Beccard, Marina; Albrecht, Sven; Schmied, Josef

Learning outcomes with text-to-speech synthesis of native and non-native English varieties

Brno studies in English. 2025, vol. 50, iss. 2, pp. 7-33

ISSN 0524-6881 (print); ISSN 1805-0867 (online)

Stable URL (DOI): https://doi.org/10.5817/BSE2024-2-1
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.82125

Learning Outcomes with Text-to-Speech Synthesis of L1 and L2 English Varieties

Brno Studies in English
Volume 50, No. 2, 2024

ISSN 0524-6881 | e-ISSN 1805-0867
https://doi.org/10.5817/BSE2024-2-1

Marina Beccard, Sven Albrecht and Josef Schmied

Abstract
English is a common instruction medium of learning tools using Text-to-Speech (TTS), yet most solutions incorporate only L1 varieties like Standard American English (AmE). At the same time, some research suggests that educational content personalized in the learner's variety is beneficial. We tested this hypothesis with students from the Masaryk University Brno, who listened to a lecture synthesized in Czech English (CzE) and AmE, rated the TTS speaker based on the Robotic Social Attributes Scale and answered questions regarding the contents of the lecture. The learners improved their knowledge similarly with both TTS varieties. Characteristics from the INTELLIGENCE cluster were rated higher than ANTHROPOMORPHISM and LIKABILITY, and the AmE voice was rated more competent than the CzE voice. While the results indicate that the narration may need to be made more engaging, the present study provides first insights into the perceptions of L2 students' own variety and recommendations for customizing credible, learning-facilitating TTS.

Key words
L2 text-to-speech system; Czech English; American English; familiarity; credibility; Robotic Social Attributes Scale (RoSAS)

1. Introduction

English has become a widespread medium of instruction in educational institutions like universities as well as in online courses and tutorials. The new technological developments in Text-to-Speech (TTS) solutions allow the rapid synchronization of course material, making the preparation of materials more time-efficient and easily adaptable. With the rise of the use of English in education, L2 English speakers have access to more resources. Yet, they may be disadvantaged because the language of instruction is usually different from the variety they have been exposed to in their home country, i.e., their own variety. Karakaş (2017) examined 50 TTS products and found that they mostly featured American English (AmE) and British English, with few exceptions like Australian and Indian English. There are no known solutions employing L2 Englishes. While people can indeed adapt to unfamiliar varieties (Cristia et al. 2012), this takes exposure time which could be used more efficiently to learn more information on the subject

matter with less effort. Our hypothesis is that customizing the synthetic voice of pedagogical agents (PAs) to mirror the sociolinguistic variety of the learners has the potential to improve learning outcomes, as learners will have more perceptual and learning benefits from a familiar variety like their own.

In this context, we have to admit that the different terminologies in linguistic subdisciplines may create issues: in traditional language teaching and learning, native vs. non-native was used, in second language acquisition, L1 vs. L2, and in sociolinguistics, e.g. Kachru's three-circle-concept, norm-providing native varieties are in the Inner Circle, norm-developing second-language varieties in the Outer Circle and norm-dependent foreign-language varieties in the Expanding Circle (e.g. English in the Czech Republic) - and all concepts have strong attitudinal implications today. Our original concept (in 2019) was to test the accommodation of non-native English in an academic lingua-franca framework. However, now we use the less controversial concepts of L1 and L2, as in recent years, the concept of native speakers in ELT has been discussed critically (e.g. Kiczkowiak 2018 in Poland) and the Common European Framework of Reference for Languages (CEFR) changed its terminology and descriptors to avoid "native": "Changes are also proposed to certain descriptors that refer to linguistic accommodation (or not) by 'native speakers', because this term has become controversial" (Council of Europe 2020: 24). Generally, the 'intelligibility principle' is clearly distinguished from the 'nativeness principle', where idealised models usually ignored the retention of accent, disregarding educational context, sociolinguistic aspects and learners' needs. The contrast between the new and previous CEFR descriptions is very explicit (Council of Europe 2020: 37):

> It should be emphasised that the top level in the CEFR scheme, C2, has no relation whatsoever with what is sometimes referred to as the performance of an idealised "native speaker", or a "well-educated native speaker" or a "near native speaker". Such concepts were not taken as a point of reference during the development of the levels or the descriptors. C2, the top level in the CEFR scheme, is introduced in the CEFR as follows:
> Level C2, whilst it has been termed "Mastery", is not intended to imply native-speaker or near native-speaker competence. What is intended is to characterise the degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners. (Council of Europe 2001 Section 3.6)

A linguistic theory supporting customization via a familiar variety is the Interlanguage Speech Intelligibility Benefit (ISIB) hypothesis. It expects L2 speakers to perform better in foreign variety processing than L1 speakers, especially if the input is congruent with their language background (Bent and Bradlow 2003). Based on Electroencephalography (EEG) evidence, English learners of Spanish were found to recruit greater cognitive resources when perceiving L1 Spanish compared to English-accented Spanish (Gosselin et al. 2022). Still, evidence on the ISIB is mixed (e.g., Dokovova et al. 2022, Fishero et al. 2023, Hayes-Harb et al. 2008) and more research is needed to support the theory.

In addition to linguistics, our hypothesis is also based on evidence from instructional psychology showing advantages of learning from teachers with a familiar accent. Investigating the personalization effect on the learning outcomes of Austrian pupils at a lower secondary school, Rey and Steib (2013) found that Austrian German was more beneficial than Standard German (Rey and Steib 2013). Listeners who heard the lecture in their local dialect scored higher in retention than those who listened to it in standard German; however, this was not the case for transfer learning (Rey and Steib 2013: 2026), which reflects how well students apply the acquired knowledge to other contexts. In addition to familiarity, language attitudes also play a role in learning. While Ahn and Moore (2011) did not find a negative effect of differently accented voice on learning outcomes, they found that learning outcomes were influenced by attitudes towards the presented varieties. Namely, only students who shared negative attitudes towards Asian English accents prior to the experiment also performed worse when presented with input produced in Korean English (Ahn and Moore 2011).

Looking at synthesised language varieties, familiar local dialects are preferred. In an experiment with robots speaking standard Arabic and a local dialect, when the robots showed high knowledge and rhetorical ability, the participants complied more with the robot speaking the participant's own local dialect (Andrist et al. 2015: 163). In the New Zealand context, a robot speaking New Zealand English was preferred, as it was rated more natural than the robot speaking American English (Tamagawa et al. 2011). Irish children gave preference to the UK than the US voice (Sandygulova and O'Hare 2015).

It should be emphasised that the present study does not investigate L2 TTS in the teaching *of* English but teaching *in* English. In the context of the teaching *of* English, accent can interfere with listening exercise outcomes. For instance, while L1 speakers of Spanish scored significantly higher in a listening task when listening to Spanish-accented English, L1 speakers of Chinese scored significantly lower when listening to Chinese-accented English (Major et al. 2002). Regarding the teaching *in* English of subject matter like Linguistics or Biology to L2 speakers, there is a research gap which we aim to address in the context of Czech English (CzE).

It should also be noted that our lecture scenario features a synthetic voice without a human visual representation. This differs from the many embodied agents in the literature featuring an animated character (see Dai et al. 2022). Note that the lecture is also scripted, i.e., it requires a text which is then synthesised, in contrast to a chatbot. The TTS lecturer is only an information source and does not interact with the users (Schroeder and Gotch 2015, p. 187). To sum up, we test the application of a TTS system in pedagogical scenarios.

In addition to the learning outcomes, we focus on the users' evaluation of the speaker. Conversational and pedagogical agents are perceived as social partners (Louwerse et al. 2009, Mayer et al. 2003b) and as such, they are attributed certain characteristics based on the social and linguistic cues they send. Regardless of whether personality is a relevant factor in the robot's task, humans ascribe personality traits to it (Nass and Lee 2001). Agents employing TTS should therefore be constructed to evoke a favourable social perception. One major factor contributing to this perception is voice, and the sociophonetic cues and features

of this voice can play a decisive role when designing speech systems (Sutton et al. 2019). Another important characteristic of a synthetic voice is credibility. To test how our synthetic voice was perceived, we gathered evaluations based on the widespread Robotic Social Attributes Scale (RoSAS) (Bartneck et al. 2009, Carpinella et al. 2017) which includes characteristics in the field of ANTHROPOMORPHISM, LIKEABILITY, and (PERCEIVED) INTELLIGENCE. We excluded items from RoSAS which are not likely to be ascribed to an artificial voice but rather to a physical robot, namely ANIMACY (e.g., Dead-Alive) and PERCEIVED SAFETY (e.g., Anxious-Relaxed). While there could be other relevant social dimensions which are not investigated in the present study (for instance relating to the perceived confidence of the speaker), the use of a validated scale like RoSAS is necessary to ensure that the questionnaire is a reliable and valid tool.

All in all, this study addresses the following research questions:

1. How does a TTS speaking the users' own L2 English variety impact the learning outcomes of the said users?
2. How do users evaluate the L1 and L2 TTS voices in terms of ANTHROPOMORPHISM, INTELLIGENCE, and LIKEABILITY?

The study simulates a learning scenario involving a pre-test, a short lecture in Linguistics, evaluation of the TTS voice as well as the participants' performance on retention and transfer learning questions. The findings can be applied in the design of customizable PAs that can be adapted to the user's variety and thus facilitate learning of input produced in the language variety speakers are most familiar with – the variety spoken in their sociolinguistic environment.

2. Background

High-quality e-learning technologies are grounded in pedagogical theory (Dalsgaard 2005). The design of PAs is based on several thoroughly researched effects and principles such as the social agency, cognitive effort, personalization, and voice principles (Mayer 2014, Mayer et al. 2003b). Social agency theory argues for the importance of social cues in multimedia learning (Atkinson et al. 2005, Mayer et al. 2003b: 419). Cognitive effort theory suggests that students work harder in a cognitively demanding situation (Mayer et al. 2003b: 421). The personalization principle supports the use of a more conversational instead of a formal style (Mayer 2014: 348). The voice principle gives importance to the voice of the system, originally supporting the use of a human voice (Mayer 2014: 357) but with later studies providing more support for computer voices (Craig and Schroeder 2017, 2019). We will focus on the ways these principles inform the conceptualization of pedagogical assistants' synthetic voices.

2.1 Personalization principle

The personalization principle has often been realized in the use of personal pronouns, conversational style, and direct comments to the reader (Reichelt et al. 2014). Still, studies have shown that the personalization effect is not universally advantageous but is culture dependent. In contrast to American students, Czech students have been found to prefer a computerized tutor to address them with more direct and formal statements rather than polite and informal statements (Brom et al. 2017). In comparison, an 'everyday' style was beneficial for German (Reichelt et al. 2014: 207) and Turkish test takers (Kartal 2010). Therefore, decisions on the application of the personalization principle should be made with the cultural and social context in mind.

2.2 Voice principle

The agent's voice has been found to be more important than its physical presence on the screen (Mayer et al. 2003a: 811). In Mayer et al. (2003b), the accented voice resulted in the same retention outcomes as the nonaccented voice, but the nonaccented voice led to better transfer. In addition, the human recording outperformed the machine synthesis in retention, transfer, and subjective characteristics like dynamics and attractiveness (Mayer et al. 2003b: 422–423). Atkinson et al. (2005: 136) also found a voice effect in favour of the human voice as opposed to the machine voice. Do et al. (2022) found an interaction effect between the learner's gender and the synthetic speech accent on retention scores, as females performed worse under the unfamiliar accent (Indian English) condition compared to the familiar one (AmE). They also noted that females rated the human-likeness of the unfamiliar accent lower (Do et al. 2022). However, other recent studies showed that with improved voice technologies, the performance with machine voices has become similar to or better than human voices (Craig and Schroeder 2017). Craig and Schroeder argue that "the voice effect is a by-product of technological limitations rather than a binding, persistent limitation an instructional designer should be concerned with" (Craig and Schroeder 2019: 1537). Indeed, it has been reported that TTS can be perceived similarly likeable, trustworthy, and intelligible as a human (Abdulrahman and Richards 2022; Bione and Cardoso 2020). In terms of retention, Davis et al. (2019: 9) found no significant difference between human and machine voices. The study compared different prosodic variations of human voice and showed that a human weak prosodic voice (i.e., incorporating weak prosodic cues, such as having a flat intonation) was rated better than a machine voice in comparison to a human strong prosodic voice (Davis et al. 2019). Thus, linguistic considerations like the features to be included in the synthesis of the voice are essential.

The influence of linguistic cues and features on learning has been examined on different scales from small-scale perceptual differences in sounds to large-scale differences in rhetorical structure and pragmatic dimensions like politeness (e.g., Lin et al. 2020, Ylinen et al. 2010). Social information regarding dialect, age, and gender can influence the perception of sounds (Drager 2010) and accordingly

affect speech perception. Cristia et al. (2012: 4) point out that accented speech initially perturbs word recognition and sentence processing, yet with exposure, people can adapt to it. Thus, language speakers can adapt to any accent with sufficient exposure. However, it may be helpful to ease this adaptation.

The psycholinguistic insights on perceptual learning are interesting for us on two levels – the human vs computer dimension and the language variety dimension. Looking at the human vs computer dimension, in McAuliffe et al. (2016), perceptual learning of fricatives by L1 English speakers took place regardless of whether participants were exposed to resynthesized or naturally produced speech. It seems that synthesised language does not hinder perception. Still, it should be noted that this is a perceptual learning task applied to sounds and thereby not fully comparable with our learning scenario aiming to test the retention and transfer of knowledge. Regarding the language variety dimension, it has been shown that dialect information (African American English vs. Standard American English) is extracted pre-attentively and is used to determine whether the speaker belongs to the same or different group as the listener (Scharinger et al. 2011). If the cues of the TTS indicate a particular variety, this is likely to be rapidly perceived and categorised. It is also expected that the variety will be associated with attitudes towards it.

Language attitudes are built in a complex manner. Speakers can express more positive attitudes towards familiar varieties (Gill 1994), yet speakers can also prefer other varieties, e.g., when L2 speakers have more positive attitudes towards L1 speakers. In a survey of Czech English students by Brabcová and Skarnitzl (2018), most students wanted to acquire an L1 English accent like British English, though they believed that students should be exposed to different accents in English lessons.

Attitudes are not restricted to general preference, but can feature characteristics like content credibility, i.e., whether information delivered by a speaker of the variety can be trusted. For example, in a study evaluating an artificial tour guide speaking Standard Austrian German or a Viennese variety, the guide speaking the standard variety was classified as educated, trustworthy, competent, polite, and serious, whereas the Viennese guide was rated more natural, emotional, relaxed, open-minded, with the highest sense of humour, but also as more aggressive (Krenn et al. 2017: 75). Voice assistants are consistently attributed age and race traits based on their vocal cues and personality traits, which correlate with common stereotypes (Holliday 2023). Dahlbäck and colleagues (2001, 2007) found evidence for the similarity-attraction effect in the perception of information read in Swedish and American English, as listeners trusted information given by a speaker of their own variety regardless of whether it was trustworthy due to signs of higher expertise or not. Lev-Ari and Keysar's well-replicated 2010 study found that trivia statements produced by speakers of L2 varieties are less trusted than those produced by speakers of L1 varieties. Czech students also rated the L1 speakers more credible than the Czech L2 speakers of English (Hanzlíková and Skarnitzl 2017, Podlipský et al. 2016). A later study showed that with increased exposure, people who have had no contact to a L2 variety like Polish English can trust it even more than a L1 variety like British English (Boduch-Grabka and Lev-Ari 2021). Thus, the incorporation of L2 TTS in PAs may have promising outcomes for evoking trust.

2.3 Credibility and intentionality

Credibility has been viewed to contain several dimensions – "competence", "trustworthiness" and "goodwill" ("intent toward receiver") (McCroskey and Teven 1999: 90). The good intent when someone is credible evokes another concept – intentionality. According to Prinz (2017), intentionality indicates that "conscious experience always implies the experience of a particular kind of access to a particular kind of content" (Prinz 2017: 348); for example, hearing implies hearing *something*, such as hearing a sound. In the case of designing TTS for PAs, intentionality motivates the conscious integration of features that aim to evoke something, e.g., an impression of a credible agent. The use of an L2 variety in our study aims to give the PA some intentionality – it invokes pre-formed intentional representations in users' minds. As users perceive the PA's intent to improve their experience, they are more likely to attribute it positive characteristics, to trust it, and to want to work with it. Many L2 English speakers have learned English in a formal education environment such as a local school, so they are likely to have had most contact to their classmates and teachers, who are also likely to be speakers of their local English variety. Thus, the local variety is their familiar variety, so when it is incorporated in a PA learning environment, this may have a social mirroring effect (Prinz 2013). The agent mirrors its user and if they share a conceptual framework, this act is perceived by the user (Prinz 2013: 1107). The user attributes intentionality to the PA even though it does not have the prerequisites (i.e., a brain equivalent to the human brain) to develop intentionality (cf. Searle 1980). As a result, the user understands the intentionality of the agent in its attempt to be like the user and be liked by the user, which can lead to a positive perception on the side of the user.

Linguistic credibility becomes an expression of goodwill, a form of caring. As previous research has shown, the (human) teacher's care for the students and their good relationship impacts the teacher's credibility (Teven 2007: 435). So how can intended goodwill be attributed to a machine? There are of course many factors to consider, but from our linguistic perspective, it is important for the agent to mirror the sociolinguistic environment of the learners. The studies described above have repeatedly demonstrated the positive effect of familiarity. In the context of PAs, Maes (1994) has also argued that the agent should adapt to the user needs by imitating the user and implementing their feedback (Maes 1994: 40).

Intentional design means taking care of individual differences of the users. For example, for groups preferring directness, the text should not be made unnecessarily indirect, as this could lead to positive face threat (Brown and Levinson 1987). Also, when a L2 variety is mirrored, it is important to avoid face threat through linguistic stereotypes, i.e., language features that are "overt topics of social comment" (Labov 1994: 78), such as the lacking /r/-/l/ distinction in Chinese English. The TTS voice should be a credible representation of the variety and not a stereotype (Ivanova and Schmied 2023). Finally, integrating L2 varieties in educational products has the potential to improve the attitudes towards them and increase their acceptance, leading to diversification and challenging negative

stereotypes (Sutton et al. 2019: 10–11). Attributing L2 varieties credibility is again a form of goodwill and caring.

Overall, creating a credible agent with a credible synthetic voice that gives the impression of intentional goodwill for the user means making the agent customised to the needs of the user. The design has to be suitable for individualisation but should also be context-aware (Sutton et al. 2019: 6–7). It has become clear that more attention should be paid to the influence of socio-cultural variables when designing language technologies. We attempted to apply the principles discussed in this section to the design of a TTS system speaking L2 varieties and tested it in a lecture scenario. The following sections describe the implementation and discuss the results of the study.

3. Methodology

3.1 Participants

We determined the sample size for the present study based on Atkinson et al. (2005), who recruited 50 participants, 25 in the human voice group and 25 in the machine voice group. The study did not include effect sizes for the relevant measures, which makes an *a priori* power analysis for sample size estimation unfeasible. Compared to other related studies (e.g., Craig and Schroeder 2017, 2019) using about 50 participants per group, the sample size of 25 participants per group is small; therefore, the results of the present study should be taken as a preliminary exploration and require replication in a further study with a larger sample.

We recruited 54 Czech speakers of English at the Masaryk University Brno in the Czech Republic by distributing the survey via mailing lists. Of them, 28 were randomly assigned to the group listening to a lecture synthesised in AmE and 26 in CzE. The participants were not informed about the variety they are about to listen. Table 1 outlines the demographic characteristics of the participants. The students were mostly between 21–23 years old and mostly females (38 females, 16 males). They mostly came from the field of Humanities and were mostly in their Bachelor studies. Most students were in their first year.

Table 1. Demographic characteristics of the study participants

Age	N (%)	Field	N (%)	Degree	N (%)	Year	N (%)
18–20	9 (17%)	Humanities	40 (74%)	Bachelor	36 (67%)	First	37 (69%)
21–23	24 (44%)	MINT	2 (4%)	Master	17 (32%)	Second	9 (17%)
24–26	10 (19%)	Other	12 (22%)	Other	1 (2%)	Third	7 (13%)
> 26	11 (20%)					Fourth	1 (2%)

The students had an advanced level of English as they had studied English for 15.3 years on average (*SD* = 5.17). Figure 1 visualises the high exposure to English and communication in English of the participants. Moreover, many communicate in

English on social media and at university. Only six students had spent time in an English-speaking country, thus the level of English should be similar across the sample.

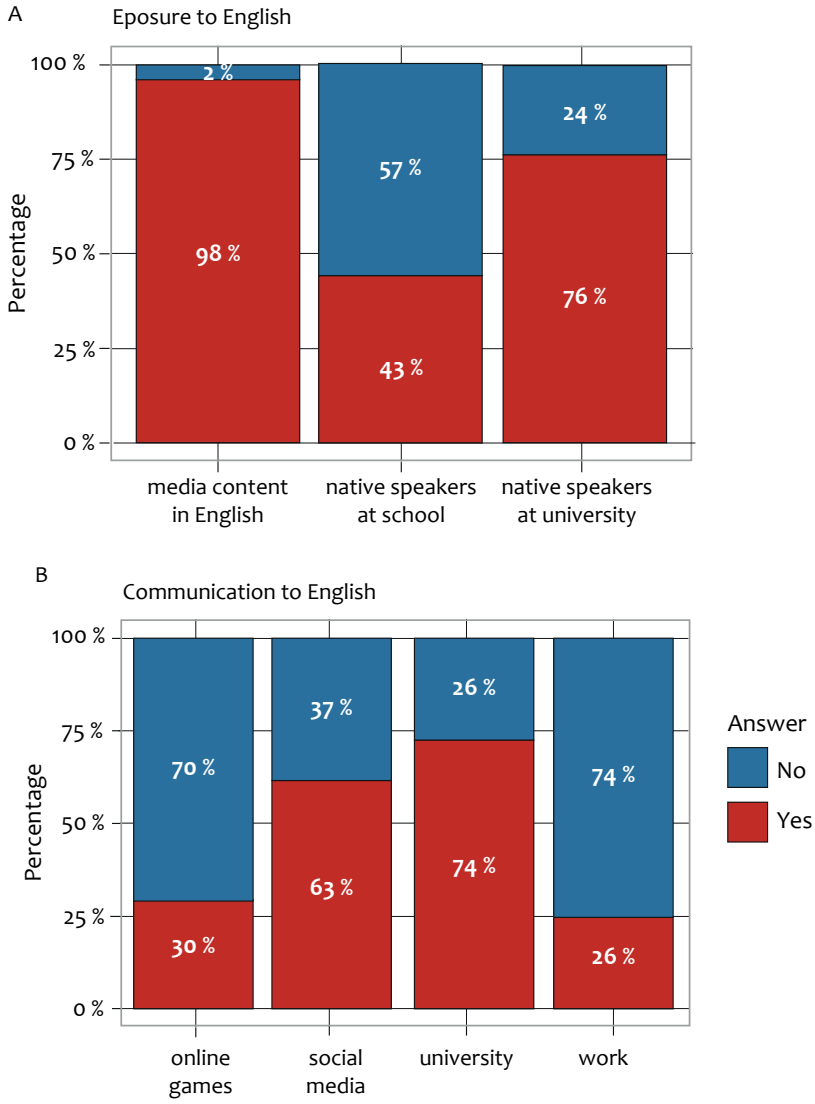

Figure 1. Situations of exposure to English and (B) places of communication in English of the Czech participants

The questionnaire also gathered data on technology affinity. 24% of the participants use virtual assistants, mostly on their phones and computers. Still, the majority of the students did not consider themselves to be technology proficient (Appendix Figure 8).

The students' background knowledge with regard to the topic of content mapping was tested in a pre-test with the open question "What do you know about the concept of content mapping in linguistics?". Prior knowledge is a variable difficult to control but essential for keeping a comparable cognitive load which would help avoid confounds and explain the learning outcomes and performance (Davis et al. 2019: 3).

3.2 Materials

The lecture incorporated an excerpt from a Linguistics article by Delibegović Džanić and Berberović (2021) "Lemons and Watermelons: Visual Advertising and Conceptual Blending" (see Appendix). The excerpt explains cognitive content mapping on the basis of a joke "If life gives you lemons, a simple surgery can give you melons" (Delibegović Džanić and Berberović 2021: 124). The text was chosen because, based on the authors' previous teaching experience, it has been shown to capture students' attention. While the body-shaming marketing slogan is admittedly sexist, it is provocative and encourages critical engagement with linguistic and social issues. Regarding the text's complexity, it is difficult to find appropriate measures because most of them are designed for L1 English texts, and are readability measures, which may not transfer to listening. Using the Python library *textstat* (Bansal and Aggarwal 2022), we compared different readability scores. However, the scores were relatively contradictory (Appendix Table 5), as EFLAW, a score usually used for second-language learners was 24.2, which is between 20.49 and 25.49 and therefore quite easy to understand, whereas many of the L1 readability scores placed the text in the 11th grade, which is relatively complex. Still, despite these discrepancies in automatic complexity measurements, a reading of the text shows that the lecture is complex due to its high-tier academic vocabulary and structure and even more so when it is only heard. Thus, the lecture was able to challenge the participants and prompt learning, which can be measured through their responses on the test.

For the synthesis of the text, we used the TTS system developed within the project "Credible Conversational Pedagogical Agents" part of the Collaborative Research Centre "Hybrid Societies". The text was transcribed and synthesised in Czech English with the tacotron-TZ-IPA-6000 model and American English with the tacotron-LJS-IPA-101500 model. The Czech English model was first trained on the American English data set for general speech synthesis capabilities. Then, it was trained on the Czech English dataset to learn the specific pronunciation features of the variety. The text was first synthesised to a Mel spectrogram using *tacotron* (Taubert 2022a), then the Mel spectrogram was synthesised to a WAVE audio file with *waveglow* (Taubert 2022b) and finally, the audio was denoised with *denoiser* (Défossez et al. 2022). The models were trained on a corpus of annotated sociolinguistic recordings with Czech speakers. The final CzE audio contained typical CzE features (see Skarnitzl and Rumlová 2019) such as /ɑ/-/o/ in *cosmetic* /kɑzˈmɛtɪk/ pronounced as /kozˈmɛtɪk/ as well as a lack vowel reduction, e.g., in *projected* /prəˈd͡ʒɛktəd/ pronounced as /prodˈ͡ʒɛktəd/. Since the recordings used in the training were of female university students in their 20s, the synthetic speaker

has a relatively young-sounding voice, which may impact its perceived credibility (Beege et al. 2017), but this is probably unlikely, since the lecture is not designed to evoke mistrust. Based on previous studies, we do not expect gender to affect learning outcomes (Castro-Alonso et al., 2021, p. 1002). Thus, we expect language variety to be the most influential variable.

The synthesised lecture lasted 2:30 min. The short length of the listening text was justified by factors such as attention span and memory capacity (Craig and Schroeder 2019: 1545). While there is some evidence for the benefits of simultaneously presenting audio and visual representation (Davis et al. 2019: 3), we incorporated only audio to avoid confounding multimodal effects that would be generated by complementary text, images, or embodied speaker, and to focus only on the voice's variety effects. As Davis et al. (2019: 3) point out, an advantage of presenting only one mode is that cognitive processing channels are not overloaded by simultaneous input.

The participants were asked to identify the origin of the speaker ("Where do you think the speaker comes from?") and were mostly able to identify the speaker of their own variety but had issues recognizing the AmE variety (Figure 2). Among the wrong responses, the AmE speaker was considered to come from Great Britain or the Czech Republic. Thus, some of the participants may have been expecting to hear a speaker related to their own environment, yet this cannot be directly concluded from our data. Future studies can ask the participants to justify their response. The wrong responses on the CzE speaker's origin were mostly broad classifications as eastern/central Europe. There were two answers indicating that this is a computer-generated voice.

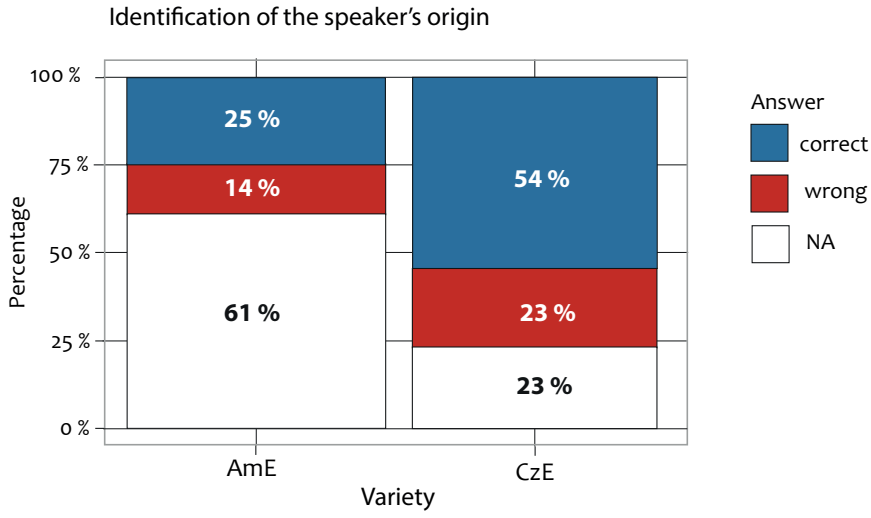

Figure 2. Percentage of correct, wrong, and lacking responses on the TTS speaker's origin

The cues in the AmE TTS voice allowing it to be classified as AmE were either not strong enough or the participants may not have had sufficient exposure to

the variety. While this may be considered a disadvantage for our TTS system, it is partially an advantage for the experiment, as we expect that the origin of the speaker has not biased the ratings of the characteristics in the AmE versions.

3.3 Procedure

The survey was conducted online on the LimeSurvey servers of the Chemnitz University of Technology. As the survey was distributed via email, the participants took it in their preferred setting and device. All questions were obligatory. The survey opened with information on the background and goals of the research and on data protection together with an informed consent form. Then the participants were randomly assigned to listen to the lecture produced either in CzE or in AmE. They were instructed to listen to the lecture and rate 13 statements on a five-point scale (see Figure 3, Table 2, and Appendix).

* Pelase rate your impression of the voice on thse scales

	1	2	3	4	5	
fake	○	○	○	○	○	natural
machinelika	○	○	○	○	○	humanlike
unconscious	○	○	○	○	○	conscious

Figure 3. Screenshot of the lecture part of the survey

Table 2. Overview of the rated TTS speaker characteristics

Group	Characteristic
ANTHROPOMORPHISM	[fake | natural]
	[machinelike | humanlike]
	[unconscious | conscious]
	[artificial | lifelike]
INTELLIGENCE	[incompetent | competent]
	[ignorant | knowledgeable]
	[unintelligent | intelligent]
	[foolish | sensible]
LIKEABILITY	[dislike | like]
	[unfriendly | friendly]
	[unkind | kind]
	[unpleasant | pleasant]
	[awful | nice]

In the test component, the participants responded to five general vocabulary questions, five academic vocabulary questions and three transfer learning questions (see Appendix). We aimed to combine different question types to test retention and transfer (see e.g., Buck 2007). The general vocabulary items were multiple-choice questions with one correct and two wrong answers, which inquired about high-tier vocabulary from the text such as "body contouring" and "astringent". These questions served to test and compare the English level of the participants. The second set of questions, the academic vocabulary type, were again multiple choice, but aimed to test the understanding of core linguistic concepts from the lecture such as "content mapping" and "input space". The final three questions on transfer learning included one short and one long open question and one multiple choice question, which tested the application of the knowledge acquired in the lecture. The participants had to explain jokes like the one used as example in the lecture (see Appendix), which prompts the application of the new concepts. The jokes presented images of a surgeon and a woman with the statements "If life gives you lemons, a simple operation/plastic surgery can give you lemons". The grading of the responses is explained in the following section.

After the test section, a demographic questionnaire was presented. It covered background data like age, field of study, gender, country of origin, and foreign languages as well as questions on tech-savviness and linguistic experience (see Appendix).

3.4 Pilot study

In order to test the experimental setting and identify potential issues with the study design, we carried out a pilot study with 74 Italian English (ItE) and 15 CzE speakers. The Italian students allowed us to test the experiment with our ItE TTS model in a different L2 English environment. The participants had a similar demographic profile like the present study: mostly between 21–23 years old (76%) and mostly females (70 females, 19 males). They mostly came from the field of Humanities (80% of the Czechs, 85% of the Italians) and were in their Master's. Almost all students were in their first year. The students had a similar level of English – about 15 years on average and high exposure to English. The students listened to the same lecture but in AmE or in their own variety (ItE or CzE) and answered the same questions. Yet, they rated the speaker on a different scale that was developed by the authors, hence only the learning outcomes results will be compared.

For the main study, the well-replicated RoSAS scale was used, which was validated in Bartneck et al. (2009) with Cronbach's alpha values around .89. The similarly high alpha values in Carpinella et al. (2017) support the scale's validity. The use of a validated scale in the present study thus improves on the pilot study in terms of validity and reliability.

3.5 Analysis

While the multiple-choice questions could be evaluated in a simple true-false manner, the pre-test and the open questions were rated on a scale from 0 to 5 (where 5 is the highest) based on the correctness of the responses and the grasp of the

information provided in the lecture. When grading the participants' explanations of the similar jokes, the first author paid attention to the comparison between the conceptual domains and the wordplay. For instance, an answer describing the metaphorical connection between the size of fruit and breasts receives a 5, while an answer just describing the fruit size receives a 4 (example answers in Appendix Table 6). A couple of responses only partly acknowledged the comparison and were given 3 points. Many responses explained the function of the jokes but not their composition (e.g., "It is funny, so you will remeber [sic] it better.") or simply re-phrased the joke's statement and were accordingly given 1 point. Refusal to give a response or answers which do not fit the question were given 0 points. This procedure allowed us to give credit to responses which have only partly grasped the ideas of the lecture. The responses from the test and the demographic data were analysed in R via RStudio (R Core Team 2022, RStudio Team 2021) with the *tidyverse* (Wickham et al. 2019) and *rstatix* packages as well as the core *stats* package. Chi square tests were applied to the categorical data of the correct and wrong responses of the multiple-choice questions. For the grades of the open questions and the rating of the voice's characteristics, after testing the assumptions for normality via Shapiro-Wilk tests and homogeneity of variances with Levene's test, the comparisons were conducted with the non-parametric Wilcoxon rank-sum test on unpaired samples (also known as the Mann-Whitney U-test) with Holm correction.

4. Results

a. Learning outcomes

In the pilot study, there was no difference between the multiple-choice score based on the variety (AmE or ItE/CzE): ([χ^2 (1, N = 979) = 1.04, p = .309]). There was also no difference between the grades of the open questions (W = 938, p = .959).

Henceforth the results of the main study will be discussed. The learning outcomes were evaluated based on the participants' pre-test and test results. Almost none of the participants knew anything about content mapping based on the pre-test (except for one participant with grade 3/5, one with 2/5, and two with 1/5). An overview of the proportion of correct and wrong responses per lecture variety is presented in Figure 4. In the multiple-choice question section (Figure 4A, 4B), there seems to be no difference between AmE and CzE [χ^2 (1, N = 594) = 0.0751, p = .784]. The general vocabulary, academic vocabulary, and transfer learning questions also have a similar correctness ratio in the two English varieties. Looking at the results from the open questions (Figure 4C), there is no difference between the grades of the AmE lecture (M = 2.2, SD = 1.75) and the CzE lecture (M = 3.13, SD = 1.53) [W = 258, p = .0668].

We tested whether other variables influence the test results. There was no difference between the responses of the students who use virtual assistants and those who do not [χ^2 (1, N = 594) = 0.323, p = .57]. The study degree also did not play a role as students from the Bachelor, Master and PhD level had a similar response ratio of about 3:2 correct vs wrong responses.

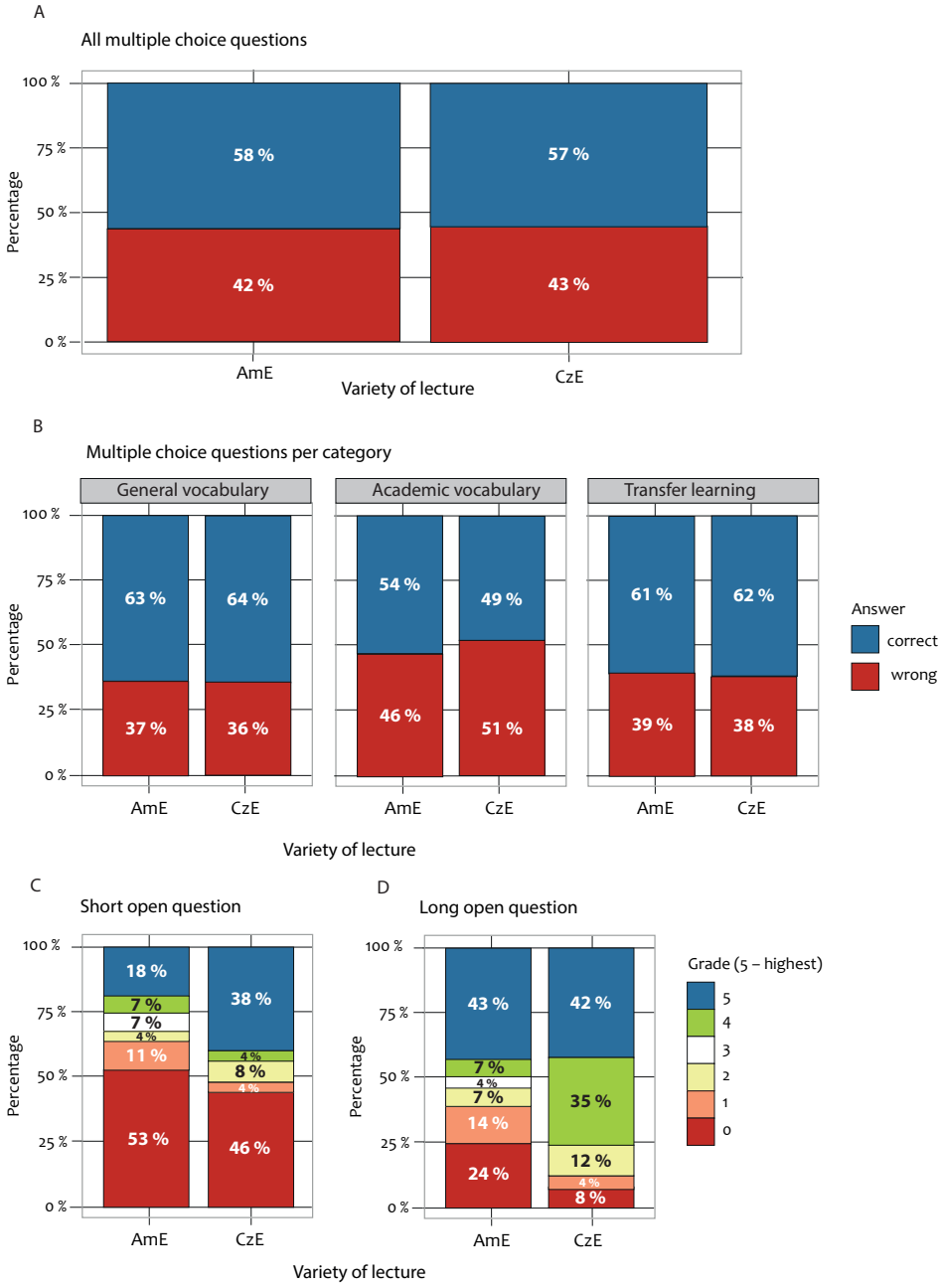

Figure 4. Answer correctness of the multiple choice and open questions per lecture variety

b. Characteristics

The Czech participants rated the CzE and AmE speaker characteristics similarly on average [$W = 436$, $p = .215$]. The mean rating ($M = 2.85$, $SD = 1.09$) is close to the middle of the five-point scale.

In terms of variety of the lecture, CzE and AmE were rated significantly differently in terms of INTELLIGENCE characteristics but similarly in terms of most ANTHROPOMORPHISM and LIKEABILITY characteristics, except for *pleasantness*. This is evident from the results of the pairwise Wilcoxon tests in Tables 3 and 4 and the boxplots with error bars in Figures 5, 6, and 7 below.

Table 3. Results of Wilcoxon rank-sum tests comparing characteristic ratings per variety summarised per group

Characteristic group	Var. 1	Var. 2	N1	N2	*W*	*p*	Sign.	Effect *r*	Magnitude
anthropomorphism	AmE	CzE	28	26	319.5	0.443	ns	0.11	small
intelligence	AmE	CzE	28	26	519.5	0.00682	**	0.37	moderate
likeability	AmE	CzE	28	26	465	0.0803	ns	0.24	small

Table 4. Results of Wilcoxon rank-sum tests comparing characteristic ratings per variety

Characteristic	Var. 1	Var. 2	N1	N2	*W*	*p*	Sign.	Effect *r*	Magnitude
competence	AmE	CzE	28	26	531.5	0.0024	**	0.41	moderate
consciousness	AmE	CzE	28	26	409	0.423	ns	0.11	small
friendliness	AmE	CzE	28	26	450	0.117	ns	0.21	small
human-likeness	AmE	CzE	28	26	281.5	0.136	ns	0.20	small
intelligence	AmE	CzE	28	26	501	0.011	*	0.35	moderate
kindness	AmE	CzE	28	26	416	0.338	ns	0.13	small
knowledgeability	AmE	CzE	28	26	477.5	0.0402	*	0.28	small
lifelikeness	AmE	CzE	28	26	296.5	0.223	ns	0.17	small
likeability	AmE	CzE	28	26	461	0.0791	ns	0.24	small
naturalness	AmE	CzE	28	26	331	0.55	ns	0.08	small
niceness	AmE	CzE	28	26	468	0.0572	ns	0.26	small
pleasantness	AmE	CzE	28	26	487	0.0245	*	0.31	moderate
sensibility	AmE	CzE	28	26	433	0.188	ns	0.18	small

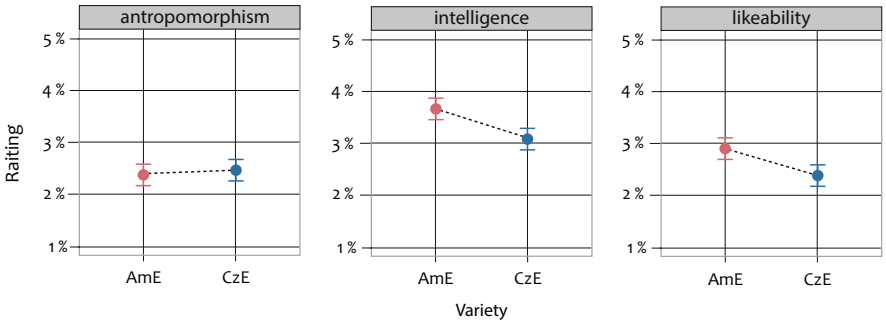

Figure 5. Voice characteristics groups rated by variety.
Error bars represent standard error

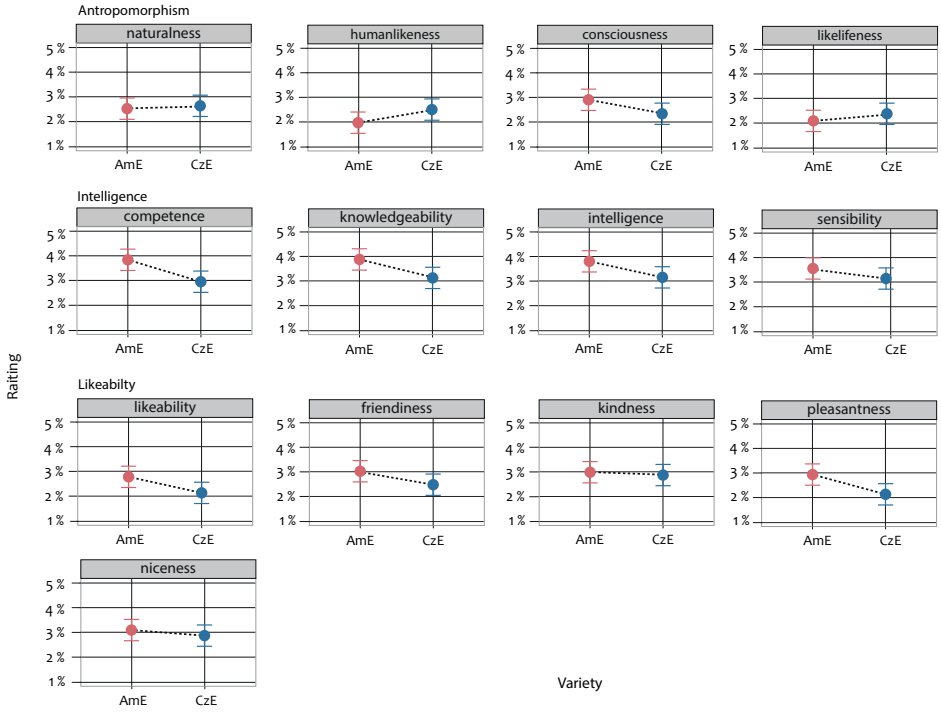

Figure 6. Voice characteristics rated by TTS variety.
Error bars represent standard error

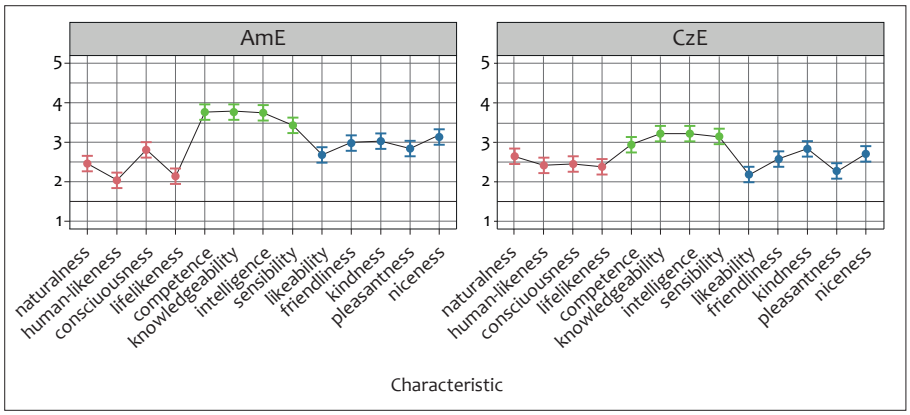

Figure 7. TTS varieties and voice characteristics ratings.
Error bars represent standard error

We tested whether other variables influence the characteristics ratings, such as the participant's gender, which was shown to affect evaluations of warmth and competence of language varieties in McKenzie et al. (2015). However, in our study, the ratings of the voice characteristics did not differ based on the participant's gender [AmE group: $W = 69.5$, $p = 0.853$; CzE group: $W = 77.5$, $p = 0.978$].

5. Discussion

This study examined the effect of the use of an L2 TTS system in teaching a lecture on content mapping in Linguistics. Although we expected to find improved learning outcomes after a lecture spoken in the L2 variety, we were not able to find evidence for the effectiveness of a TTS speaking English varieties. However, considering that participants started with no knowledge on the topic and nearly half of the multiple choice and open questions were answered correctly, we observe some development in the participant's knowledge on content mapping in Linguistics. Moreover, the study replicated the results of a pilot study with Italian and Czech English learners where the variety of the lecture also had no influence on the learning outcomes.

Generally, there might be a narrow window of learning material complexity (neither too easy, nor too difficult) for which improved learning outcomes are observed based on the voice effect. Our findings are similar to Mayer et al. (2003b), where student performance under an American-accented and Russian-accented voice did not differ significantly in terms of retention (Mayer et al. 2003b: 422). One reason may lie in the complexity of the lecture, as it may have been more suitable to be read than spoken. In future studies, the agent could be adapted to various oral presentation styles, for instance like a podcast. Speaking style is decisive for the perception of instructor credibility with many different influential variables like speaking rate and the use of humour (Myers and Martin

2018). As underscored in the review by Schneider et al. (2022), emotion factors like enthusiasm expressed through voice cues like intonation are also likely to benefit the learning process, yet this depends on the mental load during the learning task (Beege et al. 2020). There are thus many adjustable parameters in the lecture presentation that could affect the learning outcomes.

The lecture format as a whole can also be modified, as instructor credibility can be influenced by structural classroom components such as course format (Myers and Martin 2018: 41). In the process of lecture improvement, the biggest challenge would be to reduce confounds. Schroeder and Gotch (2015: 189) demonstrate that one of the biggest methodological issues of studies using PA is that they fail to isolate effects of the PA, where confounds affect the results on the level of the experiment treatment and control conditions and the design of the PA.

Another limitation of this study is that while the incorporation of only one mode (audio) was useful for reducing confounds, it reduced the ecological validity of the experiment, as learning with an audio-only narration is far from a real-life lecture scenario. Students can usually assist their comprehension through notetaking and visuals like presentations or handouts. However, while the addition of text seems intuitive, it could contradict the redundancy principle and become distracting. Moreover, even though the lecture was short, it could have been difficult to follow. This complexity was necessary to test the transfer learning effects on novel input and isolate background knowledge confounds. Managing complexity remains one of the core challenges of educational design and it opens promising opportunities for follow-up research.

When it comes to the evaluation of the speaker's voice, the participants rated the CzE and AmE voice characteristics similarly on average – around the centre of the rating scale. This may be the result of neutral response bias with some influences of social desirability bias.

Our computer voice appears to be consistently perceived as competent but rather unnatural in both AmE and CzE (Figure 7), thus it would be necessary to improve its prosody and thereby its naturalness and human-likeness (e.g., Ehret et al. 2021). The computer voice in other studies was also rated less dynamic and attractive (Mayer et al. 2003b: 423), so the overall lower rating of likeableness may be reflecting attitudes towards the computer voice. Craig and Schroeder have twice found that modern computer voices do not differ in terms of ascribed credibility compared to human voices, yet human voices are rated higher in terms of human-likeness and engagement (Craig and Schroeder 2017, 2019: 1542). Schroeder and colleagues have also shown that the more human-like the voice, the more consistent it is rated in terms of trust responses (Schroeder et al. 2021: 11). Still, trust in the virtual speaker had small influence on learning (Schroeder et al. 2021: 12). These findings may be in a way reflected in our results as well, as INTELLIGENCE variables like *competence* were rated higher than ANTHROPOMORPHISM variables like *human-likeness* and LIKEABILITY variables such as *friendliness*. Since our study compared two voices of the same agent, we cannot determine whether the reason for the ratings and learning outcomes lies in the voice quality of the agent. It would be interesting to compare our TTS to other systems and to a human voice. Also, while the perceived young age of the speaker most likely did not

affect its INTELLIGENCE perception, it would also be important to control for the age variable (see Beege et al. 2017).

The higher responses in favour of INTELLIGENCE in both varieties raise the question whether the ratings have been influenced by the academic language of the lecture. One solution would be to test students from a lower study degree with an easier text. K-12 students would be expected to have less exposure to English through media and education than in the participants in the present study. The underrepresentation of K-12 students in PA research has been highlighted (Schroeder et al. 2013: 20, Schroeder and Gotch 2015: 192), as there is some evidence that they profit from PAs more than post-secondary students (Schroeder et al. 2013). Future studies should test younger students, for instance in high school, who have less exposure to AmE compared to CzE and therefore could be expected to differ in terms of both learning outcomes and language attitudes towards the tested L1 and L2 varieties. However, considering that high school students in the Czech Republic mostly study subjects like Biology in Czech and not in English, the application of the findings on the CzE teacher voice would be limited.

Our findings on the preference of American English are similar to Mayer et al. (2003b), where the American English speaker was also rated more positively than the Russian-accented speaker. In the present study, we showed that the Czechs also covertly attribute more competence to AmE. Czechs' positive attitudes and trust towards L1 varieties have also been demonstrated in Brabcová and Skarnitzl (2018), Hanzlíková and Skarnitzl (2017) and Podlipský et al. (2016). However, in contrast to Ahn and Moore (2011), the more negative attitudes towards CzE in the present study did not worsen the participants' learning outcomes. As discussed in Reichelt et al. (2014: 207), the positive effect of personalization can be modulated by the content, learning phase duration, cognitive load, and some individual variables like previous knowledge. In our case, the attitude towards English varieties has probably affected the users' evaluation of the voice but it is not necessarily obvious whether it has influenced the learning outcomes. If the preference towards AmE were beneficial, we would have observed better learning outcomes in that variety, yet the test results were not found to differ based on the lecture's variety.

While in our study cognitive load was modulated implicitly by increasing the difficulty of the tasks, psychophysiological methods like electrocardiography (ECG) or a combination of methods like ECG and EEG (Jimenez-Molina et al. 2018) can provide additional insights on the cognitive load during different language varieties and test tasks.

Overall, the results of the present study can inform the design of personalized PAs. A possible scenario incorporating such PAs could feature Czech university students who speak Czech English as L2 and study a subject like Informatics in English abroad, e.g., in Germany. Instructors can synthesize their lectures in CzE or AmE and thus foster the students' learning outcomes by giving them input in the students' own L2 variety or a widespread L1 variety. Still, future studies are necessary to test the effect of other L1 and L2 varieties on CzE speakers' learning outcomes in such scenarios.

6. Conclusion

Technology is situated in a sociolinguistic environment and needs to be adapted to it. Every speaker has an accent, so designers of speech technologies always face the choice of which model they want to incorporate. While so far the explored varieties have predominantly been L1 English varieties like American English, we tested whether a TTS speaking the user's own L2 variety can improve learning outcomes in an English-medium environment, in accordance with the personalization, voice, credibility, and intentionality principles. Experiments with Czech English speakers showed that the variety of the agent did not significantly improve the learning outcomes. These findings replicate our results from a pilot study with Italian and Czech English learners, demonstrating the reliability of our results. In terms of subjective evaluation of the speaker, we found different evaluations of characteristics related to the ANTHROPOMORPHISM, INTELLIGENCE and LIKEABILITY of the synthetic speaker. The speaker was consistently rated more competent than human-like or likable. Moreover, AmE was rated more competent in relation CzE, indicating more positive attitudes towards L1 varieties.

Thus, our original intention of measuring to what extent accent accommodation towards students' own variety (a step towards "nativisation") was perceived as helpful for learning was difficult, as our respondents were far from the *lingua franca* ideal of accepting (their own) pronunciation differences (instead of rejecting them as deficiencies). This preference for L1 accents in teaching by Czech students, reflects traditional, positive attitudes towards native speakers in Eastern Europe, which are confirmed in other studies from the Czech Republic (Brabcová and Skarnitzl 2018)and Poland: Although over 71% of Polish language school students (and all of the teachers) agreed to the statement "It Is Acceptable to Speak English With a Foreign Accent" (Kiczkowiak 2018: 152), they agreed even more (78% and only half of the teachers) to "Students Should Try to Reduce Their L1 Accent When Speaking English" (ibid: 153). It is also interesting that in this Polish sample the quality that a teacher "Speaks English as their mother tongue OR "Is a Native English Speaker" is much more important to students (around 70%; ibid: 162) than to teachers (around 20%; ibid: 166). Thus, English is indeed seen as "foreign" and not "accommodated" or "nativized" by learners, even in university contexts.

With improved TTS model prosody and a larger sample, future studies can test the influence of variety familiarity on learning. By incorporating intentionally credible voices, language technology developers can address the individual needs of learners from different backgrounds and thereby give them active control over the product and the learning process. These efforts have the potential to ease the learning of L2 English speakers with an English-medium instruction by adapting to their individual learning in English.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Olga Dontcheva-Navratilova for the dissemination of the surveys at the Masaryk University Brno and the discussion of the results. We are also grateful for all students who participated in the survey. We thank the team of the D03 subproject of the Collaborative Research Center "Hybrid Societies" at the Chemnitz University of Technology for their contributions in the development of the TTS system and their inputs on the learning experiment. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410, TP D03. This work was also supported by a PhD scholarship awarded to MI by the State of Saxony.

Data availability

The analysis code is available in a publicly accessible repository: https://osf.io/2xegt/. The survey data will be made available by the authors upon request, without undue reservation.

References

Abdulrahman, Amal and Richards, Deborah (2022) Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction* 6 (7), 51. https://doi.org/10.3390/mti6070051

Ahn, Jeahyeon and Moore, David (2011) The relationship between students' accent perception and accented voice instructions and its effect on students' achievement in an interactive multimedia environment. *Journal of Educational Multimedia and Hypermedia* 20 (4), 319–335.

Andrist, Sean, Ziadee, Micheline, Boukaram, Halim, Mutlu, Bilge and Sakr, Majd (2015) Effects of culture on the credibility of robot speech: A comparison between English and Arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. https://doi.org/10.1145/2696454.2696464

Atkinson, Robert K., Mayer, Richard E. and Merrill, Mary 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

Bansal, Shivam and Aggarwal, Chaitanya (2022) *textstat* (Version 0.7.3) [Computer software]. https://pypi.org/project/textstat/

Bartneck, Christoph, Kulić, Dana, Croft, Elizabeth and Zoghbi, Susana (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1 (1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Beege, Maik, Schneider, Sascha, Nebel, Steve, Mittangk, Jessica and Rey, Günter Daniel (2017) Ageism – Age coherence within learning material fosters learning. *Computers in Human Behavior*, *75*, 510–519. https://doi.org/10.1016/j.chb.2017.05.042

Beege, Maik, Schneider, Sascha, Nebel, Steve, and Rey, Günter Daniel (2020) Does the effect of enthusiasm in a pedagogical Agent's voice depend on mental load in the Learner's working memory? *Computers in Human Behavior*, *112*, 106483. https://doi.org/10.1016/j.chb.2020.106483

Bent, Tessa and Bradlow, Ann R. (2003) The interlanguage speech intelligibility benefit. *Journal of the Acoustic Society of America* 114 (3), 1600–1610. https://doi.org/10.1121/1.1603234

Bione, Tiago and Cardoso, Walcir (2020) Synthetic voices in the foreign language context. *Language Learning and Technology* 24 (1), 169–186. https://doi.org/10125/44715

Boduch-Grabka, Katarzyna and Lev-Ari, Shiri (2021) Exposing individuals to foreign accent increases their trust in what nonnative speakers say. *Cognitive Science* 45 (11), e13064. https://doi.org/10.1111/cogs.13064

Brabcová, Kateřina and Skarnitzl, Radek (2018) Foreign or native-like? The attitudes of Czech EFL learners towards accents of English and their use as pronunciation models. *Studie Z Aplikované Lingvistiky* 1 (38-50).

Brom, Cyril, Hannemann, Tereza, Stárková, Tereza, Bromová, Edita and Děchtěrenko, Filip (2017) The role of cultural background in the personalization principle: Five experiments with Czech learners. *Computers & Education*, *112*, 37–68. https://doi.org/10.1016/j.compedu.2017.01.001

Brown, Penelope and Levinson, Stephen C. (1987) *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Buck, Gary (2007) *Assessing listening* (7. print). *The Cambridge language assessment series*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Carpinella, Colleen M., Wyman, Alisa B., Perez, Michael A. and Stroessner, Steven J. (2017, March). The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction* (pp. 254–262).

Castro-Alonso, Juan C., Wong, Rachel M., Adesope, Olusola O. and Paas, Fred (2021) Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review* 33 (3), 989–1015. https://doi.org/10.1007/s10648-020-09587-1

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe (2020) *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

Craig, Scotty D. and Schroeder, Noah L. (2017) Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, *114*, 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

Craig, Scotty D. and Schroeder, Noah L. (2019) Text-to-Speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research* 57 (6), 1534–1548. https://doi.org/10.1177/0735633118802877

Cristia, Alejandrina, Seidl, Amanda, Vaughn, Charlotte, Schmale, Rachel, Bradlow, Ann and Floccia, Caroline (2012) Linguistic processing of accented speech across the lifespan. *Frontiers in Psychology*, *3*, Article 479, 1–15. https://doi.org/10.3389/fpsyg.2012.00479

Dahlbäck, Nils, Swamy, Seema, Nass, Clifford, Arvidsson, Fredrik and Skågeby, Jörgen (2001) Spoken interaction with computers in a native or non-native language - same or different? In *Proceedings of INTERACT*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=58a34aa5fd752e29e434cbf9e9d72891715a2d8e

Dahlbäck, Nils, Wang, QuanYing, Nass, Clifford and Alwin, Jenny (2007) Similarity is more important than expertise: Accent effects in speech interfaces. In *CHI 2007 Proceedings,* San Jose, CA, USA. https://doi.org/10.1145/1240624.1240859

Dai, Laduona, Jung, Merel M., Postma, Marie and Louwerse, Max M. (2022) A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. *Computers & Education*, *190*, Article 104607, 1–28. https://doi.org/10.1016/j.compedu.2022.104607

Dalsgaard, Christian (2005) Pedagogical quality in e-learning. *Eleed*(1). https://www.eleed.de/archive/1/78

Davis, Robert O., Vincent, Joseph and Park, Taejung (2019) Reconsidering the voice principle with non-native language speakers. *Computers & Education*, *140*, Article 103605, 1–12. https://doi.org/10.1016/j.compedu.2019.103605

Défossez, Alexander, Synnaeve, Gabriel and Adi, Yossi (2022) *denoiser* (Version 0.1.4) [Computer software]. facebookresearch. https://github.com/facebookresearch/denoiser

Delibegović Džanić, Nihada and Berberović, Sanja (2021) Lemons and watermelons: Visual advertising and conceptual blending. In Larissa D'Angelo, Anna Mauranen and Stefania Maci (Eds.), *Metadiscourse in digital communication* (pp. 115–132) Springer. https://doi.org/10.1007/978-3-030-85814-8_6

Do, Tiffany D., Akter, Mamtaj, Choudhary, Zubin, Azevedo, Roger and McMahan, Ryan P. (2022) The effects of an embodied Pedagogical Agent's synthetic speech accent on learning outcomes. In *International Conference on Multimodal Interaction* (pp. 198–206). ACM. https://doi.org/10.1145/3536221.3556587

Dokovova, Marie, Scobbie, James M. and Lickley, Robin (2022) Matched-accent processing: Bulgarian-English bilinguals do not have a processing advantage with Bulgarian-accented English over native English speech. *Laboratory Phonology* 13 (1), Article 12, 1–40. https://doi.org/10.16995/labphon.6423

Drager, Katie (2010) Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4 (7), 473–480. https://doi.org/10.1111/j.1749-818X.2010.00210.x

Ehret, Jonathan, Bönsch, Andrea, Aspöck, Lukas, Röhr, Christine T., Baumann, Stefan, Grice, Martine, Fels, Janina and Kuhlen, Torsten W. (2021) Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception* 18 (4), 1–15. https://doi.org/10.1145/3486580

Fishero, Sheyenne, Sereno, Joan A. and Jongman, Allard (2023) Perception and production of Mandarin-Accented English: The effect of degree of Accentedness on the Interlanguage Speech Intelligibility Benefit for Listeners (ISIB-L) and Talkers (ISIB-T). *Journal of Phonetics*, *99*, Article 101255. https://doi.org/10.1016/j.wocn.2023.101255

Gill, Mary M. (1994) Accent and stereotypes: Their effect on perceptions of teachers and lecture comprehension. *Journal of Applied Communication Research*, *22*, 348–361.

Gosselin, Leah, Martin, Clara D., Martín, Ana González and Caffarra, Sendy (2022) When a nonnative accent lets you spot all the errors: Examining the syntactic interlanguage benefit. *Journal of Cognitive Neuroscience* 34 (9), 1650–1669 https://doi.org/10.1162/jocn_a_01886

Hanzlíková, Dagmar and Skarnitzl, Radek (2017) Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language* 15 (3), 285–298. https://doi.org/10.1515/rela-2017-0016

Hayes-Harb, Rachel, Smith, Bruce L., Bent, Tessa and Bradlow, Ann R. (2008) The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics* 36 (4), 664–679. https://doi.org/10.1016/j.wocn.2008.04.002

Holliday, Nicole (2023) Siri, you've changed! acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, 8. https://doi.org/10.3389/fcomm.2023.1116955

Ivanova, Marina and Schmied, Josef (2023) From cues to features: Bridging psycho- and sociolinguistics in the development of non-native English stimuli. *TESOL Communications* 2 (2), 1–17. https://doi.org/10.58304/tc.20230201

Jimenez-Molina, Angel, Retamal, Cristian and Lira, Hernan (2018) Using psychophysiological sensors to assess mental workload during web browsing. *Sensors* 18 (2). https://doi.org/10.3390/s18020458

Karakaş, Ali (2017) English voices in 'Text-to-speech tools': Representation of English users and their varieties from a World Englishes perspective. *Advances in Language and Literary Studies* 8 (5), 108. https://doi.org/10.7575/aiac.alls.v.8n.5p.108

Kartal, Günizi (2010) Does language matter in multimedia learning? Personalization principle revisited. *Journal of Educational Psychology* 102 (3), 615–624. https://doi.org/10.1037/a0019345

Kiczkowiak, Marek (2018) *Native Speakerism in English Language Teaching: Voices from Poland.* PhD thesis, University of York. https://etheses.whiterose.ac.uk/id/eprint/20985/

Krenn, Brigitte, Schreitter, Stephanie and Neubarth, Friedrich (2017) Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI & SOCIETY* 32 (1), 65–77. https://doi.org/10.1007/s00146-014-0569-0

Labov, William (1994) *Principles of linguistic change: Internal factors*. Oxford: Blackwell.

Lev-Ari, Shiri and Keysar, Boaz (2010) Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology* 46 (6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

Lin, Lijia, Ginns, Paul, Wang, Tianhui and Zhang, Peilin (2020) Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education*, *143*, Article 103658, 1–11. https://doi.org/10.1016/j.compedu.2019.103658

Louwerse, Max M., Graesser, Arthur C., McNamara, Danielle S. and Lu, Shulan (2009) Embodied conversational agents as conversational partners. *Applied Cognitive Psychology* 23 (1244-1255).

Maes, Pattie (1994) Agents that reduce work and information overload. *Communications of the ACM* 37 (7), 31–40. https://doi.org/10.1007/SpringerReference_85143

Major, Roy C., Fitzmaurice, Susan F., Bunta, Ferenc and Balasubramanian, Chandrika (2002) The Effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly* 36 (2), 173–190.

Mayer, Richard E. (2014) Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 345–368). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.017

Mayer, Richard E., Dow, Gayle T. and Mayer, Sarah (2003a) Multimedia learning in an interactive self-explaining environment: What Works in the Design of Agent-Based Microworlds? *Journal of Educational Psychology*, 95 (4), 806–812. https://doi.org/10.1037/0022-0663.95.4.806

Mayer, Richard E., Sobko, Kristina and Mautone, Patricia D. (2003b) Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology* 95 (2), 419–425. https://doi.org/10.1037/0022-0663.95.2.419

McAuliffe, Michael, Babel, Molly and Vaughn, Charlotte (2016) Do listeners learn better from natural speech? In *Interspeech,* San Francisco, USA.

McCroskey, James C. and Teven, Jason J. (1999) Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* 66 (1), 90–103.

McKenzie, Robert M., Kitikanan, Patchanok and Boriboon, Phaisit (2016) The competence and warmth of Thai students' attitudes towards varieties of English: The effect of gender and perceptions of L1 diversity. *Journal of Multilingual and Multicultural Development* 37 (6), 536–550. https://doi.org/10.1080/01434632.2015.1083573

Myers, Scott A. and Martin, Matthew M. (2018) Instructor credibility. In Marian L. Houser and Angela Hosek (Eds.), *Handbook of instructional communication: Rhetorical and relational perspectives* (pp. 38–50). Routledge.

Nass, Clifford and Lee, Kwan M. (2001) Does computer-synthesised speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7 (3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171

Podlipský, Václav J., Šimáčková, Šárka and Petráž, David (2016) Is there an interlanguage speech credibility benefit? *Topics in Linguistics* 17 (1), 30–44. https://doi.org/10.1515/topling-2016-0003

Prinz, Wolfgang (2013) Self in the mirror. *Consciousness and Cognition* 22 (3), 1105–1113. https://doi.org/10.1016/j.concog.2013.01.007

Prinz, Wolfgang (2017) Modeling self on others: An import theory of subjectivity and selfhood. *Consciousness and Cognition*, *49*, 347–362. https://doi.org/10.1016/j.concog.2017.01.020

R Core Team. (2022) *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reichelt, Maria, Kämmerer, Frauke, Niegemann, Helmut M. and Zander, Steffi (2014) Talk to me personally: Personalization of language style in computer-based learning. *Computers in Human Behavior*, *35*, 199–210. https://doi.org/10.1016/j.chb.2014.03.005

Rey, Günter D. and Steib, Nadine (2013) The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior* 29 (5), 2022–2028. https://doi.org/10.1016/j.chb.2013.04.003

RStudio Team. (2021) *RStudio: Integrated Development for R* (Version 2021.09.0) [Computer software]. http://www.rstudio.com/

Sandygulova, Anara and O'Hare, Gregory (2015) Children's perception of synthesised voice: Robot's gender, age and accent. *Social Robotics. 7th International Conference, ICSR 2015*. Springer International Publishing. https://doi.org/10.1007/978-3-319-25554-5_59

Scharinger, Mathias, Monahan, Philip J. and Idsardi, William J. (2011) You had me at "Hello": Rapid extraction of dialect information from spoken words. *NeuroImage* 56 (4), 2329–2338. https://doi.org/10.1016/j.neuroimage.2011.04.007

Schneider, Sascha, Beege, Maik, Nebel, Steve, and Rey, Günter Daniel (2022) Psychologische Befunde zum Lernen mit digitalen Medien – ein Überblick [Psychological findings on learning with digital media - an overview]. In Mario A. Pfannstiel & Peter F.-J. Steinhoff (Eds.), *E-Learning im digitalen Zeitalter* (pp. 581–605). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36113-6_28

Schroeder, Noah L., Chiou, Erin K. and Craig, Scotty D. (2021) Trust influences perceptions of virtual humans, but not necessarily learning. *Computers & Education*, *160*, 1–15. https://doi.org/10.1016/j.compedu.2020.104039

Schroeder, Noah L. and Gotch, Chad M. (2015) Persisting Issues in Pedagogical Agent Research. *Journal of Educational Computing Research* 53 (2), 183–204. https://doi.org/10.1177/0735633115597625

Searle, John R. (1980) Minds, brains, and programs. *The Behavioral and Brain Sciences*, *3*, 417–457. https://doi.org/10.7551/mitpress/3080.003.0009

Skarnitzl, Radek, & Rumlová, Jana (2019) Phonetic aspects of strongly-accented Czech speakers of English. *AUC PHILOLOGICA*, *2*, 109–128. https://doi.org/10.14712/24646830.2019.21

Sutton, Selina J., Foulkes, Paul, Kirk, David and Lawson, Shaun (2019) Voice as a design material: Sociophonetic inspired design strategies in Human-Computer Interaction. In *CHI 2019,* Glasgow, Scotland, UK. https://doi.org/10.1145/3290605.3300833

Tamagawa, Rie, Watson, Catherine I., Kuo, I. Han, MacDonald, Bruce A. and Broadbent, Elizabeth (2011) The effects of synthesised voice accents on user perceptions of robots. *International Journal of Social Robotics* 3 (3), 253–262. https://doi.org/10.1007/s12369-011-0100-4

Taubert, Stefan (2022a) *tacotron-cli* (Version 0.0.3) [Computer software]. https://github.com/stefantaubert/tacotron

Taubert, Stefan (2022b) *waveglow-cli* (Version 0.0.1) [Computer software]. https://github.com/stefantaubert/waveglow

Teven, Jason J. (2007) Teacher caring and classroom behavior: Relationships with student affect and perceptions of teacher competence and trustworthiness. *Communication Quarterly* 55 (4), 433–450. https://doi.org/10.1080/01463370701658077

Wickham, Hadley, Averick, Mara, Bryan, Jennifer, Chang, Winston, McGowan, Lucy D., François, Romain, Grolemund, Garrett, Hayes, Alex, Henry, Lionel, Hester, Jim,

Kuhn, Max, Pedersen, Thomas L., Miller, Evan, Bache, Stephan M., Müller, Kirill, Ooms, Jeroen, Robinson, David, Seidel, Dana P., Spinu, Vitalie, . . . Yutani, Hiroaki (2019) Welcome to the tidyverse. *Journal of Open Source Software* 4 (43), 1686. https://doi.org/10.21105/joss.01686

Ylinen, Sari, Uther, Maria, Latvala, Antti, Vepsäläinen, Sara, Iverson, Paul, Akahane-Yamada, Reiko and Näätänen, Risto (2010) Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience* 22 (6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Marina Beccard (formerly Ivanova) is a researcher in English Linguistics at Chemnitz University of Technology, working in the areas of foreign language perception and production, corpus linguistics, and academic writing. Her PhD project used evidence from electroencephalography (EEG) experiments and phonetic analyses to investigate the perception and production of word stress cues by speakers of Slavonic Englishes, such as Czech English. She is also interested in the incorporation of language variation in learning through technology.

Address: Marina Beccard, English and Digital Linguistics, Chemnitz University of Technology, 09107 Chemnitz, Germany. [email: marina.beccard@phil.tu-chemnitz.de]

Sven Albrecht is a researcher in English Linguistics at Chemnitz University of Technology, with research interests in variationist sociolinguistics, World English's (with a special interest in phonetic variation), corpus linguistics, and academic writing. His PhD project focused on features of Chinese English and used evidence from phonetic analyses to investigate the production of vowels by speakers of Chinese English. He is not only interested in language produced by humans but also by digital agents. He is currently working at Mittweida University of Applied Sciences.

Address: Sven Albrecht, English Language and Linguistics, Chemnitz University of Technology, 09107 Chemnitz, Germany. [email: sven.albrecht@phil.tu-chemnitz.de]

Josef Schmied was the Chair of English Language & Linguistics at Chemnitz University of Technology from 1993 to 2021. His main research interests are in Language & Culture (sociolinguistics, English in Africa and SE Asia, Academic English) and in Language & Computers (corpus-linguistics, e-learning, large language models). His current research projects focus on the synthesis and perception of artificial non-native accents, the use of internet data in linguistic analysis, innovation in remote online learning, disciplinary conventions of academic writing, and national and subnational variation of Englishes in Africa and China. He enjoys the academic discourse with his PhD students and (Alexander-von-Humboldt) guest professors and his guest professorships in Italy and China.

Address: Prof. Dr. Josef Schmied, Emeritus, English Language & Linguistics, Chemnitz, University of Technology, 09107 Chemnitz, Germany. [email: josef.schmied@phil.tu-chemnitz.de]