Journal of Phonetics 86 (2021) 101029

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Phonetic accommodation in interaction with a virtual language learning tutor: A Wizard-of-Oz study

Iona Gessinger [a,*], Bernd Möbius [a], Sébastien Le Maguer [b], Eran Raveh [a], Ingmar Steiner [c]

[a] Department of Language Science and Technology, Saarland University, Saarbrücken, Germany
[b] ADAPT Centre, Sigmedia Lab, EE Engineering, Trinity College Dublin, Ireland
[c] audEERING GmbH, Gilching, Germany

ARTICLE INFO

Article history:
Received 5 February 2020
Received in revised form 11 January 2021
Accepted 13 January 2021
Available online 14 March 2021

Keywords:
Phonetic accommodation
Human-computer interaction
Wizard-of-Oz experiment
Synthetic speech
German

ABSTRACT

We present a Wizard-of-Oz experiment examining phonetic accommodation of human interlocutors in the context of human-computer interaction. Forty-two native speakers of German engaged in dynamic spoken interaction with a simulated virtual tutor for learning the German language called *Mirabella*. Mirabella was controlled by the experimenter and used either natural or hidden Markov model-based synthetic speech to communicate with the participants. In the course of four tasks, the participants' accommodating behavior with respect to wh-question realization and allophonic variation in German was tested. The participants converged to Mirabella with respect to modified wh-question intonation, i.e., rising F_0 contour and nuclear pitch accent on the interrogative pronoun, and the allophonic contrast [ɪç] vs. [ɪk] occurring in the word ending ⟨-ig⟩. They did not accommodate to the allophonic contrast [ɛː] vs. [eː] as a realization of the long vowel ⟨-ä-⟩. The results did not differ between the experimental groups that communicated with either the natural or the synthetic speech version of Mirabella. Testing the influence of the "Big Five" personality traits on the accommodating behavior revealed a tendency for neuroticism to influence the convergence of question intonation. On the level of individual speakers, we found considerable variation with respect to the degree and direction of accommodation. We conclude that phonetic accommodation on the level of local prosody and segmental pronunciation occurs in users of spoken dialog systems, which could be exploited in the context of computer-assisted language learning.

1. Introduction

Phonetic accommodation as the phenomenon of adapting our own speech output to the speech input we receive from our interlocutors has been documented for human–human interaction (HHI) (e.g., Pardo, 2006; Levitan & Hirschberg, 2011; Lewandowski, 2012). The phenomenon includes convergence, where conversational partners become more similar to each other, as well as divergence, where they move away from each other.

An increasing amount of spoken interactions with computers in our everyday life raises the question whether this phenomenon is relevant for human–computer interaction (HCI) as well. One aspect that would imply such relevance is the supposed contribution of phonetic accommodation to communicative success and dialog quality (e.g., Manson, Bryant, Gervais, & Kline, 2013; Pickering & Garrod, 2004; Borrie, Lubold, & Pon-Barry, 2015; Lee et al., 2010), since making the communication with a computer a more pleasant experience for the human user is one of the overarching objectives in HCI research.

A more concrete situation in which phonetic accommodation, here specifically convergence, would be desirable for the user of a spoken dialog system (SDS), is a computer-assisted language learning (CALL) context. Provided that the speech output of the CALL application is of native-like quality, converging to it would lead to an improvement in the production of the learned language on the part of the user. Especially the pronunciation of speech segments and the realization of prosodic phenomena such as question intonation, lend themselves as targets for accommodation in this context, as there exist clearly defined standard realizations for these features. In the present study, we examine such features in a simulated CALL scenario.

We apply the Wizard-of-Oz (WOz) method, in which participants interact with a supposedly intelligent SDS while an experimenter is controlling the output of the system behind

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

https://doi.org/10.1016/j.wocn.2021.101029
0095-4470/

the scenes (Kelley, 1984; Dahlbäck, Jönsson, & Ahrenberg, 1993). We examine phonetic accommodation with respect to local prosody, more precisely the placement of the nuclear pitch accent in wh-questions and the final intonation contour following this nuclear accent, as well as with respect to segmental pronunciation, in particular the German allophone pairs [ɛː] vs. [eː] as a realization of the long vowel ⟨-ä-⟩, e.g., *Mädchen* (girl), and [ɪç] vs. [ɪk] occurring in the word ending ⟨-ig⟩, e.g., *Kön**ig*** (king). The different variants of these features are accepted in Standard German.

To motivate the interaction and simulate a situation similar to that one might encounter in everyday life, the SDS is presented to the participants as a tutoring system for learning German as a foreign language. We named the tutor *Mirabella*.

In a first condition, Mirabella's utterances consist of natural speech, pre-recorded by a female native speaker of German. Although natural speech is used, a key component of HCI is present in this scenario, i.e., the belief of the participants that they are interacting with a computer (Branigan, Pickering, Pearson, & McLean, 2010).

In a second condition, the natural stimuli are replaced by hidden Markov model (HMM)-based synthetic stimuli to investigate whether synthetic speech triggers similar accommodating behavior in human interlocutors as does natural speech. The synthetic stimuli were refined by using the pitch contours and segment durations from the natural recordings (see section 2.3).

Since we assume that the factors triggering phonetic accommodation in HHI also apply to the interaction with a virtual persona like Mirabella, whether she speaks with a genuine human or a synthetic voice, we expect accommodation with regard to the examined features to occur in both experimental groups. This expectation is motivated in the remainder of the introduction.

To the best of our knowledge, this is the first study examining accommodation to question realization and segmental pronunciation with the WOz method. The experiment was first introduced in Gessinger, Möbius, Fakhar, Raveh, and Steiner (2019b) with a smaller number of experimental subjects. Gessinger, Möbius, Andreeva, and Raveh (2019a) reports the results of the natural speech condition only. The present article describes the WOz experiment in its entirety and extends it by the results of the synthetic speech condition. We also present an analysis on the influence of the speakers' personality traits on phonetic accommodation for the participants of both conditions, which has not previously been published.

1.1. Theoretical frameworks

There are two main theories about the underlying reasons for the occurrence of phonetic accommodation in spoken interaction. The *Interactive Alignment Model (IAM)* (Pickering & Garrod, 2004; Pickering & Garrod, 2013) postulates a priming mechanism that automatically leads to convergence between interlocutors during a conversation. The *Communication Accommodation Theory (CAT)* (Giles, 1973; Giles, Coupland, & Coupland, 1991; Shepard et al., 2001) suggests that phonetic accommodation has a strong social component: by converging to, or diverging from, our conversational partner, we are communicating a closer or more distant relationship, respectively. Specific social factors that have been found to influence the strength and direction of phonetic accommodation in HHI include, for example, the perceived attractiveness and likability of an interlocutor (e.g., Babel, McGuire, Walters, & Nicholls, 2014; Schweitzer & Lewandowski, 2014; Michalsky & Schoormann, 2017) and the hierarchical relationship between speaker and interlocutor (e.g., Gregory & Webster, 1996).

Assuming an underlying model of phonetic accommodation that combines the *automatic approach* (IAM) and the *social approach* (CAT), as for example suggested by Krauss and Pardo (2004), Babel (2010), Lewandowski (2012), Coles-Harris (2017), leads us to believe that convergence represents the unmarked behavior. Divergence would then be expected in cases where a speaker either aims to increase social distance or to counteract extreme behavior of an interlocutor, presumably hoping for them to converge, such as in slowing down a very fast-talking speaker. In these cases, the unmediated tendency to converge might be superseded by a more dominant social motivation to diverge.

While it does not matter for the *automatic approach* whether we are communicating with a fellow human or a computer – converging behavior is expected in both cases –, the *social approach* suggests that it is crucial for the interlocutor to be perceived as a social actor, an attribute that we may not intuitively assign to a computer. Nass et al., 1994, however, argue that the latter is indeed the case for computers, too. This concept was established as the *Computers are Social Actors* paradigm (Reeves & Nass, 1996; Nass & Moon, 2000). As speech synthesis development strives for more naturalness and interactions with SDSs evolve from simple commands to free conversations, it can be assumed that this status is becoming more established. We therefore believe that a virtual interlocutor should in principle be able to trigger phonetic accommodation in a human speaker.

If the speaker believes that convergence is particularly beneficial for successful communication with a computer, for example because the computer relies on a certain speaking style to understand, the accommodation effect in HCI may be even greater than in communication with a fellow human (Branigan et al., 2010).

1.2. Related work

In HHI phonetic accommodation was observed for the perceptual similarity of utterances (e.g., Miller, Sanchez, & Rosenblum, 2013; Babel et al., 2014; Dias & Rosenblum, 2016); holistic acoustic measures such as the long-term average spectrum (Gregory & Webster, 1996), mel-frequency cepstral coefficients (Delvaux & Soquet, 2007), and amplitude envelopes (Lewandowski, 2012; Lewandowski & Jilka, 2019); global acoustic-prosodic measures such as turn-based F_0, intensity, or speaking rate (e.g., Ward & Litman, 2007; Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014); and local phenomena such as vowel quality (e.g., Babel, 2012; Dufour & Nguyen, 2013), voice onset time (VOT) (e.g., Nielsen, 2011; Yu, Abrego-Collier, & Sonderegger, 2013), pitch accents (Schweitzer et al., 2017), or allophonic variation, for example, the realization of the German word ending ⟨-ig⟩ as [ɪç] or [ɪk] (Mitterer & Müsseler, 2013).

The body of literature exploring whether humans also accommodate to the speech output of spoken dialog systems (SDSs) is growing. However, the phonetic features that have been examined in this context so far are mainly of global acoustic-prosodic nature, such as global F_0, intensity, pitch range, and speaking rate (e.g., Bell, Gustafson, & Heldner, 2003; Oviatt, Darves, & Coulston, 2004; Staum Casasanto, Jasmin, & Casasanto, 2010; Gijssels, Staum Casasanto, Jasmin, Hagoort, & Casasanto, 2016; Raveh, Siegert, Steiner, & Gessinger, 2019).

With the exception of Raveh et al. (2019), who studied a commercially available SDS without manipulating its speech output, all of the mentioned HCI studies applied the Wizard-of-Oz (WOz) method to simulate intelligent SDSs. In a WOz setup, users think that they are interacting with an autonomous system, but in reality it is the *wizard*, i.e., the experimenter, who takes the decisions about the system's responses (Kelley, 1984; Dahlbäck et al., 1993). This allows to create a dynamic conversational exchange between the users and the simulated system while having direct control over the speech output of the latter.

These studies demonstrated that humans exhibit accommodating behavior with respect to global acoustic-prosodic features when conversing with virtual interlocutors.[1] Combined with the above-mentioned results from HHI research, this supports our assumption that such behavior may also occur with the more locally anchored phonetic features investigated in the present study.

1.2.1. Voice type

Studies comparing the use of natural and synthetic speech in SDSs for tutoring showed that the pre-recorded natural version of a system is sometimes favored by users and can even be more conducive to learning than its synthetic counterpart (e.g., Baylor, Ryu, & Shen, 2003; Atkinson, Mayer, & Merrill, 2005). Forbes-Riley, Litman, Silliman, and Tetreault (2006), in contrast, found almost no influence of a virtual tutor's voice on learning gain, system usability, or dialog efficiency.

The perception of the virtual interlocutor's voice is also influenced by whether the agent is graphically represented. In Baylor et al. (2003), students were most motivated when interacting with a graphically animated agent that spoke with a synthetic voice, or with an agent that had a natural human voice and was not graphically animated.

The above mentioned WOz studies examining phonetic accommodation used either manipulated natural speech recordings (Bell et al., 2003; Staum Casasanto et al., 2010; Gijssels et al., 2016) or synthesized speech (Oviatt et al., 2004), and all of them used embodied graphical agents to represent the computer interlocutor, be they humanoid (*Cloddy Hans* in Bell et al., 2003; *VIRTUO*/*VIRTUA* in Staum Casasanto et al., 2010; Gijssels et al., 2016) or zoomorphic (various marine animals in Oviatt et al., 2004).

In the HHI context, it has been suggested that the atypicality of an interlocutor's voice may promote phonetic convergence (Babel et al., 2014). A synthetic voice may be interpreted by listeners as untypical of a human voice and therefore lead to a stronger accommodation effect than a natural voice – unless there is a limit to how untypical a voice can sound before convergence is inhibited or even turned into divergence.

In this context, it is of interest to investigate the influence of the voice type on accommodating behavior in a direct comparison of the same SDS, while excluding the possible effect of the virtual interlocutor's visual appearance by using only their voice for the interaction.

Thomason et al. (2013) compared the accommodation of intensity and F_0 features for students interacting with the ITSPOKE tutoring dialog system (Litman & Silliman, 2004) using either a pre-recorded, i.e., natural, or a synthesized voice. They reported a tendency for F_0 related features to show more convergence in the natural voice condition.

In a previous shadowing experiment where participants listened to and then repeated natural and synthetic stimuli, we found significant convergence effects for all examined features (allophonic variations [ɪç]/[ɪk] and [ɛː]/[eː], schwa epenthesis, realization of pitch accents, word-based temporal structure, distribution of spectral energy) when shadowing natural stimuli. Shadowing synthetic stimuli, while partly reducing the strength of the effects found for the natural voices, triggered accommodating behavior as well (Gessinger et al., 2018; Gessinger, Raveh, Steiner, & Möbius, 2021; Gessinger, Raveh, Le Maguer, Möbius, & Steiner, 2017). We conclude that humans are generally responsive to phonetic variation in synthetic voices, yet there might be limitations to the perceptibility of phonetic detail in synthesized utterances. This concern will likely become less relevant with improving quality of text-to-speech synthesis.

1.2.2. Speaker disposition

It is commonly observed that different speakers exhibit different degrees of phonetic accommodation (e.g., Pardo et al., 2018). Exploring the individual differences between speakers causing this variation is becoming a central point of accommodation research.

One factor that may contribute to the individual differences in accommodating behavior is the general speaker disposition, which includes aspects such as innate phonetic talent, personality traits, and cognitive abilities. Only a few studies have investigated these aspects to date.

Yu et al. (2013) found, for example, that openness and a strong attention focus were positively correlated with the degree of word-initial VOT convergence of speakers in 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 observed 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.

These results suggest that it is promising to further investigate the influence of speaker disposition on phonetic accommodation.

[1] Another line of research focuses on acoustic-prosodic accommodation on the part of the SDS and its effect on the way the virtual agents are perceived by human users with respect to traits such as social presence, likability, competence, or trustworthiness (e.g., Lubold, Walker, & Pon-Barry, 2016; Levitan et al., 2016; Gauder, Reartes, Gálvez, Beňuš, & Gravano, 2018; Beňuš et al., 2018). Developing computers who are themselves able to phonetically accommodate to the user is complementary to the research presented here. Specifically for the application in CALL, a synergy of the computer recognizing erroneous productions of the user, diverging from them to give room for accommodation and, eventually, the user converging to the computer, would probably be an ideal solution.

1.3. Hypotheses and predictions

The virtual language learning tutor Mirabella was designed to lead a friendly conversation, i.e., she explains the tasks to the participants, asks whether everything was understood, praises and encourages the participants, and does not exhibit extreme behavior that would provoke counteraction. Therefore, we have no reason to believe that the participants would show divergence in conversation with Mirabella, for example in order to increase the social distance to her. We expect mainly converging behavior on the part of the participants. To assess the impression that the participants have of Mirabella, we collect simple scores for her perceived likability and competence, as well as her intelligibility and response time after the experiment (see section 3.1).

The present study compares two voice types, i.e., a natural and a synthetic voice, in their ability to trigger accommodating behavior in users of a SDS. As discussed above, it was shown in WOz experiments using embodied graphical agents that both voice types can individually lead to phonetic accommodation of global acoustic-prosodic features (Bell et al., 2003; Oviatt et al., 2004; Staum Casasanto et al., 2010; Gijssels et al., 2016). We expect that accommodation also occurs for the more locally anchored phonetic features investigated in the present study. See section 2.1 for more details on the tested features and specific predictions.

In the case of Mirabella, the two voice types are directly compared using the same SDS and a possible effect of the virtual interlocutor's visual appearance is excluded, as she communicates only through her voice. We expect both versions of Mirabella to trigger accommodating behavior in the participants. The natural version may be at an advantage, since it has been shown that natural voices are often preferred in tutoring settings, specifically so when there is no accompanying graphical representation of the virtual interlocutor (Baylor et al., 2003; Atkinson et al., 2005). Moreover, the natural version may be more readily perceived as a social actor, which according to CAT would promote accommodation. In addition, our own prior shadowing experiment has shown a stronger accommodation effect for natural voices compared to different synthetic voices (Gessinger et al., 2018; Gessinger, Raveh, Steiner, & Möbius, 2021; Gessinger, Raveh, Le Maguer, Möbius, & Steiner, 2017). However, the synthetic version of Mirabella may have an advantage in that it sounds rather atypical, which has been shown to increase convergence for some speakers (Babel et al., 2014). Furthermore, the synthetic version may be perceived as more machine-like and therefore more likely to benefit from convergence (Branigan et al., 2010).

Although the experiment is situated in a language learning context, the participants of the present study are native speakers of German (see section 2.4). We therefore essentially investigate L1-L1 communication.[2] The question remains open, however, whether Mirabella is actually perceived as a "native speaker" of German by the participants. It is conceivable that a SDS which does not possess complete linguistic flexibility is not regarded as a fully competent speaker of the language in question and that, with respect to accommodation, similar mechanisms apply as in dialogs with non-native speakers (see Costa, Pickering, & Sorace (2008) for an overview). The belief in the limited linguistic competence of the addressee may, for example, lead to a higher degree of adaptation on the part of the participants. In contrast, native speakers are likely to be confident in their own pronunciation and may perceive the SDS as hierarchically inferior – two aspects that contradict a strong tendency towards convergence (see Gregory & Webster (1996) for hierarchy).

Apart from the general expectation to find convergence to Mirabella at the group level, we predict that the individual participants will differ considerably in their behavior, as was the case in previous studies (Gessinger, Raveh, Steiner, & Möbius, 2021; Pardo et al., 2018). To further investigate a possible source of this variation, we include the "Big Five" personality traits collected with the NEO Five Factor Inventory (NEO-FFI) in the analysis (see section 3.6). Openness and neuroticism have been suggested to promote convergence in the context of phonetic accommodation (Yu et al., 2013; Lewandowski & Jilka, 2019). We therefore expect a possible influence of these factors on our data.

2. Material and methods

The WOz experiment is presented to the participants as an interaction with an application for learning the German language. This resembles a realistic use case as it simulates a scenario from the growing field of CALL. The experiment is disguised as a test of the application before it is deployed to learners of German, which motivates the situation for the participants and shifts the focus from the participants being tested themselves to the system being under scrutiny.

The system introduces itself as a female tutor for German as a foreign language called *Mirabella*. During the experiment, the participants only interact with Mirabella's voice; she is not represented by an avatar. All utterances available to the *wizard*, i.e., the experimenter, to choose from during the experiment were either pre-recorded by a native speaker of German or pre-synthesized (see section 2.3). These stimuli are manually played back to the participants by the experimenter, while the participants believe to interact with a fully automatic SDS which understands their speech input and reacts accordingly.

During the interaction with Mirabella, the participants are seated in front of a monitor in a sound-attenuated booth and recorded with a sampling rate of 48kHz using a stationary cardioid microphone. Mirabella's utterances are played to the participants over headphones. The recordings are followed by a questionnaire about the participants themselves and their opinion about Mirabella, as well as the German version of the NEO-FFI (Borkenau & Ostendorf, 2007) to collect information about their personality traits.

2.1. Tasks and tested features

The interaction with Mirabella consists of four tasks and lasts about 30 min (including short breaks after tasks 1 and 3). Mirabella explains the tasks to the participants and takes part in them. The interaction is supported by visualization of the tasks on a screen. The features tested for accommodating

[2] For an extension of this experiment to L1-L2 communication, see Gessinger, Möbius, Andreeva, Raveh, & Steiner, 2020.

behavior are the intonation of constituent questions such as *"Wo hat sich der Hase versteckt?"* (Where did the rabbit hide?) and the variation of the German allophone pairs [ɛː] vs. [eː] as a realization of the long vowel ⟨-ä-⟩[3] in stressed syllables, e.g., *Käse* (cheese), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩[4], e.g., *Honig* (honey).

The first two tasks familiarize the participants with the system and the text material occurring in the experiment and elicit baseline productions of the target utterances.

The two tasks testing for accommodation are a *question-and-answer game* of two rounds (task 3), in which the participants and Mirabella take turns asking each other questions about the location of the animals on the screen, and a *map task* of four rounds (task 4), in which the participants have to describe their way to a destination while asking Mirabella about the hidden objects they encounter.

2.1.1. Task 1 – allophonic variation, baseline

This task ensures that the participants know all 71 German words (24 targets – 12 per allophonic contrast – and 47 fillers) they need to recognize during the experiment (see Appendix B) and reveals which versions of [ɛː] vs. [eː] and [ɪç] vs. [ɪk] they produce naturally.

The set of words contains 35 nouns, which are presented to the participants as pictures, and 36 adjectives, which are presented in their English translations. The participants name the pictures and translate the English adjectives to German by pronouncing them in the carrier sentence "*Das Wort* ⟨**item**⟩ *kenne ich.*" (I know the word ⟨item⟩.).

In case they do not recognize an item, they state: "*Das Wort kenne ich nicht.*" (I do not know the word.).

In the event that an item is not recognized (correctly), the participants are provided with the initial letter of the word in question and the opportunity to try again (see Fig. 1). If they fail a second time, the word is presented in written form and needs to be read out loud to move on with the task. That way, while avoiding to present the written form as long as possible, all items are uttered by every participant. In 88% of all cases the participants of the present study recognized the item correctly at the first attempt, for 9% of the cases the first letter was provided, and in 3% of the cases, only about half of which were target items, the word was read.

In this first task, Mirabella accepts allophonic variation in order to avoid that the participants change their pronunciation simply because they were not understood. But she only accepts the expected target words, i.e., no synonyms, in order to be perceived as a non–human interlocutor who does not have the full range of human linguistic flexibility.

The individual realizations of ⟨-ä-⟩ and ⟨-ig⟩ are auditorily categorized as [ɛː] or [eː] and [ɪç] or [ɪk], respectively, by the experimenter, in the present case the first author of the study. The categorization has to be performed in real-time and on the basis of the phonetician's auditory impression in order to ensure a smooth and seamless interaction with Mirabella for the participant. The validity of these online annotations is evaluated in section 3.4 for [ɛː]/[eː] and in section 3.5 for [ɪç]/[ɪk].

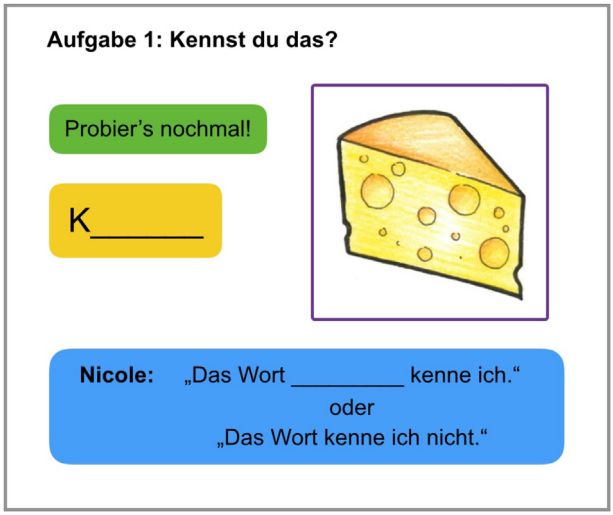


Fig. 1. Task 1 – Do you know that? Picture naming and translation task to familiarize the participants with the text material and elicit baseline productions of the target items. Here: second attempt to name the *cheese* picture. The first letter of the target word *Käse* is provided as a hint. The blue box contains the carrier sentence. *Probier's nochmal* (try again)*!*. (For a color version of the figure, the reader is referred to the web version of this article).

Note that we consider fricative variants such as [ʃ] or [ɕ] as part of the [ɪç] category.

The occurrence of the allophones under examination varies regionally throughout the German-speaking region of Europe. The codified Standard German variants of each pair are [ɛː] (predominant in the South) and [ɪç] (predominant in the North) (Dudenredaktion, 2015; Kleiner, 2011).[5] However, Kiesewalter (2019) has shown that the respective non-standard forms are perceived as subjectively corresponding to the standard (for [eː]; predominant in the North and Eastern Austria) or only slightly dialectal (for [ɪk]; predominant in the South) by native listeners of German. Therefore, we do not expect dialectality to influence accommodating behavior for these features.

While it is possible for a speaker to use both forms interchangeably, we expected the participants of the present study to have a preference for one of the two forms. The preference was determined for each participant as the majority variant produced for the 12 items per allophonic variation.[6] It was stored in the system and retrieved in task 4 to test for accommodation to the respective non-preferred variant.

2.1.2. Task 2 – question intonation, baseline

The participants formulate five wh-questions in random order whose components are given as fragments, e.g., *wer – die erste Frau – im Weltall – sein* (who – the first woman – in space – be) (see Fig. 2). Mirabella talks for the first time when she answers these questions.

This task familiarizes the participants with Mirabella's voice and reveals the intonation they usually apply when producing constituent questions. See Appendix C for the expected questions and the corresponding answers given by Mirabella.

[3] In one of twelve target items the vowel occurs word initially; in two target items the graphematic form is ⟨äh⟩. For simplification, we are referring to all of these with ⟨-ä-⟩.

[4] In many cases this ending constitutes a morphemic suffix.

[5] Often, the opposite is thought to be the case by speakers, since the written form of the word ending ⟨-ig⟩ hints towards [ɪk] being the standard and there is a tendency of long, stressed ⟨-ä-⟩ merging to [eː] across the German-speaking regions.

[6] In the event of a tie, the Standard German variant was set as the speaker preference.

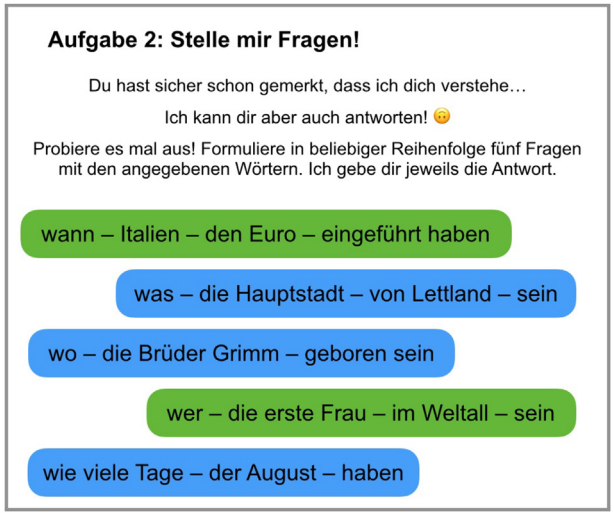


Fig. 2. Task 2 – Ask me questions! Participants formulate five wh-questions in random order from the given fragments. They are answered by Mirabella. This familiarizes them with Mirabella's voice and elicits baseline question intonation patterns. Green questions have already been asked. (For a color version of the figure, the reader is referred to the web version of this article).

Fig. 3. Task 3 – Where did the animals hide? Question-and-answer game testing accommodation to the intonation of constituent questions. Here: second round of the game; both players have asked each other seven questions so far; the animals that Mirabella has already asked for are marked by green frames, those that the participant has asked for by blue frames; it is Mirabella's turn. (For a color version of the figure, the reader is referred to the web version of this article).

Slight variations to the expected questions are accepted by the experimenter to show a certain flexibility. In the case of bigger deviations or disfluencies, Mirabella encourages the participants to try again (see Appendix A, 11,12,13,14). This behavior in combination with utterances such as *"Lass mich überlegen…"* (Let me think…) or *"Sehr gute Frage!"* (Great question!) interspersed in the dialog, aims to reinforce the impression of talking to a non–human yet social interlocutor.

The general unmarked expectation is for German wh-questions to be produced with falling intonation. Rising intonation is mainly applied in the case of echo questions, i.e., when the answer was not understood and the question is uttered again (cf. Grice & Baumann, 2002; Möbius, 1993; Wochner et al., 2015). We therefore expected to find mainly falling intonation contours for the questions asked in this task.

2.1.3. Task 3 – question intonation, test

In this task Mirabella and the participants take turns asking (Q) and answering (A) each other about ten animals hiding in ten houses (see Fig. 3), in the following form:

Q: *Wo hat sich* ⟨***the animal***⟩ *versteckt?*
Where did ⟨the animal⟩ hide?
A: ⟨***the animal***⟩ *hat sich in Haus Nummer* ⟨***number***⟩ *versteckt.*
⟨the animal⟩ hid in house number ⟨number⟩.

The order in which Mirabella and the user ask for the animals on the screen is free. The task includes two rounds of 20 turns, with Mirabella and the participants each asking and answering 10 questions per round.[7] The realization of questions on the part of the system differs between round one and round two with respect to pitch accent placement and intonation, giving room for accommodation. In round one (R1), Mirabella produces all questions with a nuclear pitch accent on the ⟨*animal*⟩ followed by a final F_0 fall, whereas in round two (R2), all questions are produced with a nuclear pitch accent on the interrogative pronoun *wo* (where) followed by a final high F_0 rise – here illustrated using the example "Where did the lion hide?":

R1: *Wo hat sich der* ***Löwe*** *versteckt?* ↘
R2: ***Wo*** *hat sich der Löwe versteckt?* ↗

The latter version constitutes the typical shape of an echo question asking for information that was already given, but not understood. Such echo questions are unlikely to occur naturally in the context of the question-and-answer exchange at hand, since the answers do not necessarily have to be understood by the participants: the correct pictures are always visually marked on the screen as well.

In the second round of the game, all animals stay paired with the same house numbers as before, however the arrangement of the houses on the screen differs from that of the first round. Therefore, it is unexpected, yet not pragmatically wrong, to ask for the location of the animals in the form of an echo question.

We expected to find falling intonation contours for the first round and a substantial increase of rising contours from the first to the second round. Additionally, we expected the nuclear pitch accent to be shifted from the ⟨*animal*⟩ in the first round to the interrogative pronoun *wo* in the second round.

2.1.4. Task 4 – allophonic variation, test

In this map task the participants describe the path from leaving a house until reaching a destination on the map while walking past different objects (see Fig. 4). To that end, they are using the prepositions given on the right side of the screen (see Appendix D for details). Additionally, the participants describe the object in question with the adjective given next to it at every step. This results in two-part utterances of the following type:

- *Ich gehe um die* ***Säge*** *herum. Die* ***Säge*** *ist schwer.*

[7] We did not include explicit filler material in this task, e.g., questions with different intonation contours, since we assume that accommodation requires a certain amount of repetition. The answers uttered between the questions serve as filler material for the questions themselves, in that they have a different intonation contour, thus providing a certain amount of variety and distraction.

Fig. 4. Task 4 – How do you reach the destination? Map task testing accommodation to allophonic variation. Here: The participant has made her way through the map up until the position marked by the yellow frame. She will ask Mirabella for the item behind the yellow box, use the preposition *um…herum* to say that she goes *around* the item, and use the given adjective *müde* to further describe the item as *tired*. (For a color version of the figure, the reader is referred to the web version of this article).

I walk around the saw. The saw is heavy.
► bold target contains the [ɛː] vs. [eː] contrast

- *Ich gehe an dem Pferd vorbei. Das Pferd ist* ***mutig****.*
  I walk past the horse. The horse is brave.
  ► bold target contains the [ɪç] vs. [ɪk] contrast

Some of the objects (**O**) and adjectives (**A**) are hidden behind boxes. The participants ask Mirabella about these items: "*Mirabella, was ist hinter der* ⟨***color***⟩ *Box?*" (Mirabella, what is behind the ⟨color⟩ box?)

The information about the participants' preference with respect to the [ɪç] vs. [ɪk] and [ɛː] vs. [eː] contrasts is automatically retrieved from the results of task 1 before the map task. Mirabella then uses the non-preferred variants when providing the requested information:

O: *Hinter der* ⟨***color***⟩ *Box ist* ⟨***the object***⟩. Behind the ⟨color⟩ box is ⟨the object⟩.
A: *Das Wort hinter der* ⟨***color***⟩ *Box ist* ⟨***adjective***⟩. The word behind the ⟨color⟩ box is ⟨adjective⟩.

Given this information, the participants can formulate the required two-part utterance. Subsequently, the hidden item is revealed.

The task consists of four maps with nine object-adjective pairs each and contains a total of 12 occurrences per allophonic contrast. Each map contains:

- three pairs including an [ɪç] vs. [ɪk] target
  e.g., *Honi**g*** (honey) – *süß* (sweet); *Baum* (tree) – *schatti**g*** (shady)
- three pairs including an [ɛː] vs. [eː] target
  e.g., *M**ä**dchen* (girl) – *schlau* (smart); *Bus* (bus) – *verspä**t**et* (delayed)
- three filler pairs not including a target[8]
  e.g., *Haus* (house) – *leer* (empty); *Autos* (cars) – *laut* (loud)

If the target item is an object, it occurs twice in the two-part utterance (e.g., *Honig*, *Mädchen*; see *Säge* in the example above); if the target item is an adjective, it occurs only once, in the second part of the utterance (e.g., *schattig*, *verspätet*; see *mutig* in the example above).

We expected to find a substantial increase of the non-preferred variant for the [ɪç] vs. [ɪk] contrast and a substantial shift in the F1-F2 space in the direction of the non-preferred variant for the [ɛː] vs. [eː] contrast during the map task as compared to the baseline task.

2.2. Text material

The text material used in the experiment pertains to two different categories. The first category contains structural utterances, which are either used to explain the tasks or to guide the conversation. While the explaining utterances are presented at the beginning of a new task and follow a chronological order that is the same for all participants, the guiding utterances are available to the experimenter at any time during the experiment and may be used to react to the participants' behavior if needed. Examples of such guiding utterances can be found in Appendix A.

The second category contains utterances which are part of the actual tasks testing for phonetic accommodation, either as target or filler material. More details about these utterances were given above, together with the explanations of the individual tasks in section 2.1.

Since the experiment is designed as an application for learning the German language, the text material used in the experiment was chosen to be accessible to advanced learners of German. This constrains the selection of possible target items substantially.

2.3. Stimuli

The first set of Mirabella's utterances was pre-recorded by a female native speaker of German (aged 26 years). The recordings were carried out with a sampling rate of 48kHz using a stationary cardioid microphone in a sound-attenuated booth. The speaker was instructed to speak in a friendly tone, basing her performance on experience with the usual tone of commercial language assistance systems. She produced the target stimuli in their different forms. The best versions in terms of target feature clarity were selected for use in the experiment.

The second set of utterances consists of synthesized speech. As the idea of the present study is to extend the analysis presented in (Gessinger, Raveh, Le Maguer, Möbius, & Steiner, 2017), we rely on the same paradigm with an updated process. This updated process uses three main toolkits: MaryTTS (Le Maguer, Steiner, Tombini, Deb, & Basu, 2018) as the front-end, HMM-based speech synthesis system (HTS) (Zen and Toda, 2005) to achieve the modeling, and the vocoder WORLD (Morise, Yokomori, & Ozawa, 2016) to render the signal from the acoustic parameters generated by HTS.

The HTS models were trained using the BITS corpus (Ellbogen, Schiel, & Steffen, 2004). We used the samples recorded by speaker *spk1*, which in total correspond to about 3h of speech sampled at 48kHz. The provided alignment was discarded, as our voice building pipeline (Steiner & Le Maguer, 2018) already includes an automatic alignment step.

[8] The [ɪç]/[ɪk] items additionally serve as fillers for the [ɛː]/[eː] items and vice versa.

In order to achieve German based synthesis, we defined a feature set derived from the one proposed for English (Tokuda, Zen, & Black, 2002). The major modification is the adaptation of the phonetic part for German. This adaptation corresponds to the extension of the phonetic alphabet and the addition of corresponding questions in the question file.

Within the synthesis pipeline we imposed three main parameters. On the one hand, we modified the front-end decision by inducing the allophonic contrasts [ɪç]/[ɪk] and [ɛː]/[eː]. This enabled HTS to produce the different variants in the map task stimuli. On the other hand, we extracted the segment durations and F_0 contours from the natural stimuli and applied these values in the synthesis process – the durations at the phone level and the fundamental frequency at the frame level. By imposing these parameters, it was possible to generate the variations of prosodic structure in the synthetic utterances of the question-and-answer game.

Imposing the duration at the phone level is straightforward as this option is directly implemented in HTS. To impose F_0, we had the choice between two main solutions: using the voicing prediction from the system or creating a new voicing prediction using the generated spectral information in combination with a simple neural network. After informal subjective evaluation, we concluded that using the voicing information predicted by HTS leads to a more consistent quality and is less likely to introduce artifacts. Applying this voicing mask when imposing the fundamental frequency avoided mismatches between F_0 and the harmonic structure of the spectrum.

Both versions of Mirabella thus use the natural source signal, but they differ with respect to the filter applied to the latter: the human vocal tract for the natural stimuli and HTS for the synthetic stimuli.

HTS produces speech with a degraded voice quality, which is often described as buzzy or muffled (Zen, Tokuda, & Black, 2009). We can therefore assume that the synthetic version of Mirabella is clearly perceived as non–human by the participants, whereas in the case of Mirabella's natural version, the impression of talking to a computer is mainly caused by the interaction itself. The process of imposing the natural segment durations and F_0 contours during synthesis, however, ensured that the synthetic version of Mirabella was still as similar as possible to the natural version in its perceived personality, insofar as the latter is conveyed through prosody (e.g., Smith, Brown, Strong, & Rencher, 1975; Apple, Streeter, & Krauss, 1979; Nass & Lee, 2001; Trouvain, Schmidt, Schröder, & Schmitz, 2006). This is relevant since the perceived personality of the interlocutor can influence the accommodating behavior towards them (e.g., Yu et al., 2013; Lewandowski & Jilka, 2019).

2.4. Participants

The participants were recruited on the Saarland University campus and were paid for taking part in the experiment. All 42 participants were native speakers of German and four spoke more than one native language (English ($n = 2$), Polish ($n = 1$), Greek ($n = 1$)). All had learned at least one foreign language, the majority two or more. The most frequent foreign languages were English ($n = 42$), French ($n = 31$), and Spanish ($n = 16$). Thirty-nine participants were students and three had non-academic jobs. The participants came from eleven German states with 61% from central regions, 22% from southern regions, and 17% from northern regions.

In a questionnaire completed after the experiment, which asked the participants to assess their general communicative behavior, 98% answered affirmatively to the question whether they change the way they speak depending on their respective interlocutor; 69% believed they would converge to an interlocutor of the same dialectal background; 26% claimed they would do the same with an interlocutor of a different dialectal background; only 17% 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.

Each participant was presented with only one of the two stimulus types – natural or HMM. This resulted in two experimental groups: the natural group with 20 participants (16 female, 4 male; mean age 25.8 years; age range 18 to 55 years) and the HMM group with 22 participants (15 female, 7 male; mean age 23.7 years; 18 to 32 years).

3. Analysis and results

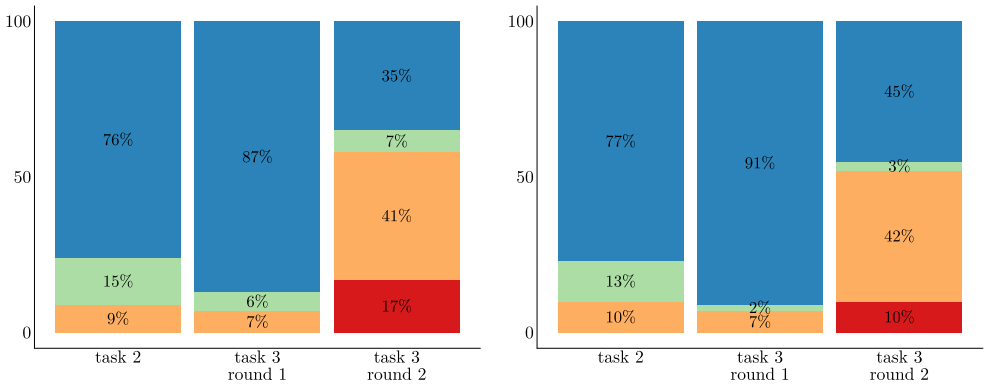
3.1. Rating of Mirabella

After the experiment, the participants rated Mirabella on 5-point scales with regard to her likability (*unpleasant* to *very likable*), competence (*incompetent* to *very competent*), intelligibility (*bad* to *very good*), and response time (*too slow* to *too fast*). 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 ratings of the two versions of Mirabella differed most for intelligibility, with the synthetic version (mean = 3.9, *SD*= 0.8) being less intelligible than the natural version (mean = 5, *SD*= 0.2). In addition, the synthetic version of Mirabella was judged to be less likable (synthetic: mean = 3.8, *SD*= 1; natural: mean = 4.5, *SD*= 0.6), but only slightly less competent (synthetic: mean = 4, *SD*= 0.9; natural: mean = 4.3, *SD*= 0.4). Mirabella's response time, i.e., the response time of the experimenter, was considered equally appropriate in both cases (synthetic: mean = 2.9, *SD*= 0.9; natural: mean = 2.9, *SD*= 0.6).

3.2. Modeling

The dependent variables are analyzed using linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) formulated with the lme4 package (1.1–21) (Bates, Mächler, Bolker, & Walker, 2015) and evaluated with the lmerTest package (3.1–0) (Kuznetsova, Brockhoff, & Christensen, 2017) in RStudio (1.1.463) (RStudio Team, 2016) with R (3.5.2) (R Core Team, 2018).



(a) Natural group. (b) Synthetic group.

Fig. 5. Percentages of questions realized with falling, falling-rising, rising(a), or rising(w) intonation contour during the baseline production (task 2) and the two rounds of the question-and-answer game (task 3). (For a color version of the figure, the reader is referred to the web version of this article).

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 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 model fit improves significantly and the AIC value decreases by at least two points as compared to the model without the factor in question. Factors kept in the model are being considered significant predictors of the respective dependent variable at $\alpha = 0.05$.

The data and the analysis scripts are available as supplementary materials to this article.

3.3. Question intonation

The intonation contours of the 1 094 questions uttered in tasks 2 and 3 (natural: $n = 526$, synthetic: $n = 568$)[9] were perceptually classified by two trained phoneticians, taking the position of the nuclear pitch accent into account. Three contour types were found in the data: *falling*, *falling-rising*, and *rising* (cf. Grice & Baumann, 2002). The latter occurs in two variants: first, as *rising(a)* contours with a nuclear pitch accent on the respective *animal* in task 3 or an equivalent word in focus in task 2, and second, as *rising(w)* contours with a nuclear pitch accent on the interrogative pronoun *wo*. Fig. 5 shows the results of the evaluation for the two experimental groups.

In task 2, where the participants formulate wh-questions from given fragments, the *falling* contours are predominant in both groups (natural: 76%, synthetic: 77%), but *falling-rising* (natural: 15%, synthetic: 13%) and *rising(a)* (natural: 9%, synthetic: 10%) contours are produced as well.

In the first round of task 3, where Mirabella produces exclusively *falling* contours, the predominance of *falling* contours on the part of the participants becomes more pronounced in both groups (natural: 87%, synthetic: 91%), yet *falling-rising* (natural: 6%, synthetic: 2%) and *rising(a)* (natural and synthetic: 7%) contours still occur.

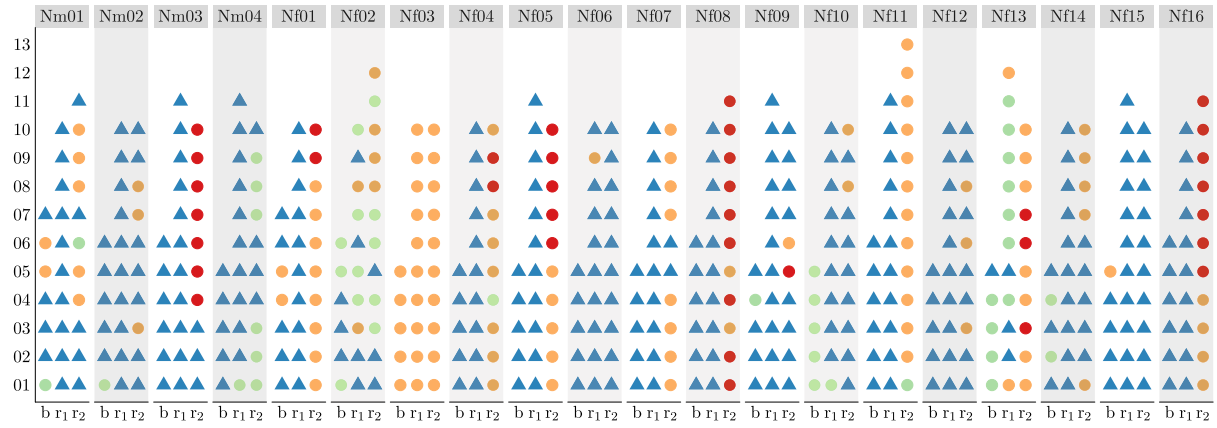
In the second round of task 3, where Mirabella produces exclusively *rising(w)* contours, the amount of *rising(a)* contours increases in both groups (natural: 41%, synthetic: 42%) and *rising(w)* contours emerge in both groups as well (natural: 17%, synthetic: 10%). While the amount of *falling-rising* contours stays about the same in both groups (natural: 7%, synthetic: 3%), the number of *falling* contours is considerably smaller in the second round of task 3 (natural: 35%, synthetic: 45%).

The increase of rising contours (this includes *falling-rising*, *rising(a)*, and *rising(w)* contours) from round 1 to round 2 of task 3 per experimental group was evaluated by fitting GLMMs with the binary response *falling/rising* as dependent variable and testing the factors TASK (round1/round2) and SPEAKER SEX (female/male) following the method described in section 3.2.

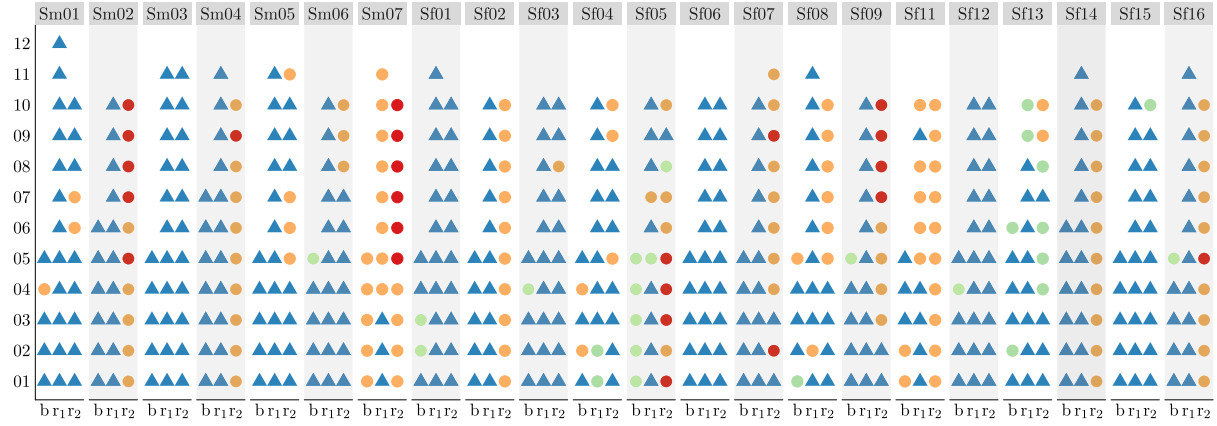
The model of the natural data set did not converge when random intercepts for ITEM, i.e., the different animals, were included, therefore the models for both experimental groups were fitted only including random intercepts for USER. The factor TASK is a significant predictor of the dependent variable in the natural group (Estimate $= -4.87$, SE $= 1.24, z = -3.94, p < 0.001$) and the synthetic group (Estimate $= -2.73$, SE $= 0.8, z = -3.44, p < 0.001$) indicating an increase of rising contours in round 2 of task 3. The models include random slopes for TASK by USER to account for the individual reactions of the participants. The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

Fig. 6 shows the individual question realizations in chronological order by each speaker of the natural and synthetic group. Note that some speakers never deviate from their preferred question intonation, e.g., speaker *Sm03* always produces the expected *falling* pattern, while speaker *Nf03* only utters *rising(a)* questions. In contrast, *Nm03*, *Nf05*, *Sm02*,

[9] Theoretically expected number of data points: (5 base questions + 2 × 10 animal questions) × number of participants. Small deviations due to repetitions.



(a) Natural group.



(b) Synthetic group.

Fig. 6. Individual question realizations with falling ▲, falling-rising ●, rising(a) ●, or rising(w) ● intonation contour in their order of occurrence during the baseline production (b), as well as during round 1 (r_1) and round 2 (r_2) of the question-and-answer game. (For a color version of the figure, the reader is referred to the web version of this article).

and *Sf09*, are examples of speakers who have a clear preference to produce the *falling* pattern, but ultimately converge to the *rising(w)* pattern produced by Mirabella, either directly or via instances of *rising(a)*.

To evaluate the accommodating behavior on the individual level we classified all participants according to the following thresholds, comparing the number of *rising(a)* or *rising(w)* occurrences in round 2 to round 1:

- increase of $\geqslant 5 \longrightarrow$ substantial convergence
- increase of $\geqslant 2 \longrightarrow$ moderate convergence
- in-/decrease of $1 \longrightarrow$ maintenance
- decrease of $\geqslant 2 \longrightarrow$ moderate divergence
- decrease of $\geqslant 5 \longrightarrow$ substantial divergence

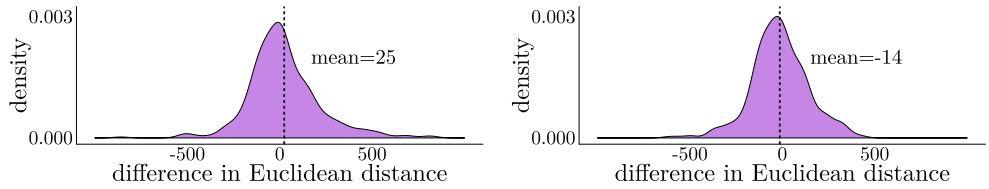
According to these criteria, 21 participants show substantial convergence (natural: 11, synthetic: 10), moderate convergence is found in 11 participants (natural: 5, synthetic: 6), and 10 participants do not change their question intonation (natural: 4, synthetic: 6). Divergence on the individual level was not found.

3.4. Allophones [ɛː]/[eː]

As auditorily determined by the experimenter during the baseline task, 25 of the 42 speakers participating in the present experiment had a preference for [ɛː] (20 female, 5 male) and 17 speakers had a preference for [eː] (11 female, 6 male). In order to validate the online annotations, all baseline [ɛː]/[eː] targets were annotated again by the original annotator, i.e., the experimenter, and an additional phonetically trained annotator without time pressure and with the option to visualize the spectrogram. The inter-rater agreement between these two offline annotations and the intra-rater agreement between the online and offline annotations of the experimenter were both found to be almost perfect (Cohen's kappa = 0.90). Although the auditory classification of vowel quality in a binary way poses a certain challenge in the experimental procedure, because ambiguous forms can be difficult to assign to a category, we conclude from this validation that the participants' preference with respect to [ɛː]/[eː] was determined correctly.

For all 1 357 realizations of long, stressed ⟨-ä-⟩ uttered by the participants in tasks 1 (natural: $n = 247$, synthetic: $n = 264$) and task 4 (natural: $n = 391$, synthetic: $n = 431$) as well as by Mirabella ($n = 12$ per natural and synthetic version), the first and second formants were measured at the temporal midpoint of the vowel using Praat's Burg algorithm (Boersma & Weenink, 2017).

In a second step, the Euclidean distance (*dist*) in the F1-F2 space between each participant realization (*U*) and the



(a) Natural group. (b) Synthetic group.

Fig. 7. Difference in Euclidean distance in the F1-F2 space (in Hz) between participant realizations of ⟨-ä-⟩ and the respective realizations by Mirabella in the baseline compared to the map task. Positive values indicate convergence, negative values divergence. The distribution means are shown by the dashed lines. They do not differ significantly from zero for either of the groups.

respective realization by Mirabella (*M*) was calculated for the baseline task (Eq. 1) and the map task (Eq. 2), e.g., for ⟨-ä-⟩ in *Käse*:

- participant's base production vs. Mirabella's production
- participant's map production vs. Mirabella's production

Finally, the difference in Euclidean distance (*dDist*) between the baseline task and the map task was calculated (Eq. 3), resulting in data sets of 403 values for the natural and 431 values for the synthetic group.[10]

$$dist(b) = \sqrt{(U_{baseF1} - M_{F1})^2 + (U_{baseF2} - M_{F2})^2} \quad (1)$$

$$dist(m) = \sqrt{(U_{mapF1} - M_{F1})^2 + (U_{mapF2} - M_{F2})^2} \quad (2)$$

$$dDist = dist(b) - dist(m) \quad (3)$$

Difference in Euclidean distance has the following potential outcomes:

- $dDist > 0$, if the participants shift their productions in the direction of Mirabella (convergence);
- $dDist = 0$, if the participants do not shift their productions in the F1-F2 space (maintenance);
- $dDist < 0$, if the participants shift their productions away from Mirabella (divergence).

The difference in Euclidean distance measure contains the information about the experimental task, since it is calculated as a comparison of the baseline and map task. It is therefore the model intercept that provides insight about accommodating behavior. The intercept is considered to significantly differ from 0 at $\alpha = 0.05$.

Fig. 7 shows the distributions of *dDist* for the two experimental groups. The distribution of the natural group has a mean of 25 which is positive and therefore suggests convergence; the distribution of the synthetic group has a mean of -14 which is negative and therefore suggests divergence.

However, fitting LMMs with *dDist* as dependent variable and testing the factors SPEAKER SEX (female/male) and PREFERENCE ([ɛː]/[eː]) following the method described in section 3.2, revealed that the means do not differ significantly from zero for the natural group (Estimate $= 26.32$, SE $= 24.54$, df $= 20.36, t = 1.07, p = 0.3$), as well as for the synthetic group (Estimate $= -19.31$, SE $= 18.4$, df $= 24.46, t = -1.05$, $p = 0.3$). These models include random intercepts for USER and ITEM, i.e., the target words. The factor PREFERENCE was a significant predictor in the model of the synthetic group, indicating that the participants with a baseline preference for [eː] have a stronger tendency to diverge than the participants preferring [ɛː], whose group intercept is slightly above zero (Estimate $= 37.06$, SE $= 14.38$, df $= 19.55, t = 2.6, p < 0.05$). The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

Fig. 8 shows the individual productions of ⟨-ä-⟩ by each speaker in the natural and synthetic groups relative to the vowels they heard from Mirabella. To evaluate the accommodating behavior on an individual level, two complementary tests were carried out per participant.[11] First, a kernel density based global two-sample comparison test for 2-dimensional data was performed to determine whether the set of baseline vowels differed significantly from the set of map task vowels ($\alpha = 0.05$). Second, a two-sided one-sample Wilcoxon signed-rank test evaluated whether the individual *dDist* distribution differed significantly from zero ($\alpha = 0.05$). If both tests reach significance, we consider the individual participant to accommodate to Mirabella, since their map task productions are substantially farther from their original baseline distribution while being substantially closer to (convergence) or farther from (divergence) Mirabella's vowels. This approach suggests three cases of convergence with respect to vowel quality (*Nm02*, *Nf04*, and *Nf15*) and five cases of divergence (*Nm01*, *Nm03*, *Sm01*, *Sm05*, and *Sf12*).

Since these tests were carried out for all 42 participants separately, we have to consider adjusting the p-values to control the false discovery rate (Benjamini & Hochberg, 1995). For all participants mentioned above, the adjusted p-values of the Wilcoxon signed-rank test remain below 0.05. However, only for *Nf15*, *Sm01*, *Sm05*, and *Sf12*, the same is true for the kernel density based comparison, as well. While keeping this limitation in mind, we still consider all eight speakers to show accommodating behavior with regard to [ɛː]/[eː].

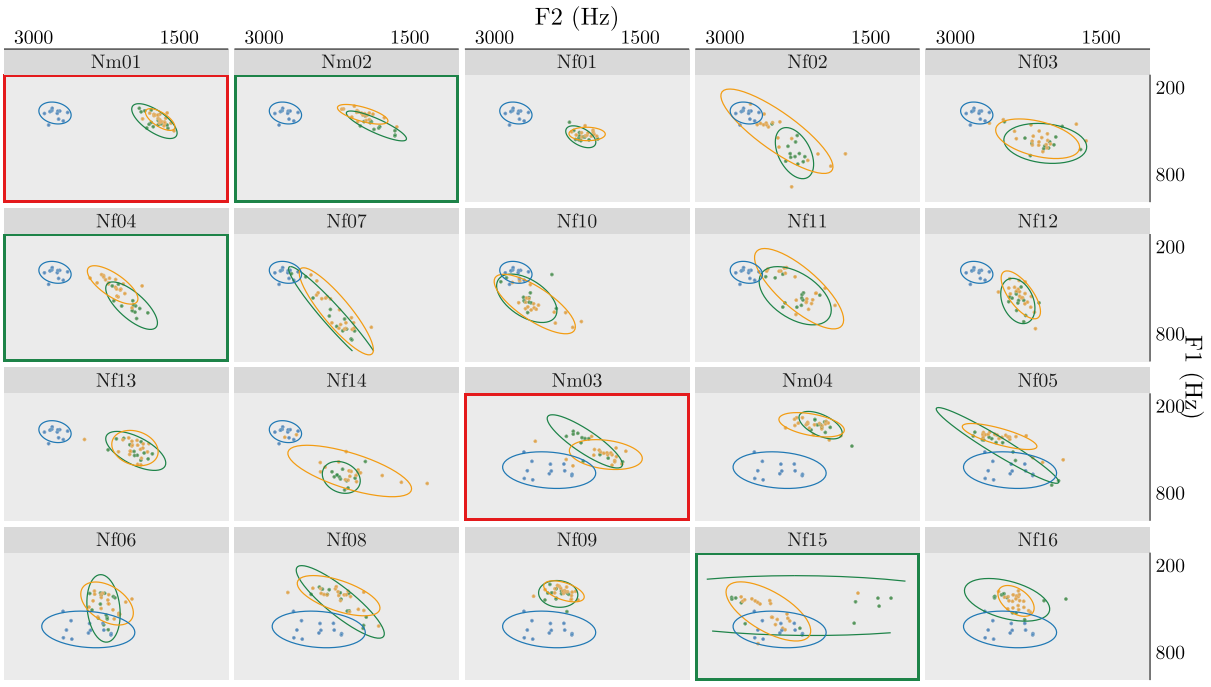
3.5. Allophones [ɪç]/[ɪk]

The 1 088 realizations of the word ending ⟨-ig⟩ uttered in tasks 1 and 4 (natural: $n = 518$, synthetic: $n = 570$)[12] were auditorily and visually classified as belonging to the fricative or plosive category by the first author of the study and an additional phonetically trained annotator. The two resulting annotations of

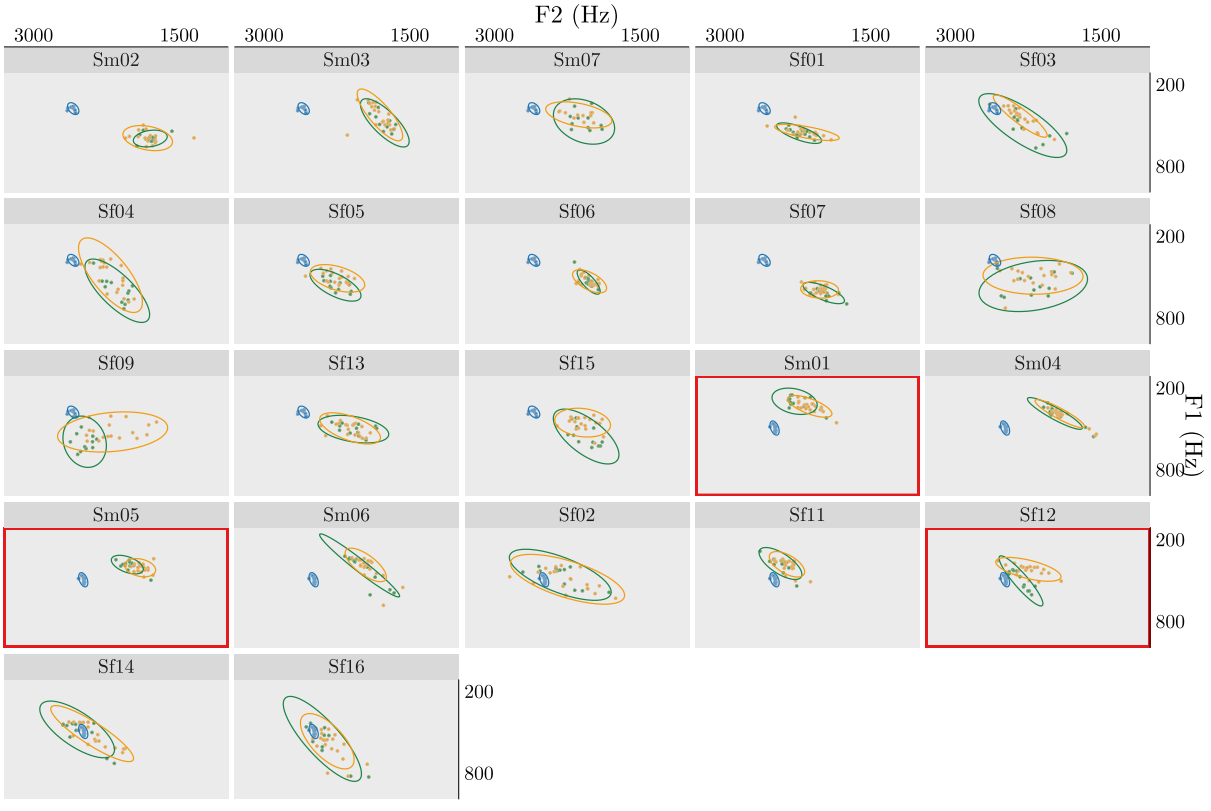
[10] Theoretically expected number of data points: 20 map items [i.e., 2 × 8 nouns + 4 adj.] compared with their base counterpart × number of participants. Small deviations due to missing values and repetitions.

[11] The analysis differs from the one in Gessinger et al. (2019a), where only the second test was carried out.

[12] Theoretically expected number of data points: (12 base items + 14 map items [i.e., 2 × 2 nouns + 10 adj.]) × number of participants. Small deviations due to missing values.

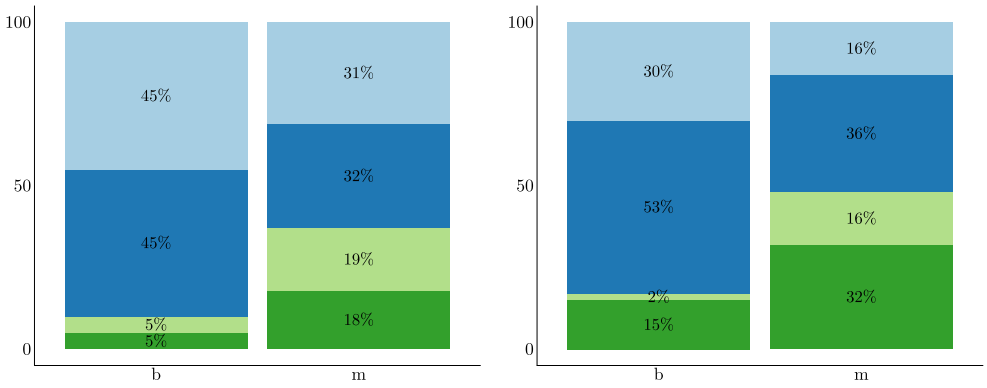


(a) Natural group.



(b) Synthetic group.

Fig. 8. Individual participant realizations of ⟨-ä-⟩ in the F1-F2 space (in Hz) from the baseline task and the map task, relative to the vowels the participants heard from Mirabella. The ellipses indicate the 95% confidence interval. Framed participants were found to converge to or diverge from Mirabella. (For a color version of the figure, the reader is referred to the web version of this article).



(a) Natural group. (b) Synthetic group.

Fig. 9. Percentages of the word ending ⟨-ig⟩ realized as the *same* variant (▲) or a *different* variant (■) as Mirabella in the baseline production (b) and the map task (m) split by participants whose baseline preference is [ɪç] (light tones) or [ɪk] (dark tones). (For a color version of the figure, the reader is referred to the web version of this article).

the baseline items were compared with the online annotation performed by the first author of the study during the experiment on a purely auditory basis and under time pressure. The two offline annotations did not differ from each other and the online and offline annotations of the first author differed in a single instance only. We conclude from this validation that the participants' preference with respect to [ɪç]/[ɪk] was determined correctly.

Since speakers are not always consistent in using only one variant during the baseline task, preference reflects the majority variant produced during task 1. Of the 42 speakers participating in the present experiment, 17 had a preference for the [ɪç] variant (13 female, 4 male) and 25 for the [ɪk] variant (18 female, 7 male). Individual realizations were further classified as being the *same* as or a *different* variant than the one produced by Mirabella.

Fig. 9 shows the results of the [ɪç] vs. [ɪk] evaluation for both experimental groups. The clear majority of all baseline instances is produced with a *different* variant of the target contrast than the one the participants hear from Mirabella in the map task (natural: 90%, synthetic: 83%). This is expected, since the variant used by Mirabella is selected to be the opposite of each participant's preference. In the remaining cases (natural: 10%, synthetic: 17%), the participants uttered the non-preferred variant in the baseline task, hence the *same* variant as Mirabella.

While the participants in the natural group are split equally between those preferring [ɪç] and those preferring [ɪk] and in each of the subgroups the *different* variant of the target contrast is produced in 90% of the baseline instances, the synthetic group contains more participants preferring [ɪk] (68%), and within this subgroup only 78% of the baseline instances are of the *different* type (compared to 94% for the [ɪç]-preference subgroup). This means that there is more variation in the baseline productions of the synthetic [ɪk]-preference subgroup than in the three other subgroups.

In the map task, the amount of non-preferred variants uttered by the participants increases by 27% to a total of 37% in the natural group and by 31% to a total of 48% in the synthetic group. In the natural data, the occurrences of *same* variants quadruple for both subgroups ([ɪç]: 19%, [ɪk]: 18%). In the synthetic data, the [ɪç] and [ɪk] subgroups contribute to the increase to different proportions: There are eightfold as many *same* variants in the [ɪç] subgroup (16%) while the occurrences only double (32%) in the [ɪk] subgroup.

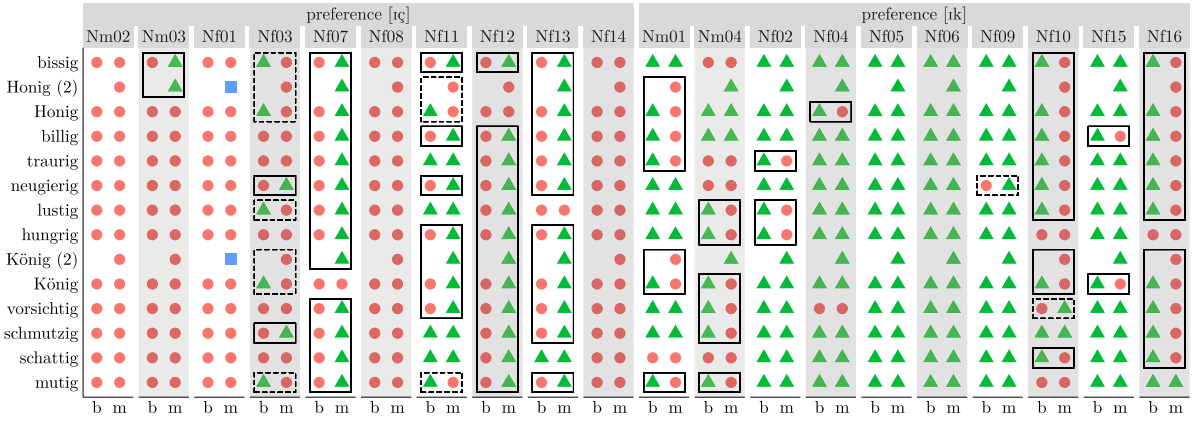
The increase of non-preferred variants per experimental group was evaluated by fitting GLMMs with the binary response *different/same* as dependent variable and testing the factors TASK (base/map), SPEAKER SEX (female/male), and PREFERENCE ([ɪk]/[ɪç]) following the method described in section 3.2.

The models did not converge when random intercepts for ITEM, i.e., the target words, were included, therefore the models were fitted only including random intercepts for USER. The factor TASK is a significant predictor of the dependent variable in the natural group (Estimate $= -0.91$, SE $= 0.44, z = -2.05, p < 0.05$) and the synthetic group (Estimate $= -0.73$, SE $= 0.23, z = -3.23, p < 0.01$) indicating an increase of *same* variants of the target contrast in the map task. The models include random slopes for TASK by USER to account for the individual reactions of the participants. The factor PREFERENCE is a significant predictor in the model of the synthetic group establishing the above made observation that the group of participants preferring [ɪk] is larger (Estimate $= 0.69$, SE $= 0.28, z = 2.47, p < 0.05$). However, there is no significant interaction of TASK and PREFERENCE. The factor SPEAKER SEX did not improve the fit of the models and was therefore not included.

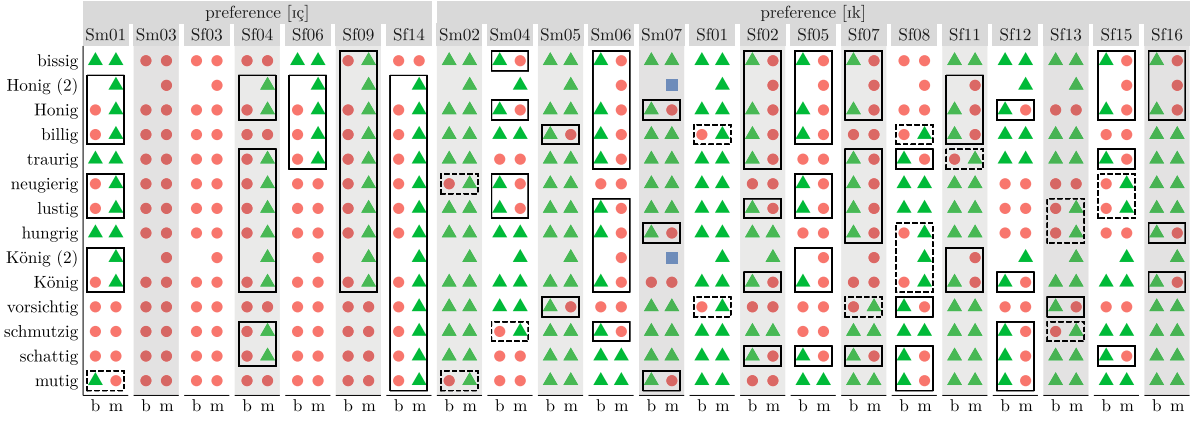
Fig. 10 shows the individual realizations of the word ending ⟨-ig⟩ in chronological order by each speaker of the natural and synthetic group. Note that some speakers never deviate from their preferred allophonic variant, e.g., speakers *Nm02* and *Sf03* always produce the fricative variant, while speakers *Nf05* and *Nf06* only produce the plosive variant. In contrast, *Nf07* and *Sf14* are examples of speakers who have a clear preference for one variant in the baseline task, but converge almost entirely to Mirabella during the map task.

To evaluate the accommodating behavior on the individual level we classified all participants according to the following thresholds, comparing the number of *same* instances in task 4 to task 1:

- increase of $\geqslant 7 \longrightarrow$ substantial convergence
- increase of $\geqslant 2 \longrightarrow$ moderate convergence



(a) Natural group.



(b) Synthetic group.

Fig. 10. Individual results for the realization of the word ending ⟨-ig⟩ as [ɪç] ● or [ɪk] ▲ in the baseline production (b) and the map task (m). Target words are given in the order of occurrence in the map task, starting with *mutig*. Solid boxes show cases of convergence, dashed boxes cases of divergence. The ■ indicates missing values. Participants are grouped by their baseline preference for [ɪç] or [ɪk].

- in-/decrease of 1 ⟶ maintenance
- decrease of ⩾2 ⟶ moderate divergence
- decrease of ⩾7 ⟶ substantial divergence

According to these criteria, 13 participants show substantial convergence (natural: 6, synthetic: 7), moderate convergence is found in 14 participants (natural: 5, synthetic: 9), 11 participants do not increase nor decrease the number of *same* instances (natural: 8, synthetic: 3), and 4 participants moderately diverge from Mirabella (natural: 1, synthetic: 3). Substantial divergence on the individual level was not found.

3.6. Personality scores

To explore the influence of different personality traits on the accommodation occurring in the present study, we collected personality scores of all participants using the German version of the NEO-FFI (Borkenau & Ostendorf, 2007). This self-description questionnaire measures the "Big Five" personality traits, i.e., neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. It uses a total of 60 items (12 items per trait) and takes approximately 10 min to complete. The questionnaire was administered after the experiment.

Raw values were calculated for each personality trait and converted into standard T-values according to the guidelines provided by NEO-FFI. These standard values take the sex and age of the participants into account.

For each personality trait, we selected the 35% of all participants (natural and synthetic group) with the lowest values and the 35% with the highest values. This resulted in balanced subsets of 29 to 30 participants. For each of the five subsets and three phonetic features we fitted the statistical models described above again, always including random intercepts for USER and ITEM.

We tested the factors TASK (where applicable), PREFERENCE (where applicable), and PERSONALITY TRAIT (high/low).

For one subset a significant effect of PERSONALITY TRAIT emerged:

In the group of participants with very high or very low values for neuroticism, question intonation was influenced by TASK, i.e., round 1 or round 2 of the question-and-answer game, (Estimate $= -4.76$, SE $= 1.14, z = -4.02, p < 0.001$) and there was a significant interaction of TASK and PERSONALITY TRAIT (Estimate $= -1.27$, SE $= 0.57, z = -2.22, p < 0.05$), indicating that participants who scored high values for neuroticism were more likely to produce questions with rising intonation

in round 2 of the game and therefore more likely to converge to Mirabella. The model includes random slopes for TASK by USER.

The reported p-values are not adjusted for the fact that the general hypothesis of whether personality traits influence accommodation with respect to a particular phonetic feature was tested for five subsets of the same data set, which increases the probability of a false positive result. If we adjust the p-values to control the false discovery rate (Benjamini & Hochberg, 1995), TASK is still a significant predictor of the intonation contour ($p < 0.001$). However, with $p = 0.1$ only a trend remains of the interaction with PERSONALITY TRAIT.

4. Discussion and conclusion

We conducted a WOz experiment with 42 native speakers of German to investigate phonetic accommodation by human interlocutors in an HCI context. The participants of the experiment solved four tasks in interaction with the virtual language learning tutor *Mirabella*, who was created for this purpose. The participants were confronted with Mirabella using either natural or synthetic speech. The latter was generated using MaryTTS, HTS, and WORLD (see section 2.3). The prosodic parameters segment duration and fundamental frequency were extracted from the natural stimuli and imposed on the synthetic ones. Due to this combined process, Mirabella's synthetic voice was clearly identifiable as non-natural while the stimuli still exhibited a natural prosody.

4.1. Reception of Mirabella

After the experiment, the participants rated Mirabella with regard to her likability, competence, and intelligibility. The natural Mirabella version was rated as being more intelligible, more likable, and somewhat more competent. However, both versions of Mirabella achieved high scores on all three 5-point scales with mean values well above 3 in each case. Mirabella's response time, i.e., the response time of the experimenter, was evaluated as well. It was considered equally appropriate in both experimental groups. This is plausible, since the experimenter was the same for both versions of Mirabella. Overall, the ratings of the synthetic Mirabella version showed more variability in all scales, which means that the participants were less in agreement in her case. Future work could investigate the influence of these evaluations on the accommodating behavior in detail.

As part of the questionnaire administered after the experiment, the participants could also express their thoughts and assumptions about the experiment. None of the participants raised any doubt that Mirabella functioned fully automatically, neither in the questionnaire nor through informal comments. On the contrary, they referred to their experience in a way that suggests they believed that they were interacting with a computer, which is a key component of HCI (Branigan et al., 2010). A frequently expressed assumption about the purpose of the study was to evaluate the dialog system in terms of how well it understands different participants and how quickly it responds to speech input. The interaction was perceived in many cases as a training for Mirabella with the presumed goal of improving HCI. One participant described the system as being child-friendly and suggested that it could be used in schools.

4.2. Accommodation to Mirabella

We tested accommodation with respect to the intonation of constituent questions in a question-and-answer game, and the variation of the German allophone pairs [ɛː] vs. [eː] as a realization of the long vowel ⟨-ä-⟩ in stressed syllables, e.g., ***Käse*** (cheese), and [ɪç] vs. [ɪk] as a realization of the word ending ⟨-ig⟩, e.g., ***Honig*** (honey), in a map task.

Both the question-and-answer game and the map task are of a rather repetitive nature. However, they are structured to reflect a possible interaction of human speakers, they enabled an engaging, dynamic and meaningful exchange between the participants and Mirabella, and it is conceivable that they could occur in a real-life learning context, especially in CALL.

4.2.1. Question intonation

As expected for native speakers of German, all participants produced predominantly *falling* intonation contours when formulating constituent questions from given fragments.

When interacting with Mirabella in the first round of the question-and-answer game, where she produced her questions with a nuclear pitch accent on the ⟨*animal*⟩ followed by a final F_0 fall, this predominance was reinforced in both experimental groups. The small amount of *falling-rising* contours and *rising(a)* contours that occurred in these two tasks, could either be idiosyncratic behavior – speaker *Nf03*, for example, produced exclusively *rising(a)* contours – or an expression of insecurity or politeness – such feelings are likely to weaken in the course of the interaction, e.g., because Mirabella's behavior confirms that the task is being carried out correctly. Therefore, it is unlikely that an increase in rising contours at later points in the interaction is attributable to insecurity or politeness.

The crucial change happened in the second round of the question-and-answer game, where Mirabella produced all questions with a nuclear pitch accent on the interrogative pronoun *wo* (where) followed by a final high F_0 rise (*rising(w)*). This behavior lead to a significant increase of rising contours (this includes *falling-rising*, *rising(a)*, and *rising(w)* contours) on the part of the participants in both experimental groups. This increase can mainly be attributed to a change in intonation contour while keeping the nuclear pitch accent on the ⟨*animal*⟩. However, in a smaller number of cases, participants also shifted the nuclear pitch accent to the interrogative pronoun. This suggests that the participants were primarily receptive to the overall rising contour. It seems sensible to ask to what extent convergence can take place without giving the impression to mock the interlocutor. The question intonation in the present study may well be a case in which full convergence, i.e., a rising contour *with* a shifted pitch accent, seems to go one step too far for many participants. Since a rising contour *without* a shifted pitch accent results in a more acceptable form than a shifted pitch accent with a *falling* contour – no such cases occurred in our data –, we can observe this clear two-step convergence hierarchy.

Pragmatic context In the following we would like to return briefly to the influence of the pragmatic context on the task at hand. We have already mentioned that the echo questions that

Mirabella produces in round two do not contradict the context. But they are also not expected to occur, since the change in the pragmatic context is not very obvious. In round two, the animals are arranged differently on the screen than in round one and we can assume that this is consciously perceived by the participants. Furthermore, the animals are still paired with the same house number, which justifies an echo question, but in our opinion is probably not consciously noticed by the participants.

Could it still be accommodation to the changed pragmatic context instead of to Mirabella's speech output that we observe in our data? The majority of the participants adopted the rising intonation, but did not shift the pitch accent to the interrogative pronoun in round two. However, there is no pragmatic motivation for this, because only the shift of the nuclear pitch accent (in combination with rising intonation) changes the function of the question to suit the changed pragmatic context.

In conclusion, we do not believe that changing the pragmatic context alone would trigger the observed amount of questions with rising intonation, nor do these questions fit functionally to the changed pragmatic context. We therefore assume that the observed change in question intonation by the participants constitutes accommodation to Mirabella.

Personality scores In a separate analysis, including personality scores collected with the German version of the NEO-FFI (Borkenau & Ostendorf, 2007), neuroticism emerged as a significant predictor of accommodation to question intonation, with more neurotic participants converging more to Mirabella. When applying the Benjamini-Hochberg correction to account for multiple comparisons, this effect did not hold. However, since the finding is in line with Lewandowski and Jilka (2019), where more neurotic speakers show more convergence with respect to word-based amplitude envelope match, we would like to discuss a possible explanation for the occurrence of such an effect regarding question intonation. A high degree of neuroticism is synonymous with emotional instability. People with a high level of neuroticism are more likely to state that they are easily out of balance, more insecure and nervous, and less able to control their needs (Borkenau & Ostendorf, 2007). Lewandowski and Jilka (2019) relate the degree of neuroticism to the need of social approval and suggest that under the CAT perspective (Giles, 1973; Giles et al., 1991; Shepard et al., 2001), this might imply that neurotic people have a tendency to converge in an attempt to avoid distress. However, the degree of neuroticism was not predictive of the other features tested in the present study. A possible difference between the question intonation and the allophonic contrasts is that deviating from the expected way of formulating questions might have more potential to cause communicative distress than using another allophonic variant.

Note that we conducted an analysis of isolated personality traits, while traits may also interact with each other in influencing accommodating behavior.

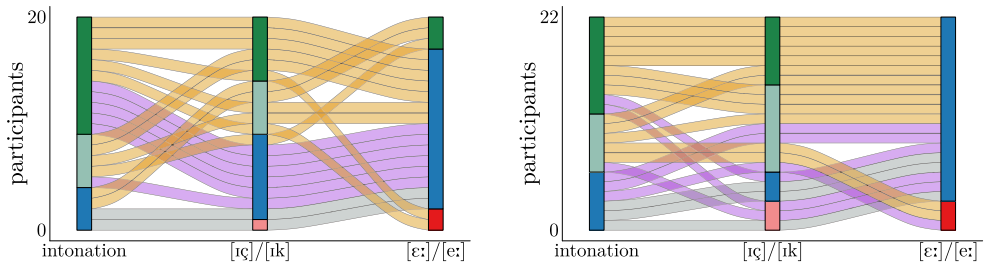
4.2.2. Allophonic contrasts

The allophonic contrasts were tested in a map task. The missing information on the map, i.e., the target words, were provided by Mirabella in full sentence contexts and had to be included in a two-part utterance by the participants. To construct an utterance, the participants had to select a suitable preposition and formulate grammatically correct sentences. This seemed to be difficult at times – although our participants were native speakers of German –, but always resulted in acceptable utterances for the purpose of the current study. In any event, the participants' attention had to be divided between different domains and we assume that pronunciation did not stand out as an obvious target.

Word ending ⟨**-ig**⟩ With respect to the [ɪç]/[ɪk] contrast, we found a significant convergence effect during the map task for both experimental groups. This effect did not depend on the baseline preference of the speaker. Although [ɪç] is codified Standard German and [ɪk] a Southern German variant, which might imply that the former is more prestigious and therefore able to trigger more convergence, an effect of baseline preference was not expected, since Kiesewalter (2019) showed that [ɪk] is perceived as being close to the standard by native listeners of German. Further evidence for the ambiguous status of the [ɪç]/[ɪk] contrast comes from the participants of the present study: In the post-experiment questionnaire, almost 40% of the participants misjudged which variant of the contrast they predominantly produce themselves.[13] When asked for their opinion about the respective other variant, the vast majority judged it as acceptable. Only five participants had a negative opinion about the variant they did not produce themselves, e.g., "wrong" or "weird". For speakers *Nm02*, *Nf01*, and *Nf14* this was [ɪk] and they did indeed not produce a single instance of it. For speakers *Sf02* and *Nf09* the disliked variant was [ɪç]. *Nf09* did produce one [ɪç] in the baseline task and then diverged to only producing [ɪk] during the map task, while *Sf02* produced three instances of [ɪç] in the baseline task and then even showed substantial convergence to Mirabella during the map task. This suggests that the attitude of a speaker towards the feature in question might influence their accommodating behavior, but does not fully predict it.

Long vowel ⟨**-ä-**⟩ The analysis of the [ɛː]/[eː] contrast by measuring the difference in Euclidean distance in the F1-F2 space did not reveal an accommodation effect for either of the two experimental groups. Only a stronger divergence tendency among the participants with a baseline preference for [eː] in the synthetic group was predicted by the statistical model. The absence of substantial accommodating behavior on the group level was not expected, since participants of a previous shadowing experiment converged with respect to the [ɛː]/[eː] contrast when shadowing natural stimuli, but also, to a smaller extent, when shadowing HMM-based stimuli (Gessinger, Raveh, Steiner, & Möbius, 2021). However, formulating a new utterance entails a higher cognitive load than repeating a given utterance. Therefore, the attention to phonetic detail at the level needed to capture the fine-grained differences in vowel quality may not have been available to the participants of the present study. In addition, it is possible that the gradual change in vowel quality is generally more difficult for speakers to access and control than the binary variation between fricative and plosive in the case of the [ɪç]/[ɪk] contrast or the different forms of question intonation. Other ways of evaluating the [ɛː]/[eː] contrast, e.g., as a categorical change between [ɛː] and [eː], may provide more insight.

[13] This is consistent with the assumption made in Mitterer and Müsseler (2013) that speakers are often unaware which variant of [ɪç]/[ɪk] they use.



(a) Natural group. (b) Synthetic group.

Fig. 11. Accommodation behavior of the 42 participants on the three examined features. The colors code substantial convergence, moderate convergence, maintenance, moderate divergence, and substantial divergence. Some participants converge with respect to two features, some only for one feature, and some do not converge at all. (For a color version of the figure, the reader is referred to the web version of this article)

4.2.3. Individual behavior

To get an impression of the individual accommodating behavior within the two experimental groups, we determined for each participant whether they converged, diverged, or maintained their preference for the three analyzed features. The accommodation to the question intonation and the [ɪç]/[ɪk] contrast was further classified as being moderate or substantial. For the [ɛː]/[eː] contrast the individual result reflects a combination of a significant shift away from the participants' own baseline vowel productions and towards/away from Mirabella.[14] Fig. 11 shows the resulting individual accommodating behavior of the 42 participants.

The majority of cases of substantial convergence are found for the question intonation in both experimental groups. Overall convergence (moderate and substantial) in the natural group is led by question intonation as well ($n = 16$), followed by the [ɪç]/[ɪk] contrast ($n = 11$), and even three individual cases of vowel convergence, i.e., for speakers *Nm02*, *Nf04* and *Nf15*. In the synthetic group question intonation and [ɪç]/[ɪk] contrast are on par (both: $n = 16$) and no individual cases of vowel convergence occurred. Occasional divergence is found for the [ɪç]/[ɪk] contrast (speakers *Nf03*, *Sm02*, *Sf01*, and *Sf13*) and the [ɛː]/[eː] contrast (speakers *Nm01*, *Nm03*, *Sm01*, *Sm05*, and *Sf12*).

According to these measures, 60% of the participants converged to two out of the three tested features (natural: $n = 12$, synthetic: $n = 13$) and 28% to one feature only (both: $n = 6$). Very few participants did not converge at all (natural: $n = 2$, synthetic: $n = 3$). This confirms that accommodating behavior with respect to one phonetic feature does not necessarily predict the behavior with respect to another feature, which was previously documented for acoustic-prosodic features in HHI (e.g., Priva & Sanker, 2018; Reichel, Beňuš, & Mády, 2018; Weise & Levitan, 2018).

In the questionnaire administered after the experiment, only very few participants stated that they had consciously perceived some of the tested features. Three participants pointed out that Mirabella produced ⟨-ig⟩ differently than they expected. Among them were *Nm02* and *Nm14*, who showed no accommodation to Mirabella with respect to this feature, and *Nm01*, who converged substantially. Two other participants commented on the varying question intonation, namely *Sm03*, who did not change their own intonation at all, and *Nm03*, who adopted both the rising intonation and the shifted pitch accent from Mirabella at the fourth trial. This illustrates on a small scale that the conscious perception of a phonetic change neither necessarily leads to nor prevents accommodation. The extent to which the other participants consciously reflected on pronunciation characteristics of Mirabella cannot be further evaluated.

4.3. Conclusion

In summary, the participants of the present study accommodated their phonetic productions to the speech of a virtual language learning tutor with respect to two out of three tested features, i.e., question intonation and the allophonic contrast [ɪç] vs. [ɪk]. This shows that accommodating behavior in users of a SDS is indeed triggered by locally anchored phonetic features. Also in line with our predictions (see section 1.3), the accommodation occurred in the form of convergence. This was the expected behavior under both the assumption that alignment between interlocutors is an automatic process (cf. *IAM*, Pickering & Garrod, 2004; Pickering & Garrod, 2013) and the assumption that we aim to decrease social distance to an interlocutor by converging to them (cf. *CAT*, Giles, 1973; Giles et al., 1991; Shepard et al., 2001), since the participants considered Mirabella to be likable. The participants did not accommodate to the allophonic contrast [ɛː] vs. [eː], which in turn demonstrates that phonetic convergence does not necessarily occur for all features.

The absence of accommodating behavior at the group level, as in the present case of [ɛː]/[eː], may be related to the salience of the feature in question: if it is not recognized as a potential target for accommodation (consciously or subconsciously), it cannot lead to convergence. Considering the motivation to reduce social distance or to facilitate communication with an interlocutor through convergence, it is possible that different phonetic features contribute to these goals to varying degrees and speakers may implement accommodation accordingly.

As expected, we found considerable variation with respect to the degree and direction of accommodation on the level of individual speakers. It has already been suggested that a model of phonetic accommodation that combines the *automatic approach* (IAM) and the *social approach* (CAT) is influenced by additional factors (e.g., Lewandowski, 2012). For example, various aspects of the speaker disposition may be associated with individual differences in accommodating behavior (Yu et al., 2013; Lewandowski & Jilka, 2019). We

[14] As pointed out in section 3.4, the individual results for the [ɛː]/[eː] contrast are based on unadjusted p-values. When adjusting the p-values to control the false discovery rate, one case of convergence remains in the natural group and three cases of divergence in the synthetic group.

tested the influence of the "Big Five" personality traits on the accommodating behavior in our data, which revealed a tendency for neuroticism to influence the convergence of question intonation. Openness, which had previously also been shown to positively correlate with convergence, did not appear as a predictor in our data.

In keeping with our predictions, the overall results did not differ between the experimental groups that communicated with either the natural or the synthetic speech version of Mirabella. Mirabella's synthetic voice was clearly identifiable as non-natural, which did not prevent nor promote accommodating behavior. It remains unclear to what extent the presumed advantages that the different voice types hold (see section 1.3) have worked in their favor, e.g., *natural voice:* potentially more straight-forwardly perceived as social actor, therefore more accommodation according to CAT; *synthetic voice:* potentially perceived as more machine-like and more likely to benefit from convergence.

Initially, we had hypothesized that accommodation might be weakened by the fact that in this study native speakers of German, who are most likely confident in their own pronunciation, interact with a virtual language learning tutor for German whom they might perceive as hierarchically inferior. This does not seem to be the case. Alternatively, our second assumption may have counteracted the latter, namely that the participants probably did not perceive the SDS as fully linguistically flexible and therefore again assumed that it could likely benefit from convergence. Extending the present study to L1-L2 communication by having non-native speakers of German interact with Mirabella is a next step to further investigate these dynamics (see Gessinger, Möbius, Andreeva, Raveh, & Steiner, 2020 for native speakers of French).

Finally, we conclude that phonetic accommodation on the level of local prosody and segmental pronunciation occurs in users of SDSs. This may be exploited, for example, in computer-assisted language learning applications in a way that is beneficial for many users of such systems.

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.

Acknowledgements

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 Nauman Fakhar for implementing the WOz platform, Bistra Andreeva for helping with the evaluation of the question intonation, Jens Neuerburg for assistance in the annotation process, Katie Ann Dunfield for assistance in the recording process, Christine Mangold for creating the illustrations for the WOz platform, and three anonymous reviewers for very helpful comments on an earlier draft of this manuscript.

Appendix A. Guiding utterances

Some examples of the utterances that are available to the experimenter during the experiment in order to react spontaneously to the behavior of the participants.

1. Ja. (*Yes.*)
2. Nein. (*No.*)
3. Das weißich leider nicht. (*Unfortunately, I do not know that.*)
4. Ok? (*Ok?*)
5. Ok. (*Ok.*)
6. Du bist dran! (*It is your turn!*)
7. Versuch's nochmal! (*Try again!*)
8. Zurück zur Aufgabe! (*Back to the task!*)
9. Sehr gut! (*Very good!*)
10. Fast geschafft! (*Almost done!*)
11. Stell' eine Frage! (*Ask a question!*)
12. Verwende die angegebenen Wörter! (*Use the given words!*)
13. Lass mich überlegen. . . (*Let me think. . .*)
14. Sehr gute Frage! (*Great question!*)

Appendix B. Target and filler words

Overview of the 71 words presented in task 1. The ten animals are used in the question-and-answer game. With the exception of *Affe* and *Hase*, all words are used in the map task. The target words contain the allophonic contrasts. Corresponding graphemes are set in bold.

1. **Target words**
   - ► [ɛː] vs. [eː]
   - (a) S**ä**ge (*saw*)
   - (b) M**ä**dchen (*girl*)
   - (c) K**ä**fer (*beetle*)
   - (d) B**ä**r (*bear*)
   - (e) Universit**ä**t (*university*)
   - (f) K**ä**se (*cheese*)
   - (g) J**ä**ger (*hunter*)
   - (h) Gl**ä**ser (*glass, pl.*)
   - (i) verspätet (*delayed*)
   - (j) **ä**hnlich (*similar*)
   - (k) gef**ä**hrlich (*dangerous*)
   - (l) gew**ä**hlt (*elected*)
   - ► **[ɪç] vs. [ɪk]**
   - (a) Kön**ig** (*king*)
   - (b) Hon**ig** (*honey*)
   - (c) mut**ig** (*brave*)
   - (d) schatt**ig** (*shady*)
   - (e) schmutz**ig** (*dirty*)
   - (f) vorsicht**ig** (*cautious*)
   - (g) hungr**ig** (*hungry*)
   - (h) lust**ig** (*funny*)
   - (i) traur**ig** (*sad*)
   - (j) neugier**ig** (*curious*)
   - (k) bill**ig** (*cheap*)
   - (l) biss**ig** (*likely to bite*)
2. **Filler words**
   - Pferd (*horse*)
   - Fisch (*fish*)
   - Kuh (*cow*)
   - Maus (*mouse*)
   - Hund (*dog*)
   - Katze (*cat*)
   - Löwe (*lion*)
   - Vogel (*bird*)
   - Hase (*rabbit*)
   - Affe (*monkey*)
   - Haus (*house*)
   - Baum (*tree*)
   - Autos (*car, pl.*)
   - Kuchen (*cake*)

- Bahnhof (*train station*)
- Bus (*bus*)
- Apfelsaft (*apple juice*)
- Blumen (*flower, pl.*)
- Zwillinge (*twin, pl.*)
- See (*lake*)
- Flughafen (*airport*)
- Computer (*computer*)
- Wald (*forest*)
- Politiker (*politician*)
- Museum (*museum*)
- leer (*empty*)
- schwer (*heavy*)
- schlau (*smart*)
- laut (*loud*)
- müde (*tired*)
- rund (*round*)
- neu (*new*)
- kalt (*cold*)
- berühmt (*famous*)
- wild (*wild*)
- schön (*beautiful*)
- groß (*big*)
- teuer (*expensive*)
- alt (*old*)
- gesund (*healthy*)
- nass (*wet*)
- modern (*modern*)
- klein (*small*)
- dunkel (*dark*)
- süß (*sweet*)
- sauber (*clean*)
- interessant (*interesting*)

Appendix C. Questions from fragments and answers

Questions to be formulated by the participants in task 2 with the provided fragments (•) and corresponding answers given by Mirabella (○):

- *Wann hat Italien den Euro eingeführt?*
  When did Italy introduce the Euro?
  - ○ *Italien hat den Euro 1999 eingeführt.*
    Italy introduced the Euro in 1999.
- *Was ist die Hauptstadt von Lettland?*
  What is the capital of Latvia?
  - ○ *Die Hauptstadt von Lettland ist Riga.*
    The capital of Latvia is Riga.
- *Wo sind die Brüder Grimm geboren?*
  Where were the Brothers Grimm born?
  - ○ *Die Brüder Grimm sind in Hanau geboren.*
    The Grimm brothers were born in Hanau.
- *Wer war die erste Frau im Weltall?*
  Who was the first woman in space?
  - ○ *Walentina Tereschkowa war die erste Frau im Weltall.*
    Valentina Tereshkova was the first woman in space.
- *Wie viele Tage hat der August?*
  How many days are in August?
  - ○ *Der August hat 31 Tage.*
    August has 31 days.

Appendix D. Map task prepositions

The prepositions used in task 4 govern either the accusative case [ACC] or the dative case [DAT].

- *um* [ACC] *herum* (around)
- *aus* [DAT] *heraus* (out of)
- *in* [ACC] *hinein* (into)
- *an* [DAT] *vorbei* (past)
- *durch* [ACC] *hindurch* (through)

Appendix E. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.wocn.2021.101029.

References

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

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology, 37*(5), 715. https://doi.org/10.1037/0022-3514.37.5.715.

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30*(1), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437–456. https://doi.org/10.1017/S0047404510000400.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phoneticimitation. *Journal of Phonetics, 40*, 177–189. https://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. *Laboratory Phonology, 5*(1), 123–150. https://doi.org/10.1515/lp-2014-0006.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Baylor, A., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. In *EdMedia + Innovative Learning* (pp. 452–458). Hawaii: Honolulu.

Bell, L., Gustafson, J., & Heldner, M. (2003). Prosodic adaptation in human-computer interaction. *International Congress of Phonetic Sciences* (pp. 2453–2456). Barcelona: (ICPhS).

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

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). https://doi.org/10.21437/SpeechProsody.2018-45.

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/.

Borkenau, P., Ostendorf F. (2007). NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (Vol. 2. neu normierte und vollständig überarb. Auflage). Göttingen: Hogrefe.

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. Frontiers in Psychology 6(1187). https://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. *Journal of Pragmatics, 42*(9), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012.

Coles-Harris, E. H. (2017). Perspectives on the motivations for phonetic convergence. *Language and Linguistics Compass, 11*(12) https://doi.org/10.1111/lnc3.12268.

Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes, 23*(4), 528–556. https://doi.org/10.1080/01690960801920545.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how. *Knowledge-based Systems, 6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N.

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

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics, 78*(1), 317–333. https://doi.org/10.3758/s13414-015-0982-6.

Dudenredaktion (2015). Duden – Das Aussprachewörterbuch: Betonung und Aussprache vonüber 132.000 Wörtern und Namen. In *Duden – Deutsche Sprache in 12 Bänden* (Vol.). Mannheim: Bibliographisches Institut GmbH.

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology, 4*(346) https://doi.org/10.3389/fpsyg.2013.00346.

Ellbogen, T., Schiel, F., & Steffen, A. (2004). The BITS speech synthesis corpus for German. In *International Conference on Language Resources and Evaluation*

(LREC) (pp. 2091–2094). Lisbon. url:http://www.lrec-conf.org/proceedings/lrec2004/pdf/72.pdf.

Forbes-Riley, K., Litman, D. J., Silliman, S., & Tetreault, J. R. (2006). Comparing Synthesized versus Pre-Recorded Tutor Speech in an Intelligent Tutoring Spoken Dialogue System. In *Florida Artificial Intelligence Research Society Conference (FLAIRS)* (pp. 509–514). Florida: Melbourne Beach.

Gauder, L., Reartes, M., Gálvez, R. H., Beňuš, Š., & Gravano, A. (2018). Testing the effects of acoustic/prosodic entrainment on user behaviour at the dialog-act level. In *International Conference on Speech Prosody. Poznań* (pp. 374–378). https://doi.org/10.21437/SpeechProsody.2018-76.

Gessinger, I., Möbius, B., Andreeva, B., Raveh, E., Steiner I. (2019a). Phonetic accommodation in a Wizard-of-Oz experiment: intonation and segments. In: Interspeech. Graz, pp. 301–305. https://doi.org/10.21437/Interspeech.2019-2445.

Gessinger, I., Möbius, B., Andreeva, B., Raveh, E., Steiner, I. (2020). Phonetic accommodation of L2 German speakers to the virtual language learning tutor Mirabella. In: Interspeech. Shanghai, pp. 4118–4122, https://doi.org/10.21437/Interspeech.2020-2701.

Gessinger, I., Möbius, B., Fakhar, N., Raveh, E., & Steiner, I. (2019b). A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction. In *International Congress of Phonetic Sciences (ICPhS) Melbourne* (pp. 1475–1479). url_:https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS1524.pdf..

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

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: behavior of groups and individuals in speech shadowing. *Speech Communication, 127*, 43–63. https://doi.org/10.1016/j.specom.2020.12.004.

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). https://doi.org/10.21437/SpeechProsody.2018-46.

Gijssels, T., Staum Casasanto, L., Jasmin, K., Hagoort, P., & Casasanto, D. (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes, 53*(4), 233–251. https://doi.org/10.1080/0163853x.2015.1023965.

Giles, H. (1973). Accent mobility: a model and some data. In: Anthropological Linguistics, pp. 87–105.

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

Gregory, S. W., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70*(6), 1231–1240. https://doi.org/10.1037/0022-3514.70.6.1231.

Grice, M., & Baumann, S. (2002). Deutsche Intonation und GToBI. *Linguistische Berichte, 191*, 267–298.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS), 2*(1), 26–41. https://doi.org/10.1145/357417.357420.

Kiesewalter, C. (2019). Zur subjektiven Dialektalität regiolektaler Aussprachemerkmale des Deutschen. Franz Steiner Verlag. isbn: 9783515124379. url_:https://elibrary.steiner-verlag.de/book/99.105010/9783515124430.

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? *Behavioral and Brain Sciences, 27*(2), 203–204. https://doi.org/10.1017/S0140525X0436005X.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Le Maguer, S., Steiner, I., Tombini, F., Deb, P., Basu, M., Kröger I. (2018). Agile MaryTTS Architecture for the Blizzard Challenge 2018. In: Blizzard Challenge. Hyderabad. url: http://festvox.org/blizzard/bc2018/MARY_BlizzardChallenge2018.pdf.

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/archivepapers/interspeech2010/i100793.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. https://doi.org/10.21437/Interspeech.2016-985.

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/archivepapers/interspeech2011/i113081.pdf.

Lewandowski, N. (2012). Talent in nonnative phonetic convergence. PhD thesis. Universität Stuttgart. https://doi.org/10.18419/opus-2858.

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality & attention. *Frontiers in Communication, 4*(18). https://doi.org/10.3389/fcomm.2019.00018.

Litman, D., & Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. *Demonstration papers at HLT-NAACL, 2004*, 5–8. https://doi.org/10.3115/1614025.1614027.

Lubold, N., Pon-Barry H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In: ACM workshop on Multimodal Learning Analytics, pp. 5–12. https://doi.org/10.1145/2666633.2666635.

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). https://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. *Evolution and Human Behavior, 34*(6), 419–426. https://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. https://doi.org/10.21437/Interspeech.2017-1520.

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks?. *Attention, Perception, & Psychophysics, 75*(8), 1817–1826. https://doi.org/10.3758/s13414-013-0517-y.

Mitterer, H., & Müsseler, J. (2013). Regional accent variation in the shadowing task: evidence for a loose perception-action coupling in speech. *Attention, Perception & Psychophysics, 75*(3), 557–575. https://doi.org/10.3758/s13414-012-0407-8.

Möbius, B. (1993). Ein quantitatives Modell der deutschen Intonation. In *Analyse und Synthese von Grundfrequenzverläufen*. Niemeyer.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems, 99*(7), 1877–1884.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171. https://doi.org/10.1037/1076-898x.7.3.171.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://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). ACM. https://doi.org/10.1145/191666.191703.

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

Oviatt, S., Darves, C., & Coulston, R. (2004). Toward adaptive conversational interfaces: modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction, 11*, 300–328. https://doi.org/10.1145/1017494.1017498.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*(4), 2382–2393. https://doi.org/10.1121/1.2178720.

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. *Journal of Phonetics, 69*, 1–11. https://doi.org/10.1016/j.wocn.2018.04.001.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190. https://doi.org/10.1017/S0140525X04450055.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347. https://doi.org/10.1017/s0140525x12001495.

Priva, U. C., & Sanker, C. (2018). Distinct behaviors in convergence across measures. In *Annual Meeting of the Cognitive Science Society (CogSci), Austin, TX* (pp. 1515–1520).

R Core Team (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. url:https://www.R-project.org.

Raveh, E., Siegert, I., Steiner, I., Gessinger, I., Möbius B. (2019). Three's a crowd? Effects of a second human on vocal accommodation with a voice assistant. In: Interspeech. Graz, pp. 4005–4009. url_:https://www.iscaspeech.org/archive/Interspeech2019/pdfs/1825.pdf.

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. *Computers & Mathematics with Applications, 33*(5), 128.

Reichel, U. D., Beňuš, Š., & Mády, K. (2018). Entrainment profiles: Comparison by gender, role, and feature set. *Speech Communication, 100*, 46–57. https://doi.org/10.1016/j.specom.2018.04.009.

RStudio Team (2016). RStudio: Integrated Development Environment for R. RStudio, Inc. Boston, MA. url:https://www.rstudio.com/.

Schweitzer, A., & Lewandowski, N. (2014). Social factors in convergence of F1 and F2 in spontaneous speech. In *International Seminar on Speech Production*. Cologne.

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. https://doi.org/10.21437/Interspeech.2017-1248.

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

Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on personality perception. *Language and Speech, 18*(2), 145–152. https://doi.org/10.1177/002383097501800203.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In *32nd Annual Meeting of the Cognitive Science Society (CogSci)* (pp. 127–132). Cognitive Science Society.

Steiner, I., & Le Maguer, S. (2018). Creating New Language and Voice Components for the Updated MaryTTS Text-to-Speech Synthesis Platform. In *Language Resources and Evaluation Conference (LREC)* (pp. 3171–3175). Miyazaki. url_:http://www.lrec-conf.org/proceedings/lrec2018/summaries/1045.html.

Thomason, J., Nguyen, H. V., Litman D. (2013). Prosodic entrainment and tutoring dialogue success. In: International Conference on Artificial Intelligence in Education. Springer, pp. 750–753. https://doi.org/10.1007/978-3-642-39112-5104.

Tokuda, K., Zen, H., & Black, A. W. (2002). An HMM-based speech synthesis system applied to English. In *IEEE Workshop on Speech Synthesis, Santa Monica, CA* (pp. 227–230). https://doi.org/10.1109/WSS.2002.1224415.

Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., Barry W. J. (2006). Modelling personality features by changing prosody in synthetic speech. In: International Conference on Speech Prosody. Dresden. url: https://www.isca-speech.org/archive/sp2006/papers/sp06088.pdf.

Ward, A., Litman D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In: Workshop on Speech and Language Technology in Education (SLaTE). Farmington, PA, pp. 57–60. url_:https://www.isca-speech.org/archiveopen/archivepapers/slate 2007/sle7057.pdf.

Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies* (vol. 2, pp. 297–302). New Orleans, LA. https://doi.org/10.18653/v1/N18-2048.

Wochner, D., Schlegel, J., Dehé, N., Braun B. (2015). The prosodic marking of rhetorical questions in German. In: Interspeech. Dresden, pp. 987–991. url:https://www.isca-speech.org/archive/interspeech2015/papers/i150987.pdf.

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. https://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.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication, 51*, 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004.