwa epenthesis, in that order. Very few cases of divergence were found for all features but the pitch accent realization for which no such cases occurred. Although almost half of the participants individually converged to at least three out of six features, this demonstrates that accommodation with respect to one particular feature does not necessarily predict the behavior with respect to another feature.

Describing accommodating behavior more broadly for different speaker groups is a step towards modeling the given individual variation for the HCI context in order to gain a better understanding of the user or even to implement such behavior in the computer.

On the group level, the participants of the natural condition converged to all features under examination, however very subtly so for schwa epenthesis. The participants of the diphone condition behaved similarly to the natural group with respect to several features ([ɪç] vs. [ɪk], DTW cost, and match value) or did not show any accommodation for other features. For the participants of the HMM condition, the effects were less clear overall. A significant convergence effect in the shadowing task only emerged for [ɛː] vs. [eː] and the PaIntE parameters. However, taking into account the post production, we conclude that the same pattern of convergence in the shadowing task and divergence after the shadowing task observed in the natural group for all features but schwa epenthesis, is underlying the HMM group, too.

The present experiment showed that German native speakers converge to various features ranging from segmental variation and local prosody to the word-based temporal structure and distribution of spectral energy when shadowing short sentences from natural voices. For segment-level features, like the ones we examined, accommodation had previously only been investigated in shorter, mono- or bisyllabic utterances (Babel, 2012; Dufour and Nguyen, 2013; Mitterer and Müsseler, 2013). We could show that such features are also picked up from longer utterances. The analysis of pitch accent realizations differed from an earlier approach investigating conversational speech (Schweitzer et al., 2017) in that it included the accent type. The assumption that nuclear pitch accents might cause a greater convergence effect due to their higher perceptual salience was not confirmed. An earlier approach to investigate the accommodation of the word-based distribution of spectral energy in conversational speech (Lewandowski, 2012) was expanded in this study to include the aspect of temporal structure, showing a convergence effect for the distribution of energy over spectral bands, even when convergence with respect to timing is already accounted for.

As the participants in the present experiment shadowed both a female and a male voice, we examined whether they showed a higher degree of accommodation to a model talker of the same or the opposite sex. However, no strong tendency could be observed, since only one case of increased convergence in mixed-sex pairs was found.

Regarding the comparison of natural and synthetic model speakers in speech shadowing, synthetic voices were found to induce accommodating behavior as well, but partly reduce the strength of effects found for the natural voices. One difference between the synthetic voices used in this study was that the diphone voices were perceived as generally more unnatural and unlikable than the HMM voices, which could be a source for different accommodating behavior towards them. The predominant pattern of accommodation for all voice types, however, was convergence during the interaction, followed by divergence after

the interaction. We conclude that phonetic accommodation does occur in human-computer interaction involving synthetic speech, but for the phonetic features and model voices examined here, to a lesser extent overall than in human-human interaction.

CRediT authorship contribution statement

Iona Gessinger: Methodology, Resources, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Eran Raveh:** Resources, Software, Investigation, Data curation, Writing – review & editing. **Ingmar Steiner:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Bernd Möbius:** Conceptualization, Funding acquisition, Project administration, Methodology, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project–ID MO 597/6–1,2 and STE 2363/1–1. We thank Sébastien Le Maguer for help with the generation of the HMM stimuli, Antje Schweitzer for the extraction of the PaIntE parameters and advice regarding their analysis, Bistra Andreeva for advice regarding the PaIntE analysis, Natalie Lewandowski for providing scripts used in the amplitude envelope analysis, Johannah O'Mahony and Jens Neuerburg for assistance in the annotation process, and two anonymous reviewers for very helpful comments on an earlier draft of this manuscript.

Appendix

The following listing gives an overview of the text material used in the experiment. The underlined graphemes correspond to the three variations of segmental pronunciation. The words in bold type were used for the amplitude envelope analysis.

I Target sentences

▶ [ɛː] **vs.** [eː]
(1) Die **Bestätigung** ist für Tanja.
(2) Der **Schädling** sieht aber komisch aus.
(3) Ich mag die **Qualität** deiner Tasche.
(4) Wie viel **Verspätung** hat der Zug?
(5) War das **Gerät** sehr teuer?

▶ [ɪç] **vs.** [ɪk]
(6) Es ist ganz schön **staubig** im Keller.
(7) Der **König** hält eine Rede.
(8) Ich bin **süchtig** nach Schokolade.
(9) Kommt **Essig** in den Salat?
(10) Kommt **Ludwig** heute Abend mit?

▶ [n̩] **vs.** [ən]
(11) Wir **reden** ohne Unterbrechung.
(12) Wir **besuchen** euch bald wieder.
(13) Sie **begleiten** dich zur Taufe.
(14) Sind die **Küchen** immer so groß?
(15) Sind die **Affen** denn zutraulich?

II Filler sentences

(16) Ich hätte gern zwei kleine **Brüder**.
(17) Das **Heft** war gestern noch da.
(18) Die **Glühbirne** ist leider kaputt.
(19) Sucht sich Karin eine neue **Arbeit**?
(20) Wird die **Wohnung** noch renoviert?
(21) **Sara** hat eine andere Meinung.
(22) Ich **täusche** mich so gut wie nie.
(23) Keiner glaubt diese **Geschichte**.
(24) Habt ihr das rote **Auto** erkannt?
(25) Kommt Fabian auch zu dem **Fest**?
(26) Die **Katze** weckt mich immer auf.
(27) Der **Kaffee** war ja schon kalt.
(28) Das wird ein schönes **Geschenk**.
(29) Wer fliegt heute in den **Urlaub**?
(30) **Warum** regt er sich denn so auf?

References

Abrego-Collier, C., Grove, J., Sonderegger, M., Alan, C.L., 2011. Effects of speaker evaluation on phonetic convergence. In: International Congress of Phonetic Sciences (ICPhS). Hong Kong, pp. 192–195, URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Abrego-Collier/Abrego-Collier.pdf.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: International Symposium on Information Theory. pp. 267–281.

Babel, M., 2010. Dialect divergence and convergence in New Zealand English. Lang. Soc. 39, 437–456. http://dx.doi.org/10.1017/S0047404510000400.

Babel, M., 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. J. Phonetics 40, 177–189. http://dx.doi.org/10.1016/j.wocn.2011.09.001.

Babel, M., McGuire, G., Walters, S., Nicholls, A., 2014. Novelty and social preference in phonetic accommodation. Lab. Phonol. 5 (1), 123–150. http://dx.doi.org/10.1515/lp-2014-0006.

Bailly, G., Martin, A., 2014. Assessing objective characterizations of phonetic convergence. In: Interspeech. Singapore, pp. 2011–2015, URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_2011.pdf.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Bell, L., Gustafson, J., Heldner, M., 2003. Prosodic adaptation in human-computer interaction. In: International Congress of Phonetic Sciences (ICPhS). Barcelona, pp. 2453–2456, URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_2453.pdf.

Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., Levitan, R., 2018. Prosodic entrainment and trust in human-computer interaction. In: International Conference on Speech Prosody. Poznań, pp. 220–224. http://dx.doi.org/10.21437/SpeechProsody.2018-45.

Bilous, F.R., Krauss, R.M., 1988. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. Lang. Commun. 8, 183–194. http://dx.doi.org/10.1016/0271-5309(88)90016-x.

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

Borrie, S.A., Lubold, N., Pon-Barry, H., 2015. Disordered speech disrupts conversational entrainment: a study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. Front. Psychol. 6 (1187), http://dx.doi.org/10.3389/fpsyg.2015.01187.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., 2010. Linguistic alignment between people and computers. J. Pragmat. 42 (9), 2355–2368. http://dx.doi.org/10.1016/j.pragma.2009.12.012.

Burnham, D., Joeffry, S., Rice, L., 2010. 'D-o-e-s-Not-C-o-m-p-u-t-e': vowel hyperarticulation in speech to an auditory-visual avatar. In: Auditory-Visual Speech Processing. AVSP, Hakone, URL: https://www.isca-speech.org/archive/avsp10/papers/av10_P18.pdf.

Cohen Priva, U., Sanker, C., 2019. Limitations of difference-in-difference for measuring convergence. Lab. Phonol. 10 (1), 15. http://dx.doi.org/10.5334/labphon.200.

Coles-Harris, E.H., 2017. Perspectives on the motivations for phonetic convergence. Lang. Linguist. Compass 11 (12), http://dx.doi.org/10.1111/lnc3.12268.

Coulston, R., Oviatt, S., Darves, C., 2002. Amplitude convergence in children's conversational speech with animated personas. In: International Conference on Spoken Language Processing. ICSLP, Denver, pp. 2689–2692, URL: https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_2689.pdf.

Delvaux, V., Soquet, A., 2007. Inducing imitative phonetic variation in the laboratory. In: International Congress of Phonetic Sciences (ICPhS). Saarbrücken, pp. 369–372, URL: http://www.icphs2007.de/conference/Papers/1318/1318.pdf.

Dias, J.W., Rosenblum, L.D., 2016. Visibility of speech articulation enhances auditory phonetic convergence. Attent. Percept. Psychophys. 78 (1), 317–333. http://dx.doi.org/10.3758/s13414-015-0982-6.

Dudenredaktion, 2015. Duden - Das Aussprachewörterbuch: Betonung und Aussprache von über 132.000 Wörtern und Namen. In: Duden - Deutsche Sprache in 12 Bänden, vol. 6, Bibliographisches Institut GmbH, Mannheim.

Dufour, S., Nguyen, N., 2013. How much imitation is there in a shadowing task? Front. Psychol. 4 (346), http://dx.doi.org/10.3389/fpsyg.2013.00346.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van der Vrecken, O., 1996. 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, Vol. 3. ICSLP, Philadelphia, PA, pp. 1393–1396. http://dx.doi.org/10.1109/icslp.1996.607874.

Ellbogen, T., Schiel, F., Steffen, A., 2004. The BITS speech synthesis corpus for German. In: International Conference on Language Resources and Evaluation. LREC, Lisbon, pp. 2091–2094, URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/72.pdf.

Fowler, C.A., Brown, J.M., Sabadini, L., Weihing, J., 2003. Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. J. Mem. Lang. 49 (3), 396–413. http://dx.doi.org/10.1016/S0749-596X(03)00072-X.

Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., Steiner, I., 2017. Shadowing synthesized speech – segmental analysis of phonetic convergence. In: Interspeech. Stockholm, pp. 3797–3801. http://dx.doi.org/10.21437/Interspeech.2017-1433.

Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., Steiner, I., 2018. Convergence of pitch accents in a shadowing task. In: International Conference on Speech Prosody. Poznań, pp. 225–229. http://dx.doi.org/10.21437/SpeechProsody.2018-46.

Giles, H., 1973. Accent mobility: a model and some data. Anthropol. Linguist. 87–105.

Giles, H., Coupland, N., Coupland, J., 1991. Accommodation theory: communication, context, and consequence. In: Giles, H., Coupland, J., Coupland, N. (Eds.), Contexts of Accommodation: Developments in Applied Sociolinguistics. Cambridge University Press, pp. 1–68. http://dx.doi.org/10.1017/cbo9780511663673.001.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105 (2), 251–279. http://dx.doi.org/10.1037/0033-295X.105.2.251.

Gregory, S.W., Webster, S., 1996. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. J. Pers. Soc. Psychol. 70 (6), 1231–1240. http://dx.doi.org/10.1037/0022-3514.70.6.1231.

Honorof, D.N., Weihing, J., Fowler, C.A., 2011. Articulatory events are imitated under rapid shadowing. J. Phonetics 39 (1), 18–38. http://dx.doi.org/10.1016/j.wocn.2010.10.007.

Jagdfeld, N., Baumann, S., 2011. Order effects on the perception of relative prominence. In: International Congress of Phonetic Sciences (ICPhS). Hong Kong, pp. 958–961, URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Jagdfeld/Jagdfeld.pdf.

Kiesewalter, C., 2019. Zur subjektiven Dialektalität regiolektaler Aussprachemerkmale des Deutschen. Franz Steiner Verlag.

King, S., Black, A.W., Taylor, P., Caley, R., Clark, R., 1999. Edinburgh speech tools library. Version 1.2. URL: http://www.cstr.ed.ac.uk/projects/speech_tools/.

Kisler, T., Reichel, U., Schiel, F., 2017. Multilingual processing of speech via web services. Comput. Speech Lang. 45, 326–347. http://dx.doi.org/10.1016/j.csl.2017.01.005.

Kleiner, S., 2011. Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG). Unter Mitarbeit von Ralf Knöbl. URL: http://prowiki.ids-mannheim.de/bin/view/AADG/.

Krauss, R.M., Pardo, J.S., 2004. Is alignment always the result of automatic priming? Behav. Brain Sci. 27 (2), 203–204. http://dx.doi.org/10.1017/S0140525X0436005X.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 82 (13), 1–26. http://dx.doi.org/10.18637/jss.v082.i13.

Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., Narayanan, S.S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Interspeech. Makuhari, pp. 793–796, URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_0793.pdf.

Levitan, R., Beňuš, Š., Gálvez, R.H., Gravano, A., Savoretti, F., Trnka, M., Weise, A., Hirschberg, J., 2016. Implementing acoustic-prosodic entrainment in a conversational avatar. In: Interspeech. San Francisco, CA, pp. 1166–1170. http://dx.doi.org/10.21437/Interspeech.2016-985.

Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., Nenkova, A., 2012. Acoustic-prosodic entrainment and social behavior. In: NAACL Conference on Human Language Technologies. pp. 11–19, URL: https://www.aclweb.org/anthology/N12-1002.pdf.

Levitan, R., Hirschberg, J., 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In: Interspeech. Florence, pp. 3081–3084, URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_3081.pdf.

Lewandowski, N., 2012. Talent in Nonnative Phonetic Convergence (Ph.D. thesis). Universität Stuttgart, http://dx.doi.org/10.18419/opus-2858.

Lewandowski, N., Jilka, M., 2019. Phonetic convergence, language talent, personality & attention. Front. Commun. 4 (18), http://dx.doi.org/10.3389/fcomm.2019.00018.

Lubold, N., Walker, E., Pon-Barry, H., 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In: ACM/IEEE International Conference on Human-Robot Interaction. HRI, pp. 255–262. http://dx.doi.org/10.1109/HRI.2016.7451760.

Manson, J.H., Bryant, G.A., Gervais, M.M., Kline, M.A., 2013. Convergence of speech rate in conversation predicts cooperation. Evol. Hum. Behav. 34 (6), 419–426. http://dx.doi.org/10.1016/j.evolhumbehav.2013.08.001.

Michalsky, J., Schoormann, H., 2017. Pitch convergence as an effect of perceived attractiveness and likability. In: Interspeech. Stockholm, pp. 2253–2256. http://dx.doi.org/10.21437/Interspeech.2017-1520.

Miller, R.M., Sanchez, K., Rosenblum, L.D., 2013. Is speech alignment to talkers or tasks? Attent. Percept. Psychophys. 75 (8), 1817–1826. http://dx.doi.org/10.3758/s13414-013-0517-y.

Mitterer, H., Ernestus, M., 2008. The link between speech perception and production is phonological and abstract: evidence from the shadowing task. Cognition 109 (1), 168–173. http://dx.doi.org/10.1016/j.cognition.2008.08.002.

Mitterer, H., Müsseler, J., 2013. Regional accent variation in the shadowing task: evidence for a loose perception-action coupling in speech. Attent. Percept. Psychophys. 75 (3), 557–575. http://dx.doi.org/10.3758/s13414-012-0407-8.

Möbius, B., 1993. Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen. Niemeyer.

Möhler, G., 1998. Describing intonation with a parametric model. In: International Conference on Spoken Language Processing. Sydney, pp. 2851–2854, URL: https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0205.pdf.

Möhler, G., Conkie, A., 1998. Parametric modeling of intonation using vector quantization. In: ESCA/COCOSDA Workshop on Speech Synthesis. SSW, Blue Mountains, Australia, pp. 311–316, URL: https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_311.pdf.

Namy, L.L., Nygaard, L.C., Sauerteig, D., 2002. Gender differences in vocal accommodation: the role of perception. J. Lang. Soc. Psychol. 21 (4), 422–432. http://dx.doi.org/10.1177/026192702237958.

Nass, C., Moon, Y., 2000. Machines and mindlessness: social responses to computers. J. Soc. Issues 56 (1), 81–103. http://dx.doi.org/10.1111/0022-4537.00153.

Nass, C., Steuer, J., Tauber, E.R., 1994. Computers are social actors. In: SIGCHI Conference on Human Factors in Computing Systems. pp. 72–78. http://dx.doi.org/10.1145/191666.191703.

Nguyen, N., Dufour, S., Brunellière, A., 2012. Does imitation facilitate word recognition in a non-native regional accent? Front. Psychol. 3 (480), http://dx.doi.org/10.3389/fpsyg.2012.00480.

Nielsen, K.Y., 2011. Specificity and abstractness of VOT imitation. J. Phonetics 39 (2), 132–142. http://dx.doi.org/10.1016/j.wocn.2010.12.007.

Olive, J., van Santen, J., Möbius, B., Shih, C., 1998. Synthesis. In: Sproat, R. (Ed.), Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic, Dordrecht, pp. 191–228 (Chap. 7).

Oviatt, S., Darves, C., Coulston, R., 2004. Toward adaptive conversational interfaces: modeling speech convergence with animated personas. ACM Trans. Comput.-Human Interact. 11, 300–328. http://dx.doi.org/10.1145/1017494.1017498.

Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382–2393. http://dx.doi.org/10.1121/1.2178720.

Pardo, J.S., Urmanche, A., Wilman, S., Wiener, J., 2017. Phonetic convergence across multiple measures and model talkers. Attent. Percept. Psychophys. 79 (2), 637–659. http://dx.doi.org/10.3758/s13414-016-1226-0.

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. J. Phonetics 69, 1–11. http://dx.doi.org/10.1016/j.wocn.2018.04.001.

Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behav. Brain Sci. 27 (2), 169–190. http://dx.doi.org/10.1017/S0140525X04450055.

Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behav. Brain Sci. 36 (4), 329–347. http://dx.doi.org/10.1017/s0140525x12001495.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria. URL: https://www.r-project.org.

Reeves, B., Nass, C., 1996. The media equation: how people treat computers, television, and new media like real people and places. Comput. Math. Appl. 33 (5), 128. http://dx.doi.org/10.1016/s0898-1221(97)82929-x.

Schweitzer, A., Lewandowski, N., 2014. Social factors in convergence of F1 and F2 in spontaneous speech. In: International Seminar on Speech Production. Cologne, http://dx.doi.org/10.13140/2.1.3709.5689.

Schweitzer, A., Lewandowski, N., Dogil, G., 2014. Advancing corpus-based analyses of spontaneous speech: switch to GECO!. In: LabPhon. Tokyo.

Schweitzer, A., Lewandowski, N., Duran, D., Dogil, G., 2015. Attention, please! Expanding the GECO database. In: International Congress of Phonetic Sciences (ICPhS). Glasgow, URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0620.pdf.

Schweitzer, A., Möhler, G., Dogil, G., Möbius, B., 0000. The PaIntE model of intonation. In: Barnes, J.A., Shattuck-Hufnagel, S. (Eds.), Prosodic Theory and Practice. MIT Press (in press).

Schweitzer, K., Walsh, M., Schweitzer, A., 2017. To see or not to see: interlocutor visibility and likeability influence convergence in intonation. In: Interspeech. Stockholm, pp. 919–923. http://dx.doi.org/10.21437/Interspeech.2017-1248.

Shepard, C.A., Giles, H., Le Poire, B.A., 2001. Communication accommodation theory. In: Robinson, W.P., Giles, H. (Eds.), The New Handbook of Language and Social Psychology. Wiley, pp. 33–56.

Shockley, K., Sabadini, L., Fowler, C.A., 2004. Imitation in shadowing words. Percept. Psychophys. 66 (3), 422–429. http://dx.doi.org/10.3758/BF03194890.

Suzuki, N., Katagiri, Y., 2007. Prosodic alignment in human-computer interaction. Connect. Sci. 19 (2), 131–141. http://dx.doi.org/10.1080/09540090701369125.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). Speech Cod. Synth. 497–518.

Taylor, P., 2009. Text-to-Speech Synthesis. Cambridge University Press, pp. 422–445. http://dx.doi.org/10.1017/CBO9780511816338 (Chap. 14).

Wade, T., Dogil, G., Schütze, H., Walsh, M., Möbius, B., 2010. Syllable frequency effects in a context-sensitive segment production model. J. Phonetics 38 (2), 905–945. http://dx.doi.org/10.1016/j.wocn.2009.10.004.

Walker, A., Campbell-Kibler, K., 2015. Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. Front. Psychol. 6 (546), http://dx.doi.org/10.3389/fpsyg.2015.00546.

Yu, A.C.L., Abrego-Collier, C., Sonderegger, M., 2013. Phonetic imitation from an individual-difference perspective: subjective attitude, personality and autistic traits. PLoS One 8 (9), e74746. http://dx.doi.org/10.1371/journal.pone.0074746.

Zen, H., Toda, T., 2005. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In: European Conference on Speech Communication and Technology. Eurospeech, Lisbon, URL: http://www.festvox.org/blizzard/bc2005/IS052192.PDF.