A WIZARD-OF-OZ EXPERIMENT TO STUDY PHONETIC ACCOMMODATION IN HUMAN-COMPUTER INTERACTION

Iona Gessinger[1,2], Bernd Möbius[1], Nauman Fakhar[1], Eran Raveh[1,2], Ingmar Steiner[1−3]

[1]Language Science and Technology, Saarland University, Germany
[2]Multimodal Computing and Interaction, Saarland University, Germany
[3]German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
gessinger@coli.uni-saarland.de

ABSTRACT

This paper presents a Wizard-of-Oz experiment designed to study phonetic accommodation in human-computer interaction. The experiment comprises a dynamic exchange of information between a human interlocutor and a supposedly intelligent system, while allowing for planned manipulation of the system's speech output on the level of phonetic detail. In the current configuration of the experiment, we are targeting convergence in allophonic contrasts and phenomena of local prosody. A study was conducted with 12 German native speakers. The results of a map task show highly speaker dependent behavior for the contrast [ɪç] vs. [ɪk] occurring in the German suffix ⟨-ig⟩: during the baseline production of the target items, speakers either consistently choose one allophone or use both interchangeably. When conversing with the system, some speakers converge to its speech output, while others maintain their preferred variant or even diverge. This reflects individual variation observed in previous work on accommodation.

Keywords: human-computer interaction, Wizard-of-Oz experiment, phonetic accommodation

1. INTRODUCTION

It has been shown that phonetic accommodation occurs in human-human interaction (HHI) [2,3,16,20]. Interlocutors adjust their speech output to the speech input they receive during a conversation. As a result, it can become more or less similar (*converging* or *diverging*, respectively) with respect to a variety of phonetic features.

Under the assumption that phonetic accommodation is both internally motivated, for instance in the form of automatic perception–production integration, and externally motivated, for example as a means to indicate the social relationship to an interlocutor [7], we expect convergence to be the unmarked behavior. Divergence is then 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. To account for the varying degree of accommodating behavior found in different speakers [21], we need to consider that these two mechanisms might be developed and weighted differently between speakers.

Since phonetic accommodation is assumed to contribute to the ease and success of HHI [23], it is a relevant topic for human-computer interaction (HCI) research as well. A growing number of studies is exploring whether humans accommodate to the speech output of spoken dialogue systems (SDSs) and how accommodating behavior on the part of an SDS is perceived by the user. The phonetic features that are examined in this context are mainly of global acoustic-prosodic nature, such as pitch, intensity and speaking rate [4, 5, 8, 11, 14, 15, 19, 26]. Overall results show that humans indeed accommodate to such features in HCI, too. However, it remains unclear whether the underlying factors motivating convergence are the same as in HHI.

We aim to explore whether accommodating behavior in HCI also occurs in more locally anchored prosodic phenomena, specifically pitch accent realization and intonation of constituent questions, and on the level of segmental pronunciation, such as the German allophone pairs [ç] vs. [k] occurring in the suffix ⟨-ig⟩, therefore henceforth [ɪç] vs. [ɪk], and [ɛː] vs. [eː]. Prior work has shown that speakers generally accommodate to such phenomena [1, 12, 13, 18, 25].

Up to now, SDSs themselves are not phonetically responsive to the user input. Suggestions for models intended to enable the computer to show phonetically accommodative behavior are being developed [15,24], yet, to the best of our knowledge, there is no system which could be employed to study the user side. We hence apply the Wizard-of-Oz (WOz)

method to simulate an intelligent SDS as it is good practice in HCI research [4,9–11,19]. While the user believes to interact with an autonomous system, it is in fact the *wizard*, i.e., the experimenter, who makes decisions about the system's responses.

This paper presents the structure of the WOz experiment in its entirety to give a coherent overview of the interaction. It further shows the results of a first study on the [ɪç] vs. [ɪk] contrast. Based on the individual variation observed in previous work on accommodation, we expect that the participants can be grouped into two classes: those who converge to the system and those who maintain their preferred variant of the feature.

2. EXPERIMENTAL SETUP

2.1. Setting

The WOz experiment is presented to the user as an application for learning the German language. The text material used in the experiment is therefore chosen to be accessible to advanced learners of German. This resembles a realistic use case as it simulates a scenario from the growing field of computer-assisted language learning. The experiment can therefore be disguised as a test of the application, which motivates the situation for the user and shifts the focus from the user being tested to the system.

The system introduces itself as a female trainer for German as a foreign language called *Mirabella*. The user only interacts with Mirabella's voice; she is not represented by an avatar. All utterances available to the *wizard* to choose from during the experiment were pre-recorded by a female German native speaker aged 26 years. The recordings were carried out with a sampling rate of 48 kHz using a stationary cardioid microphone in a sound-attenuated booth.

The interaction is supported by visualization of the tasks on a screen. Mirabella explains the tasks to the user and takes part in them, either by taking turns in a question-and-answer exchange with the user (cf. task 3) or by providing missing information to the user in a map task (cf. task 4).

2.2. Structure

The experiment consists of four tasks. The first two tasks familiarize the user with the system and text material occurring in the experiment, and elicit baseline productions of target items. Task 3 tests pitch accent realization and intonation of constituent questions, and task 4, the realization of the allophone pairs [ɪç] vs. [ɪk] and [ɛː] vs. [eː].

Task 1 All pictures as well as English translations of the adjectives occurring in the experiment are presented to the user. The latter names the pictures and translates the English adjectives to German by uttering them in the following carrier sentence:

Das Wort ⟨***item***⟩ *kenne ich.*
The word ⟨item⟩ is known to me.

This task thus ensures that the text material is known to the user and reveals which versions of [ɪç] vs. [ɪk] and [ɛː] vs. [eː] they naturally prefer. The individual realizations are perceptually categorized by the experimenter. Note that we consider fricative variants such as [ʃ] or [ɕ] as part of the [ɪç] category. The preference is stored in the system and retrieved in task 4.

Figure 1: Where did the animals hide?

Task 2 The user formulates five wh-questions whose components are given as fragments, e.g., *wo – die Brüder Grimm – geboren sein* (where – the Brothers Grimm – born). Mirabella talks for the first time when answering these questions.

This task familiarizes the user with Mirabella's voice and reveals the intonation they usually apply when producing constituent questions.

Task 3 Mirabella and the user take turns asking and answering each other about ten animals hiding in ten houses (Fig. 1), 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 task includes two rounds of 20 turns, with Mirabella and the user each asking and answering 10 questions per round.

The realization of questions and answers on the part of the system differs between round one and round two with respect to pitch accent placement and intonation to give room for accommodation.

In round 1, Mirabella produces all questions with nuclear pitch accent on ⟨*animal*⟩ followed by terminal *f0* fall, whereas in round 2, all questions are produced with nuclear pitch accent on the interrogative

pronoun *wo* followed by terminal high *f0* rise.

Mirabella's answers carry two pitch accents in round 1, namely on ⟨*animal*⟩ and ⟨*number*⟩, while they carry three pitch accents in round 2, namely on ⟨*animal*⟩, *Haus,* and ⟨*number*⟩.

Task 4 The information about the user's preference with respect to the [ɪç] vs. [ɪk] and [ɛː] vs. [eː] contrasts is automatically retrieved from the results of task 1. Mirabella then uses the dispreferred forms throughout the entire task.

The user describes the path from leaving the house until reaching the destination on a map (see Fig. 2) while using the prepositions given on the right side of the screen and subsequently describing the object in question with the adjective given next to it, as follows:

a) *Ich gehe um den* ***Honig*** *herum.*
I am walking around the honey.

b) *Der* ***Honig*** *ist süß.*
The honey is sweet.

Some of the objects and adjectives are hidden behind boxes. The user asks Mirabella about these items and she provides the missing information:

obj. *Hinter der* ⟨***color***⟩ *Box ist* ⟨***the object***⟩.
Behind the ⟨color⟩ box is ⟨the object⟩.

adj. *Das Wort hinter der* ⟨***color***⟩ *Box ist* ⟨***adjective***⟩.
The word behind the ⟨color⟩ box is ⟨adjective⟩.

Given this information, the user can formulate the required utterance. If the target item is an object, it will occur twice in the utterance (*Honig* in the example above); if the target item is an adjective, it will occur only once, in the second part of the utterance.

The task consists of four maps with nine object-adjective pairs each. Each map contains three pairs including an [ɪç] vs. [ɪk] target (e.g., *Computer – billig*), three pairs including an [ɛː] vs. [eː] target (e.g., *Löwe – gefährlich*), and three filler pairs not including a target (e.g., *Wald – dunkel*).

3. EXPERIMENT

3.1. Participants

12 German native speakers (9 female) with a mean age of 22 years (range 18 to 29) were recruited on the Saarland University campus and paid for taking part in the study. Two participants have a second native language, with one of them indicating that German is not dominant in his everyday life. All but two participants are students of linguistics and may therefore be more aware of the phenomenon tested in this study than the average user of an SDS.

3.2. Recordings

During the interaction with Mirabella, participants were seated in front of a monitor and recorded in the same manner as described in Section 2.1. Mirabella's utterances were played to the participants over headphones. The interaction with Mirabella lasted about 30 min, including short breaks after tasks 1 and 3. The recordings were followed by a questionnaire about Mirabella including several 5-point Likert scales.

Figure 2: How do you reach the destination?

3.3. Results

The overall interaction between Mirabella and the participants went smoothly. Minor deviations from the planned course on the part of some participants were successfully guided back on track through directions given by Mirabella. The only deviation that could not be recovered occured for participant f01 in task 4 who consistently used pronouns instead of repeating the noun in the second part of the utterance. This resulted in missing values for the respective items (Fig. 3).

The results of the 5-point Likert scales reveal that participants considered Mirabella to be likeable (*unpleasant* to *very likeable* – mean: 4.4), competent (*incompetent* to *very competent* – mean: 4.1) and very well intelligible (*badly* to *very well* – mean: 5). They also considered Mirabella's reaction time to be appropriate (*too slow* to *too fast* – mean: 2.8).

Figure 3 shows the results for the [ɪç] vs. [ɪk] contrast. Half of the participants have a preference to produce [ɪç] during task 1 (base), the other half prefer [ɪk]. Eight participants consistently choose one allophone when naming the items in this task, three participants (all with preference [ɪk]) produce the dispreferred allophone once (m01 in *schattig*, f04 in *vorsichtig*, f09 in *neugierig*), and only participant f03 uses both allophones almost equally frequently.

During task 4 (map), four behavioral patterns are discernible: five participants converge to Mirabella's realization of the suffix ⟨-ig⟩ in at least one and maximally 13 out of 14 cases. Another five participants

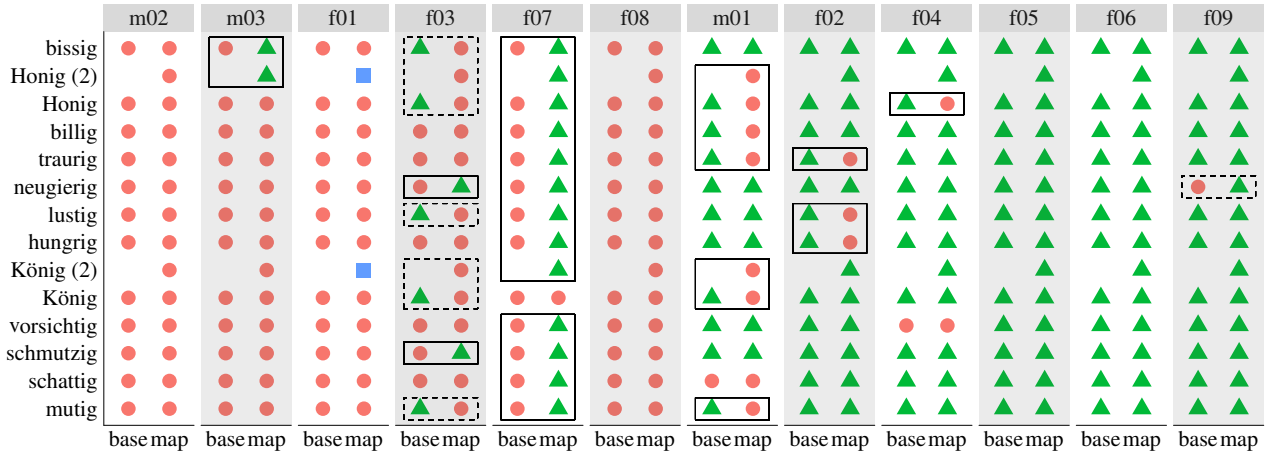

Figure 3: Realization of suffix ⟨-ig⟩ as [ɪç] ● or [ɪk] ▲ in the baseline task (1) and the map task (4). Target words are given in the order of occurrence in the map task, starting with *mutig*. Solid boxes indicate convergence, dashed boxes divergence. The ■ indicates missing values. Participants are grouped by preference: [ɪç] left and [ɪk] right.

maintain their preferred variant in all cases. Participant f09 maintains her preferred variant and diverges in the case of the only item she produced differently during the baseline task. Participant f03 maintains (5 cases), diverges (7 cases), and converges as well (2 cases).

3.4. Discussion

Of the 12 participants in the study, the two largest groups showed either maintaining or converging behavior. Occasional diverging behavior was observed as well. One participant showed a combination of all three behavior types.

As the majority of participants in this group were students of linguistics, the awareness of the German [ɪç] vs. [ɪk] contrast might have been above average. One of the two non-linguists (f07) showed the biggest convergence effect.

The questionnaire completed after the interaction with Mirabella further revealed that m02, f01, and f09 have a negative attitude towards the allophone they do not believe to produce themselves: they consider it to be either "wrong" or at least "not aesthetically pleasing". Therefore, it comes as no surprise that they did not converge or, in the case of the second non-linguist (f09), even diverged. All other participants consider the allophone they do not believe to produce themselves acceptable as well.

Participants m01, f02, and f05 believe to realize the suffix ⟨-ig⟩ as [ɪç] in their everyday life, which does not match their preference in task 1. This mistaken belief is compatible with the fact that m01 and f02 show converging behavior in task 4, yet difficult to reconcile with f05 not converging at all.

It has been shown that humans converge more readily to an interlocutor they consider more attractive [2, 17]. Furthermore, it has been suggested that users adapt their language more to that of the system if they consider the latter to be less capable and more likely to benefit from the convergence [6, 22]. However, no difference between the assessment of Mirabella's likeability and competence was found between the different participant groups in this study.

4. CONCLUSION AND FUTURE WORK

A WOz experiment simulating an intelligent SDS was conducted to elicit phonetically accommodative behavior of users in an HCI scenario. The results pertaining to the German [ɪç] vs. [ɪk] contrast were reported in this paper. Some participants converged to the SDS with respect to this contrast. We assume that awareness of and attitude towards the specific contrast in question might be a reason for some participants not to converge. Therefore, participants with average linguistic knowledge should be tested, too. Furthermore, it is possible that more convergence occurs if the SDS appears less competent, for instance because it asks the user to repeat some of their utterances. Such interruptions may be introduced judiciously in the course of the experiment. We are planning to analyze more features, extend the user group to non-native speakers of German, and apply synthetic speech instead of natural recordings in Mirabellas's utterances.

5. ACKNOWLEDGEMENTS

This research was funded in part by the German Research Foundation grants MO 597/6-2 and STE 2363/1-2. We thank Jens Neuerburg (annotations), Katie Ann Dunfield (recordings), and Christine Mangold (illustrations).

6. REFERENCES

[1] Babel, M. 2009. *Phonetic and social selectivity in speech accommodation*. PhD thesis University of California, Berkeley.

[2] Babel, M., McGuire, G., Walters, S., Nicholls, A. 2014. Novelty and social preference in phonetic accommodation. *Laboratory Phonology* 5(1), 123–150.

[3] Bailly, G., Lelong, A. 2010. Speech dominoes and phonetic convergence. *Interspeech* 1153–1156.

[4] Bell, L., Gustafson, J., Heldner, M. 2003. Prosodic adaptation in human-computer interaction. *International Congress of Phonetic Sciences (ICPhS)* Barcelona. 833–836.

[5] Beňuš, Š., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., Levitan, R. 2018. Prosodic entrainment and trust in human-computer interaction. *International Conference on Speech Prosody* Poznań. 220–224.

[6] 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.

[7] Coles-Harris, E. H. 2017. Perspectives on the motivations for phonetic convergence. *Language and Linguistics Compass* 11(12).

[8] Coulston, R., Oviatt, S., Darves, C. 2002. Amplitude Convergence in Children's Conversational Speech with Animated Personas. *ICSLP* Denver. 2689–2692.

[9] Dahlbäck, N., Jönsson, A., Ahrenberg, L. 1993. Wizard of Oz studies – why and how. *Knowledge-based systems* 6(4), 258–266.

[10] DeVault, D., Mell, J., Gratch, J. 2015. Toward natural turn-taking in a virtual human negotiation agent. *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*.

[11] Gauder, L., Reartes, M., Gálvez, R. H., Beňuš, Š., Gravano, A. 2018. Testing the effects of acoustic/prosodic entrainment on user behavior at the dialog-act level. *International Conference on Speech Prosody* Poznań. 374–378.

[12] Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., Steiner, I. 2017. Shadowing synthesized speech – segmental analysis of phonetic convergence. *Interspeech* Stockholm. 3797–3801.

[13] Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., Steiner, I. 2018. Convergence of pitch accents in a shadowing task. *International Conference on Speech Prosody* 225–229.

[14] Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., Casasanto, D. 2016. Speech accommodation without priming: The case of pitch. *Discourse Processes* 53(4), 233–251.

[15] 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. *Interspeech* San Francisco, CA. 1166–1170.

[16] Lewandowski, N. 2012. *Talent in nonnative phonetic convergence*. PhD thesis Universität Stuttgart.

[17] Michalsky, J., Schoormann, H. 2017. Pitch convergence as an effect of perceived attractiveness and likability. *Interspeech* 2253–2256.

[18] 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.

[19] 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.

[20] Pardo, J. S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4), 2382–2393.

[21] 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.

[22] Pearson, J., Hu, J., Branigan, H. P., Pickering, M. J., Nass, C. I. 2006. Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM 1177–1180.

[23] Pickering, M. J., Garrod, S. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences* 27(2), 169–190.

[24] Raveh, E., Steiner, I., Möbius, B. 2017. A computational model for phonetically responsive spoken dialogue systems. *Interspeech* Stockholm. 884–888.

[25] Schweitzer, K., Walsh, M., Schweitzer, A. 2017. To see or not to see: Interlocutor visibility and likeability influence convergence in intonation. *Interspeech* 919–923.

[26] Staum Casasanto, L., Jasmin, K., Casasanto, D. 2010. Virtually accommodating: Speech rate accommodation to a virtual interlocutor. *32nd Annual meeting of the Cognitive Science Society (CogSci 2010)* Portland, OR. 127–132.