

Listener Perceptions of Accented Synthetic Speech: Analyzing the Impact of L1

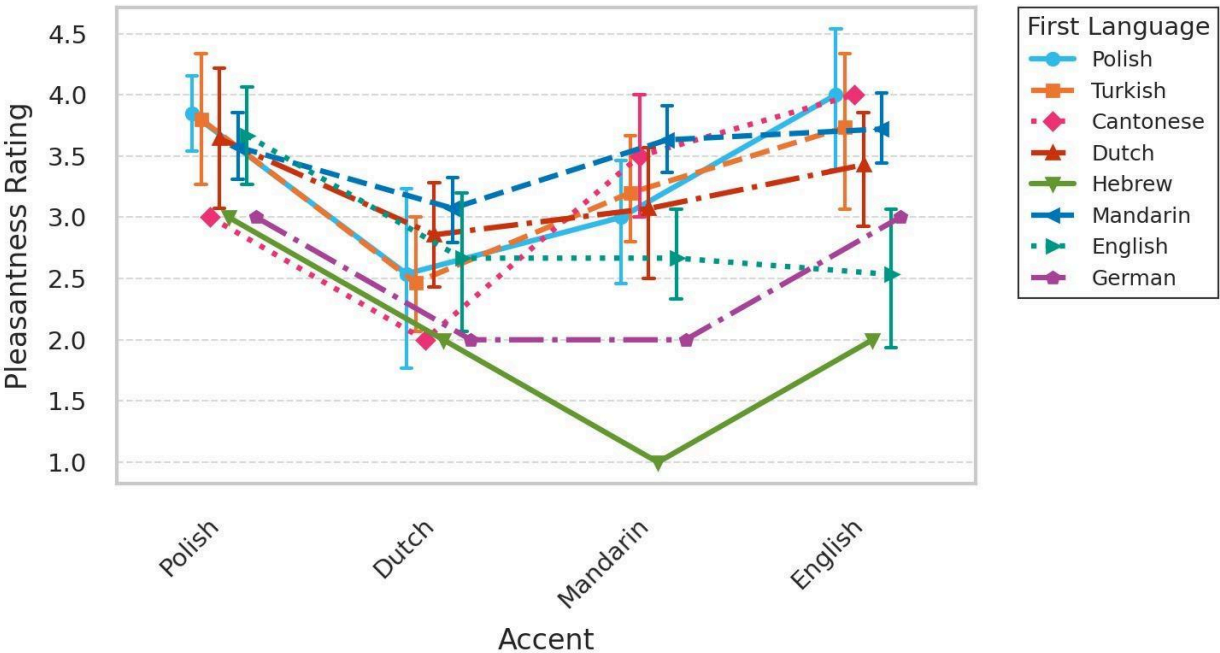
Jan Kokowski, Tiantian Zhang, Stella Siu, Vass Verkhodanova
University of Groningen

This study investigates how listeners' first language (L1) influences perceptions of pleasantness and trustworthiness in accented synthetic English speech. While research on Text-to-Speech (TTS) systems often prioritizes intelligibility and naturalness, user acceptance and emotional response — such as perceived trustworthiness and pleasantness — play a crucial role in applications like virtual assistants and customer service bots (Nordheim, 2018). Drawing on Miao (2024), our study explores whether accent familiarity affects these perceptions among listeners from diverse linguistic backgrounds.

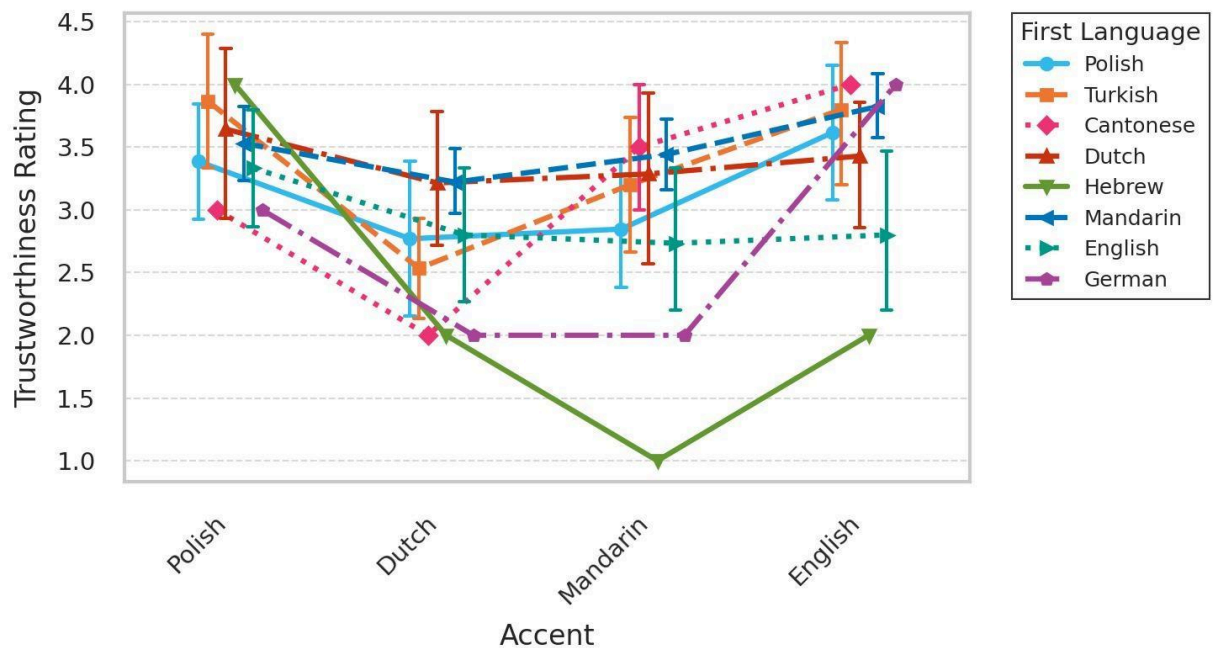
We generated synthetic speech stimuli using advanced AI-based TTS technology provided by Elevenlabs, producing voices in four distinct accents: Standard American English (native), Mandarin Chinese, Polish, and Dutch (Flemish). A total of 129 participants with eight L1 backgrounds rated these voices across standardized prompts using a 5-point Likert scale. Data were analyzed using Aligned Rank Transform (Wobbrock et al., 2011) and ANOVA to rationalize interactions between L1 and accent on pleasantness and trustworthiness ratings, with pairwise post-hoc tests identifying significant differences.

The findings indicate that listeners' L1 significantly influenced their ratings of pleasantness and trustworthiness, although the post-hoc test showed no significance for any specific accent, indicating that no single accent was universally preferred. Notably, native English-speaking participants rated the Standard American English voice lower than expected, likely due to an "uncanny valley" effect resulting from its lack of recognizable regional features. In contrast, non-native listeners rated the same voice positively, potentially associating it with perceived fluency and prestige.

Interaction Effect: L1 Background and Accent on Pleasantness Ratings



Interaction Effect: L1 Background and Accent on Trustworthiness Ratings



The results highlight the role of accent familiarity in shaping perceptions of synthetic speech, demonstrating the need for TTS systems to authentically represent regional linguistic identities. Designing voices that align with listener expectations may improve acceptance and inclusivity in speech-driven technologies. Further research could examine additional factors influencing listener perceptions, such as English proficiency levels and the ability to recognize specific accents.

References

Miao, Y. (2024). Factors Affecting Listener Perception of Accented Speech: The Role of Accent Familiarity and Linguistic Training. *International Journal of Listening*, 38(3), 203–215. <https://doi.org/10.1080/10904018.2023.2252019>

Nordheim, C. B. (2018). *Trust in chatbots for customer service – findings from a questionnaire study*. University of Oslo.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 143-146).

Listener Perceptions of Accented Synthetic Speech: Analyzing the Impact of L1

1 Introduction

While much of the focus in TTS evaluation has been on how natural or clear these voices sound, the way listeners perceive their pleasantness and trustworthiness is equally important for diverse applications (e.g. customer service bots or voice assistants) and something that many TTS systems designers aim to achieve. The perception of pleasantness and trustworthiness of a given synthetic voice can vary widely depending on a listener's linguistic background and familiarity with the accent being used.

In this study, we set out to explore how a listener's first language (L1) influences their perception of pleasantness and trustworthiness in accented-English synthetic speech. Specifically, we wanted to understand whether familiarity with an accent makes a difference and how this plays out for listeners from different linguistic backgrounds. To tackle this question, we created speech stimuli using AI voice synthesis technology. The voices represented four accents: Standard American English (native), Mandarin Chinese, Polish, and Dutch (Flemish). A total of 129 participants, speaking eight different native languages, rated these voices on a 5-point Likert scale for pleasantness and trustworthiness. Using Aligned Rank Transform (ART) analysis alongside ANOVA, we examined how L1 and accent interact to shape listener evaluations.

This research not only highlights how accent familiarity impacts perception but also provides practical insights for designing better text-to-speech (TTS) systems. By understanding what makes synthetic voices more acceptable and appealing across different linguistic groups, we can move closer to creating voices that truly resonate with diverse users.

2 Stimulus Creation

The stimuli for this study were generated using advanced AI speech synthesis technology provided by ElevenLabs, a platform known for producing natural-sounding speech and accurately replicating English accents associated with specific first-language (L1) speakers. To ensure consistency and minimize potential biases, all recordings were synthesized using male voices with uniform volume levels.

For the native English accent, Standard American English was selected (AI voice: Peter). Non-native accents were represented by languages from diverse language families, with AI voices chosen to best reflect the characteristic features of each accent: Mandarin Chinese (Jimmy), Polish (Grzegorz Turek), and Dutch (Ben van Praag). This selection aimed to investigate how familiarity with an accent influences perceptions of trustworthiness and pleasantness, providing a robust foundation for understanding the role of accent familiarity in shaping listener evaluations. Each accent was used to synthesize the same five prompts:

1. "Thank you for calling. How can I help you today?"
2. "The weather tomorrow will be sunny with a chance of rain in the evening."
3. "Please make sure to follow the instructions carefully."
4. "Your package is scheduled to arrive by 5 PM tomorrow."
5. "This event is open to everyone and starts at 10 AM."

3 Experiment Setup and Data Collection

The experiment was conducted using the [Qualtrics platform](#), where participants were recruited via social networks. Before starting the experiment, participants reviewed a consent form outlining the study's purpose, procedures, voluntary nature, and data processing policies. Consent was indicated by proceeding to the next page.

Participants then provided demographic information, including their first language, English proficiency level, and gender. The study primarily targeted native speakers corresponding to the synthesized accents but also allowed for additional responses to facilitate broader analysis if sufficient data were collected.

Participants listened to 20 audio recordings, consisting of the 5 prompts synthesized in 4 different accents. They rated each recording on pleasantness and trustworthiness using a 5-point Likert scale, where 1 represented "unpleasant" or "not trustworthy" and 5 represented "very pleasant" or "very trustworthy."

To minimize bias, the survey grouped responses by prompt while varying the order of accents presented within each prompt. The table below illustrates the randomized order of language accents across prompts:

Order	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
1	Polish	Dutch	Mandarin Chinese	American English	Polish
2	Dutch	Mandarin Chinese	American English	Polish	Dutch
3	Mandarin Chinese	American English	Polish	Dutch	Mandarin Chinese
4	American English	Polish	Dutch	Mandarin Chinese	American English

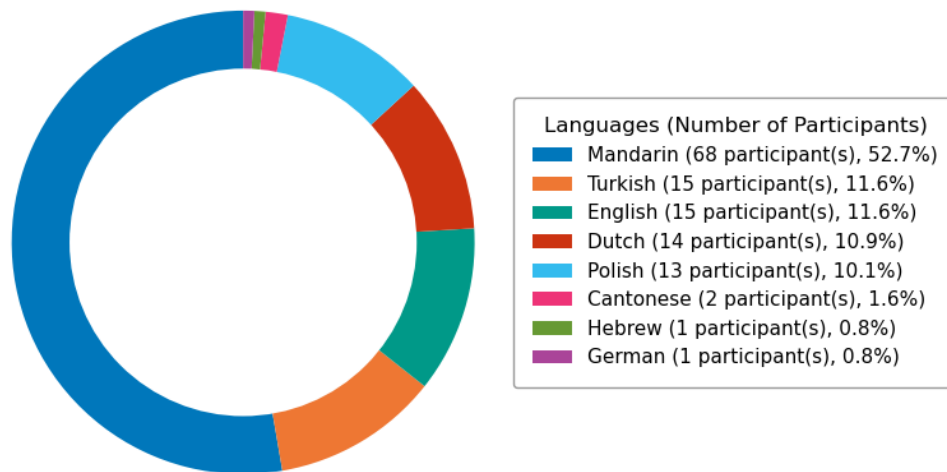
Table 1: Randomized order of language accents across prompts

4 Data Analysis

To analyze the data, we conducted an Aligned Rank Transform analysis in addition to ANOVA ("ART", Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 143-146)). While ANOVA examines the effects of independent variables, it does not account for interactions between variables such as L1 and accent on the two dependent variables: pleasantness and trustworthiness ratings. Furthermore, the prerequisites for ANOVA were not met, as the responses were not normally distributed, and the variances were not homogeneous. These violations were due to the ordinal nature of the Likert scale ratings, which were restricted to discrete values.

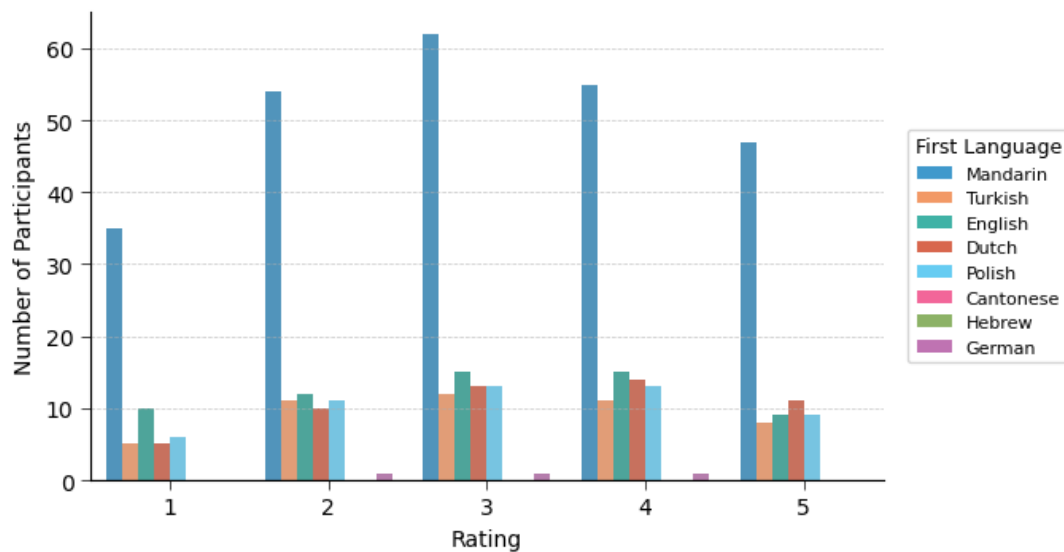
A total of 129 listeners participated in the survey, representing eight native languages. The largest cohorts included Mandarin speakers (52.7%), Turkish speakers (11.6%), and English speakers (11.6%). Each participant rated 20 accented or Standard American English sentences on a 1 to 5 Likert scale, resulting in 5,160 ratings overall.

Distribution of First Languages
Total Ratings: 5160



We applied the Shapiro-Wilk test to assess normality within pleasantness and trustworthiness ratings for each L1 group. The majority of these ratings were found to deviate from a normal distribution.

Number of Participants per Trustworthiness/Pleasantness Rating by L1 Group



Shapiro-Wilk Test Results for Pleasantness Ratings by L1

L1	n	W-statistic	p-value	Distribution
Mandarin	1360	0.909	9.6e-28	Non-normal
English	300	0.917	8.0e-12	Non-normal
Turkish	300	0.906	9.6e-13	Non-normal
Dutch	280	0.899	9.4e-13	Non-normal
Polish	260	0.915	5.4e-11	Non-normal
Cantonese	40	0.875	3.9e-04	Non-normal
German	20	0.812	0.001	Non-normal
Hebrew	20	0.907	0.055	Normal
Total	2580			

H_0 : Data follows a normal distribution ($p > 0.05$)

H_1 : Data does not follow a normal distribution ($p \leq 0.05$)

Shapiro-Wilk Test Results for Trustworthiness Ratings by L1

L1	n	W-statistic	p-value	Distribution
Mandarin	1360	0.906	3.5e-28	Non-normal
English	300	0.906	1.0e-12	Non-normal
Turkish	300	0.911	2.5e-12	Non-normal
Dutch	280	0.893	3.5e-13	Non-normal
Polish	260	0.906	1.1e-11	Non-normal
Cantonese	40	0.896	0.001	Non-normal
German	20	0.809	0.001	Non-normal
Hebrew	20	0.888	0.025	Non-normal
Total	2580			

H_0 : Data follows a normal distribution ($p > 0.05$)

H_1 : Data does not follow a normal distribution ($p \leq 0.05$)

Given this non-normality, we employed the ART method, which allows for the analysis of interactive variables (e.g., L1 and accent) using ranked data. Ranked data refers to the relative rankings of pleasantness or trustworthiness ratings within each group combination (i.e., each accent per L1). With four accents and eight L1 groups, there were 32 unique group combinations. After ranking, we performed ANOVA on the entire dataset, focusing on the ranked data to determine the significance of L1 and accent effects.

The ANOVA results revealed that L1 significantly influences listeners' perceptions of trustworthiness and pleasantness in accented English speech, as indicated by p-values below 0.001 and high F-statistics for both dependent variables. However, accents themselves did not show a significant effect, suggesting that listeners do not exhibit a preference for any particular accent, including Standard American English. The high residual sum of squares points to additional factors, such as listeners' English proficiency or gender—data we collected but reserved for analysis in future studies—as potential contributors to the ratings.

ANOVA Results for Pleasantness:

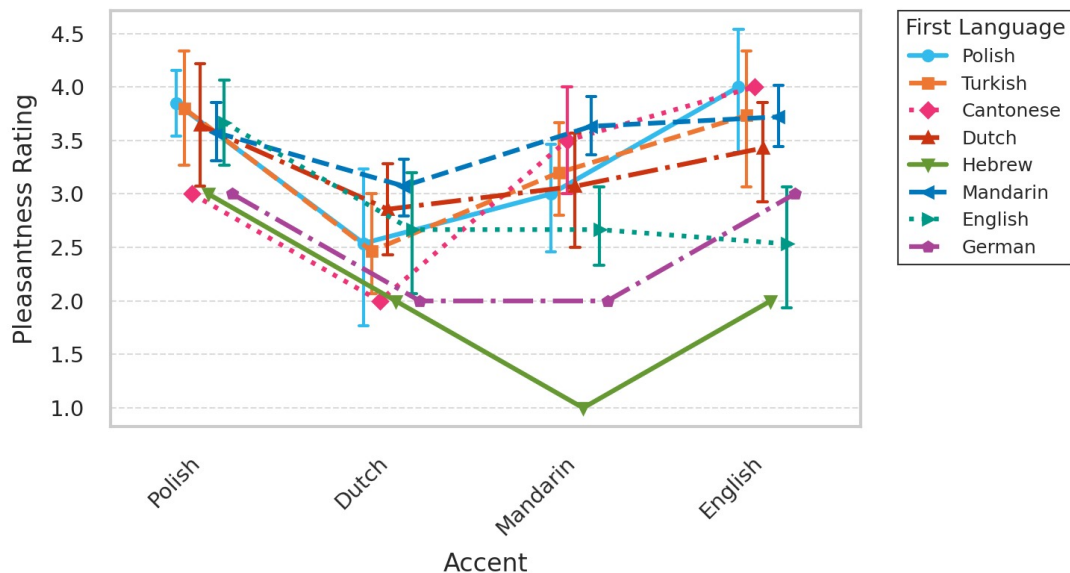
	sum_sq	df	F	PR(>F)
C(L1)	47486030.114	7.000	343.192	0.000
C(Accent)	14879.466	3.000	0.251	0.861
C(L1):C(Accent)	13296.891	21.000	0.032	1.000
Residual	50365215.329	2548.000	nan	nan

ANOVA Results for Trustworthiness:

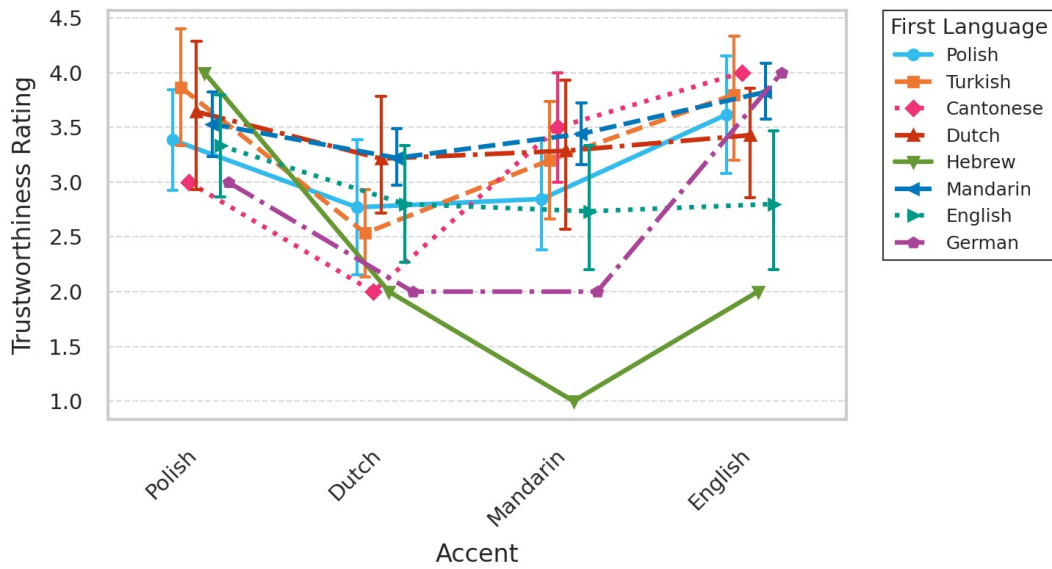
	sum_sq	df	F	PR(>F)
C(L1)	48827094.998	7.000	347.435	0.000
C(Accent)	14879.466	3.000	0.247	0.863
C(L1):C(Accent)	13296.891	21.000	0.032	1.000
Residual	51155022.329	2548.000	nan	nan

The following charts illustrate the interaction effect between L1 and accent on the two dependent variables. The symbols represent the mean ratings for each accent by each L1 cohort. Vertical bars indicate the standard deviation, reflecting the variability in ratings within each L1 cohort. The charts emphasize how various accented speech are perceived differently across L1 cohorts.

Interaction Effect: L1 Background and Accent on Pleasantness Ratings



Interaction Effect: L1 Background and Accent on Trustworthiness Ratings



To ensure the robustness of these findings, we also conducted pairwise post-hoc analyses to identify significant differences between accent pairs for both dependent variables. Consistent with the ANOVA results, no significant differences were observed between any accent pairs. This further supports the conclusion that no accent pair showed a significant difference in ratings, in line with the ANOVA results that accents do not significantly impact listeners' perceptions.

Post-Hoc Results for Pleasantness:

	group1	group2	meandiff	p-adj	lower	upper	reject
0	Dutch	English	6.625	0.929	-21.277	34.526	False
1	Dutch	Mandarin	3.800	0.985	-24.102	31.701	False
2	Dutch	Polish	4.610	0.974	-23.291	32.512	False
3	English	Mandarin	-2.825	0.994	-30.726	25.077	False
4	English	Polish	-2.015	0.998	-29.916	25.887	False
5	Mandarin	Polish	0.810	1.000	-27.091	28.712	False

Post-Hoc Results for Trustworthiness:

	group1	group2	meandiff	p-adj	lower	upper	reject
0	Dutch	English	-6.625	0.931	-34.828	21.579	False
1	Dutch	Mandarin	-3.800	0.986	-32.004	24.404	False
2	Dutch	Polish	-4.610	0.975	-32.814	23.594	False
3	English	Mandarin	2.825	0.994	-25.379	31.028	False
4	English	Polish	2.015	0.998	-26.189	30.218	False
5	Mandarin	Polish	-0.810	1.000	-29.014	27.394	False

5 Interpretation of the Results and Implications for TTS Design

Several surprising results emerged from the study, which we will now interpret and discuss in terms of their implications for TTS design.

One unexpected finding was the low score of the American English voice among native English speakers, with a mean value of 2.53. This outcome contradicted our initial hypothesis that familiarity with a given

accent would likely lead to higher scores. Since our native English participants included both Americans and Canadians, we had anticipated that the Standard American English accent would resonate positively with this group. The likely explanation for this result lies in the “uncanny valley” effect, as the voice reportedly lacked any regional accent typically associated with specific U.S. states. This absence of regional markers was only noticeable to American and Canadian participants, leading to lower ratings.

Conversely, the high score of the American English voice among non-native English speakers might reflect the perceived fluency, confidence, and prestige associated with this variety of English in non-English-speaking countries. These findings suggest that while Standard American English may be a suitable default for general-use TTS systems aimed at international users, developing region-specific TTS models is essential to enhance acceptability and user satisfaction among native English speakers.

Another noteworthy observation was the low score of the Dutch voice among Dutch participants. Although the voice exhibited a discernible accent based on a Flemish speaker, most Dutch participants seemed unable to identify its Flemish origin. This could suggest either a lack of recognition of the accent or a broader phenomenon tied to the linguistic environment in the Netherlands. Given the high level of English proficiency in the Netherlands—often characterized by minimal accents—participants may have less exposure to or tolerance for deviations from native-like English speech. This could result in a stricter evaluation of non-native accents, even subtle ones.

To explore these possibilities in future studies, it would be valuable to test participants’ ability to recognize foreign accents after completing the listening task. Additionally, investigating the relationship between participants’ English proficiency levels and their perceptions of accented speech could provide deeper insights into the impact of accent familiarity and expectations on user perception.

Additionally, we observed a strong correlation between trustworthiness and perceived pleasantness. This highlights the importance of designing voices with calm, reassuring, and friendly tones to enhance pleasantness and foster trust among users. Interestingly, the Polish voice was frequently singled out as particularly pleasant and received high scores across all participant groups. To investigate the influence of accent familiarity further, future studies could use a cloned voice (maintaining the same timbre) synthesized with different regional accents. This would enable a more precise comparison of how accent familiarity affects perceptions of pleasantness and trustworthiness. Moreover, testing participants’ ability to identify a speaker’s first language could yield valuable insights.

6 Limitations

The study revealed a slight centering bias in the evaluation results, with relatively few participants giving scores of 1 and the majority choosing scores of 3 or 4. To address this in future research, participants could be explicitly encouraged to use the full range of the evaluation scale.

Familiarity with accented English speech among speakers of different languages significantly influences perceived pleasantness and trustworthiness. Other factors, such as the voice’s timbre, speech rate, and fluency, also play a role. Future studies could standardize these variables across voices to more accurately assess the impact of accent familiarity on user perception.

7 Further Research

Future research could explore how perceptions of pleasantness and trustworthiness in English-accented speech vary based on participants’ declared levels of English proficiency and gender. Such studies would help further refine TTS designs to cater to diverse user groups effectively.

8 LLM Usage

LLM was consulted in the following tasks:

- Brainstorming research ideas; we generated the hypothesis by ourselves.
- Selecting statistical analysis algorithm.
- Proofreading the report.
- Plotting charts.