Title

The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise

Journal

JASA Express Letters, 2(4)

ISSN

2691-1191

Authors

Aoki, Nicholas B
Cohn, Michelle
Zellou, Georgia

Publication Date

2022-04-01

DOI

10.1121/10.0010274

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise

Nicholas B. Aoki,[a] **Michelle Cohn, and Georgia Zellou**
Department of Linguistics, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, California 95616, USA
nbaoki@ucdavis.edu, mdcohn@ucdavis.edu, gzellou@ucdavis.edu

Abstract: This study examined how speaking style and guise influence the intelligibility of text-to-speech (TTS) and naturally produced human voices. Results showed that TTS voices were less intelligible overall. Although using a clear speech style improved intelligibility for both human and TTS voices (using "newscaster" neural TTS), the clear speech effect was stronger for TTS voices. Finally, a visual device guise decreased intelligibility, regardless of voice type. The results suggest that both speaking style and visual guise affect intelligibility of human and TTS voices. Findings are discussed in terms of theories about the role of social information in speech perception.

[Editor: Douglas D O'Shaughnessy] https://doi.org/10.1121/10.0010274

Received: 12 January 2022 **Accepted:** 1 April 2022 **Published Online:** 20 April 2022

1. Introduction

Spoken interactions with voice-activated artificially intelligent (voice-AI) assistants, like Amazon's Alexa (Amazon, Seattle, WA), are increasingly common (Ammari *et al.*, 2019). Voice-AI assistants use text-to-speech (TTS) voices to engage with users, and although TTS voices have become more naturalistic (van den Oord *et al.*, 2016), they are consistently less intelligible than naturally produced speech (Simantiraki *et al.*, 2018). To increase the intelligibility of TTS voices, one solution is to implement a clear speaking style for TTS voices, paralleling how human talkers overcome communication barriers. Clear speech enhances intelligibility for human listeners (e.g., Cohn *et al.*, 2021) through a variety of acoustic–phonetic modifications relative to casual speech, such as a slower speaking rate and higher pitch (Uchanski, 2005). In studies of TTS intelligibility, however, speaking style is often not studied (e.g., Simantiraki *et al.*, 2018) or TTS voices are not directly compared to human voices (Cohn and Zellou, 2020).

Recent advances in neural TTS allow for the creation of different speech styles, created by training a neural network on speech samples by a particular speaker and in a specified style (Wood and Merritt, 2018). The current study examines the intelligibility of two TTS speaking styles: a clear "newscaster" style and a conversational, casual style (Pelzer and Sanchez, 2020). The newscaster style was selected as a clear speech proxy given evidence that, at least for certain English-speaking human newscasters, this style resembles clear speech. For example, Gasser *et al.* (2019) found that, relative to non-newscasters, Boston newscasters spoke with a slower speaking rate. Moreover, several newscasters report that they aim to "enhance listener comprehension", which is one of the goals of clear speech (Uchanski, 2005). By presenting both clear and casual TTS styles as well as naturally produced voices (here, referred to as "human" voices) in clear and casual styles, the present study tests whether listeners experience a clear speech benefit for TTS and compares the relative intelligibility of human and TTS voices for both speaking styles.

1.1 How is intelligibility influenced by visual guise?

Besides speaking style, visual information may also influence the intelligibility of TTS and human voices. However, how visual cues affect intelligibility is debated. A *bias* account posits that learning the social identity of the speaker activates stereotypes that can reduce listener comprehension. For example, in a speech-perception-in-noise (SPIN) task, Yi *et al.* (2013) presented sentences produced by native or Korean-accented English speakers in an audio-only or audiovisual condition to listeners (native English speakers). Although the audiovisual condition improved performance overall, the audiovisual gain was lower for the Korean-accented speakers, suggesting that seeing an Asian face induces expectations of a foreign accent and that stronger stereotype activation leads to a greater intelligibility detriment. Given that TTS voices

[a]Author to whom correspondence should be addressed.

are rated as less communicatively competent than human voices (Cowan *et al.*, 2015) and that people produce more effortful speech towards voice-AI devices (e.g., Cohn and Zellou, 2021), a *bias* account predicts that, regardless of speaking style, accuracy should decrease when the guise provides cues that the speaker is a device (i.e., when a device picture is seen).

A *congruency* account claims that paralinguistic cues activate linguistic exemplars associated with social identities, so speech input matching those exemplars should improve recognition. McGowan (2015) showed that performance on a SPIN task for a Chinese-accented voice improved when a Chinese face was shown, relative to a Caucasian face. This is not easily explained by a *bias* account as it would predict lower intelligibility for the Chinese face condition. McGowan (2015) suggested that intelligibility improved in the Chinese face condition because the visual information better matched expectations for Chinese-accented speech. In the present study, a *congruency* prediction is that seeing a device should increase intelligibility for TTS voices and decrease intelligibility for human voices, and *vice versa*. A *congruency* account might also predict an interaction between guise and speaking style. Given that "casual" TTS voices are rated as more "human-like" and "natural", than "clear" styles (Cohn and Zellou, 2020), a device guise may be more incongruous with a casual style. Thus, intelligibility may be lower for a casual style when a device picture is shown. A final *congruency* possibility is that an incongruent guise only reduces intelligibility for human voices, since anthropomorphism of TTS voices has been observed in prior work (Cohn *et al.*, 2020), but "device-ification" of humans is not an observed behavior [but see Mendelsohn *et al.* (2020) for a discussion of linguistic dehumanization of marginalized groups].

2. Methods

2.1 Participants

A total of 67 participants were recruited through the University of California, Davis Psychology subjects pool and received course credit for their participation. Data from four subjects were removed as they reported hearing difficulties. Data from two additional subjects were removed as accuracy was more than two standard deviations below the mean. The remaining 61 participants (40 female; mean age = 19.97 y, SD = 2.11) were native English speakers.

2.2 Stimuli

Stimuli consisted of 144 semantically unpredictable sentences from the Speech Perception in Noise test (Kalikow *et al.*, 1977), which all contain a phrase-final keyword (e.g., "We've spoken about the truck."). Four speakers (two human, two TTS) produced each sentence in both clear and casual styles. The human voices (one male, one female) were native speakers of American English who were recorded talking to a real listener using a head-mounted microphone (Shure™ WH20XR) and a USB audio mixer (Steinberg™ UR12). They were recorded in a quiet room at home rather than a sound booth due to COVID-19 measures. They were instructed to "say the sentences in a natural, casual manner" and to "speak clearly to someone who may have trouble understanding you" for the casual and clear styles, respectively (the productions were taken from a subset of items used in Cohn *et al.*, 2021). The TTS stimuli were generated with Amazon Polly (US-English) for two speakers (the male "Matthew" voice and the female "Joanna" voice) in the (default) neural TTS for the casual style and in the newscaster style for the clear style.

Duration (in milliseconds) and mean pitch (in Hertz) were measured on the target words using Praat (Boersma and Weenink, 2021). Paired t-tests were run between the casual and clear styles within each voice type, and the effect size for each of these comparisons was calculated with Cohen's d (Cohen, 1992). The interpretation of effect size follows Cohen's benchmarks (negligible if $|d| < 0.2$; small if $|d|$ is between 0.2 and 0.5; medium if $|d|$ is between 0.5 and 0.8; large if $|d|$ is greater than 0.8). For the human voices, the target words in clear speech had significantly higher pitch than in casual speech (mean difference = 12.26 Hz, $t = 6.40$, $p < 0.001$), but the effect size was negligible ($|d| = 0.18$); meanwhile, the target words in clear speech were significantly longer than in casual speech (mean difference = 0.16 ms, $t = 29.02$, $p < 0.001$) with a large effect size ($|d| = 1.52$). For the TTS voices, the target words in clear speech had significantly higher pitch than in casual speech (mean difference = 14.16 Hz, $t = 13.24$, $p < 0.001$) with a small effect size ($d = 0.42$); the TTS target words in clear speech were significantly shorter than in casual speech (mean difference = 0.01 ms, $t = 4.37$, $p < 0.001$) with a negligible effect size ($|d| = 0.18$). In summary, non-negligible effect sizes were found for two acoustic patterns: the longer target word duration in the clear style compared to the casual style for the human voices; the greater target word pitch in the clear style compared to the casual style for the TTS voices. Thus, although the clear speech stimuli for both the human and TTS voices were produced with acoustic features associated with more effortful speech, duration is more heavily weighted in the clear speech stimuli for the human voices while pitch is weighted more heavily in the clear speech stimuli for the TTS voices.

All human and TTS stimuli were resampled to 44.1 kHz and amplitude-normalized to 65 dB in Praat. Speech-shaped noise (SSN) was created using the long-term spectrum (LTAS) for all sentences combined (Winn, 2019). The sentences were mixed with noise at a −3 dB signal-to-noise ratio (McCloy, 2015), with noise starting 500 ms before the sentence onset and ending 500 ms after the sentence offset.

2.3 Procedure

Participants completed the experiment online *via* Qualtrics. On each trial, participants heard a sentence once and were asked to type the last word of each sentence. The 144 sentences were presented with the speaker/style combinations equally and pseudo-randomly assigned across 8 lists. Stimulus presentation order of the 144 sentences was randomized for each participant.

For half of the trials, voices were paired with static stock images of human faces while in the other half of trials, TTS voices were paired with pictures of cylindrical devices similar to Amazon's first generation Echo device (Fig. 1). Participants were randomly assigned to a "Congruent" or "Incongruent" condition (29 of 61 subjects were in the Congruent condition). In the Congruent condition, images matched the voice type (e.g., human voices were paired with human images and TTS voices were paired with device images). In the Incongruent condition, the human voices were paired with the device pictures and *vice versa*. Gender was matched in both the Congruent and Incongruent conditions. For example, in the Congruent condition, the human female face was shown with the human female voice, and in the Incongruent condition, the human female face was shown with the female TTS voice. The contrast between the male and female TTS voices was represented with the device pictures using a red/blue color contrast (background color–voice gender assignment was randomized across participants).

2.4 Statistical analysis

Keyword accuracy was coded binomially as correct (= 1) or incorrect (= 0). Only responses with all and only the correct affixes were counted as correct (e.g., "strip" was considered incorrect if the right answer was "strips"). Obvious spelling mistakes were counted as correct, however: "sleaves" for "sleeves", "jointd" for "joints", "shead" for "shed", "nob" for "knob", "heard/hurd" for "herd", "theif" for "thief", "brews" for "bruise", "funn/funp" for "fun", and "witts" for "wits."

The binomial data were modeled with mixed-effects logistic regression through the *lme4* package in R (Bates *et al.*, 2015). The model included fixed effects of Style (clear, casual), Voice (human, TTS), and Guise (human, device), and all possible interactions. Random effects included by-listener, by-talker, and by-sentence random intercepts. By-listener random slopes for Voice, Guise, and Style each introduced singularity issues and were excluded. The model was checked following Sonderegger *et al.* (2018). The binned residual plot showed that the model was appropriate (92% of residuals were within the error bounds). All Cook's distance values were below 4/n, thus showing no evidence of influential points. Random intercept distributions were approximately normal.

3. Results

Figure 2 provides aggregated accuracy proportions in each condition. The logistic regression model revealed an effect of Style ($Coef = -0.3$, $SE = 0.03$, $z = -11.01$, $p < 0.001$). As seen in Fig. 2, casual speech is less intelligible overall (28.7% correct) than clear speech (38.5%). Additionally, there was an effect of Voice ($Coef = -0.6$, $SE = 0.26$, $z = -2.34$, $p < 0.05$): TTS voices are less intelligible overall (23.4%) than human voices (43.8%). There was also an effect of Guise ($Coef = -0.09$, $SE = 0.03$,

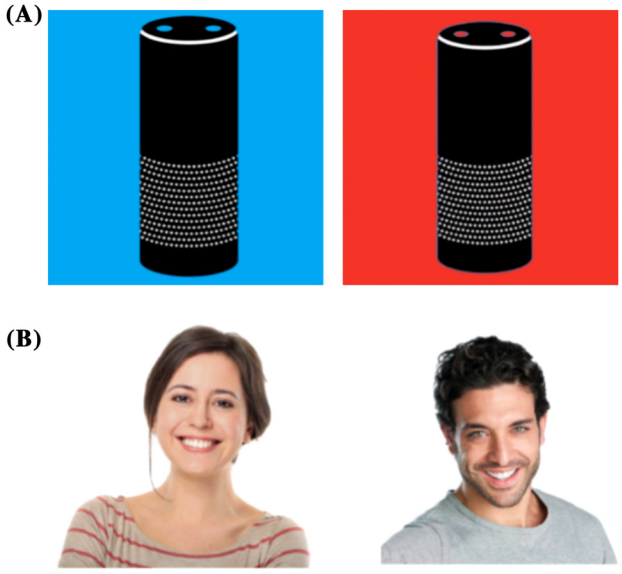

Fig. 1. The images in the experiment are shown for the (A) device and (B) human guises. Separate talkers for the device guises were indicated with a red or blue background.

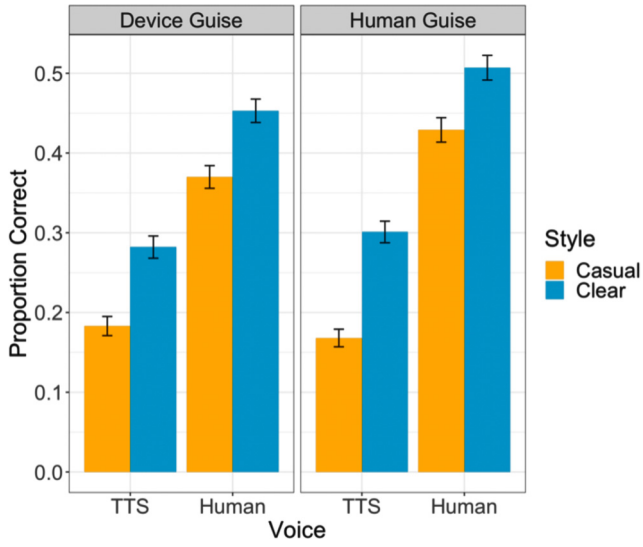

Fig. 2. Proportion of keywords correctly recognized for clear and casual speaking styles across TTS and human voices as a function of the device guise (left panel) and the human guise (right panel). Error bars represent standard errors. (Note that the logistic regression model takes into account listener, speaker, and sentence variation that is not displayed.)

$z = -3.23$, $p < 0.01$), indicating that seeing a device picture decreases intelligibility (32.6%) relative to seeing a picture of a human (34.6%).

Finally, there was an interaction between Voice and Style (*Coef* $= -0.09$, *SE* $= 0.03$, $z = -3.24$, $p < 0.01$). While clear speech is more intelligible than casual speech for both human and TTS voices, casual speech is even less intelligible than clear speech for the TTS voices. For the TTS voices, accuracy for casual speech and clear speech was 17.5% and 29.2%, respectively. For the human voices, accuracy for casual speech and clear speech was 39.8% and 47.9%, respectively. Note that although the increase in accuracy for the human guise appears to be numerically larger for the human voices compared to the TTS voices in Fig. 2, the model did not reveal a significant interaction between Voice and Guise (*Coef* $= 0.07$, *SE* $= 0.06$, $z = 1.18$, $p = 0.24$).

4. Discussion

Human talkers naturally adopt a clear speaking style in difficult listening situations, and this adaptation benefits listeners by increasing comprehension (e.g., Uchanski, 2005). Prior studies of TTS voice intelligibility have been limited as they have not compared intelligibility across speaking styles (e.g., Cohn and Zellou, 2020). The current study fills this gap by finding that, similar to human voices, a clear speech style boosts intelligibility for TTS relative to a more casual style. In other words, listeners can leverage clear speech effects for both human and TTS voices.

As in other work, TTS voices are overall less intelligible than human voices (e.g., Simantiraki *et al.*, 2018). Moreover, the current study shows that seeing an image of a device decreases intelligibility for both voice types, thus supporting *bias* accounts of how visual guises affect speech perception (e.g., Yi *et al.*, 2013). Given that devices are perceived as less communicatively competent than humans (Cowan *et al.*, 2015; Cohn *et al.*, 2022), looking at a device may trigger this stereotype and lower comprehension. More broadly, this finding builds on work showing that socio-indexical information and speech perception are intertwined (e.g., D'Onofrio, 2015) and contributes to research indicating that people have distinct mental representations for humans and devices, which affect speech perception (e.g., Zellou *et al.*, 2021).

Although both human voices and TTS voices showed a clear speech benefit, the intelligibility gain for the clear TTS relative to casual TTS was even greater than the clear speech benefit for the human voices. On the one hand, this effect could be accounted for by a *congruency* effect (e.g., McGowan, 2015): casual speech might be more incongruous with devices (e.g., Cohn and Zellou, 2020), thus resulting in lower intelligibility. However, the lack of evidence for an interaction between guise and style makes this interpretation more tenuous. Another possibility is that the pitch differences between clear and casual TTS may have supported greater intelligibility than the duration differences observed across clear and casual human speech. Future work examining how the acoustic properties of human and TTS voices relate to intelligibility differences can explore this aspect further.

There are several limitations of the present study that open directions for future work. First, although a "newscaster" speaking style may resemble clear speech for American English speakers, whether this transfers to other languages (and consequently, to TTS voices in other languages) remains an open question. Second, prior work has demonstrated that speaker gender plays a role in intelligibility (Bradlow *et al.*, 1996). In the current study, we only used one voice

per gender within each voice type (at the time of the study, only one male and one female voice were available in the "newscaster" TTS style). Future work investigating the role of speaker gender across voice types can provide a more comprehensive account of the factors that affect TTS voice intelligibility. Furthermore, the visual guises in the present study used cylindrical device silhouettes, in contrast with real human faces, in order to maximize the difference between the human and device guises. Future experiments showing more anthropomorphic robots (e.g., Cohn *et al.*, 2020) can test whether a greater degree of human-likeness reduces bias and generates different intelligibility effects.

There are several practical implications of the present findings. First, if devices can adapt appropriately between casual and clear speaking styles, similar to human talkers, then devices may be perceived as more natural. Moreover, using a clear speaking style for TTS voices and showing a human face may be helpful at improving TTS voice intelligibility in noisy environments and for individuals with communicative barriers (e.g., L2 listeners or those who are hard of hearing). As voice-AI interfaces become even more commonplace, understanding the factors that shape TTS intelligibility will be relevant for matters of accessibility, as well as our broader scientific understanding of human–computer interaction.

Acknowledgments

This research was supported by the National Science Foundation SBE Postdoctoral Research Fellowship to M.C. under Grant No. 1911855 and an Amazon Research Grant to G.Z.

References and links

Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. (**2019**). "Music, search, and IoT: How people (really) use voice assistants," ACM Trans. Comput-Hum. Interact. **26**(3), 1–28.

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.

Boersma, P., and Weenink, D. (**2021**). "Praat: Doing phonetics by computer (version 6.1.40) [computer program]," https://www.fon.hum.uva.nl/praat/ (Last viewed December 3, 2021).

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**(3), 255–272.

Cohen, J. (**1992**). "A power primer," Psychol. Bull. **112**, 155–159.

Cohn, M., Jonell, P., Kim, T., Beskow, J., and Zellou, G. (**2020**). "Embodiment and gender interact in alignment to TTS voices," in *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, Toronto, Canada (July 29-August 1), pp. 220–226.

Cohn, M., Pycha, A., and Zellou, G. (**2021**). "Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech," Cognition **210**, 104570.

Cohn, M., Segedin, B. F., and Zellou, G. (**2022**). "Acoustic-phonetic properties of Siri- and human-directed speech," J. Phon. **90**, 101123.

Cohn, M., and Zellou, G. (**2020**). "Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes," in *Proceedings of Interspeech 2020*, Shanghai, China (October 25-29), pp. 1733–1737.

Cohn, M., and Zellou, G. (**2021**). "Prosodic differences in human- and Alexa-directed speech, but similar local intelligibility adjustments," Front. Commun. **6**, 675704.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (**2015**). "Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue," Int. J. Hum. Comput. **83**, 27–42.

D'Onofrio, A. (**2015**). "Persona-based information shapes linguistic perception: Valley Girls and California vowels," J. Socioling. **19**(2), 241–256.

Gasser, E., Ahn, B., Napoli, D. J., and Zhou, Z. L. (**2019**). "Production, perception, and communicative goals of American newscaster speech," Lang. Soc. **48**(2), 233–259.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**(5), 1337–1351.

Mendelsohn, J., Tsvetkov, Y., and Jurafsky, D. (**2020**). "A framework for the computational linguistic analysis of dehumanization," Front. Artif. Intell. **3**, 55.

McCloy, D. (**2015**). "Mix speech with noise [Praat script]," https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat (Last viewed December 3, 2021).

McGowan, K. B. (**2015**). "Social expectation improves speech perception in noise," Lang. Speech **58**(4), 502–521.

Pelzer, J., and Sanchez, A. (**2020**). "Giving your content a voice with the Newscaster speaking style from Amazon Polly," AWS Mach. Machine Learning. Blog.

Simantiraki, O., Cooke, M., and King, S. (**2018**). "Impact of different speech types on listening effort," in *Proceedings of Interspeech 2019*, Graz, Austria (September 15-19), pp. 2267–2271.

Sonderegger, M., Wagner, M., and Torreira, F. (**2018**). *Quantitative Methods for Linguistic Data*, 1st ed. (digital).

Uchanski, R. M. (**2005**). "Clear speech," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. Remez (Blackwell, Malden, MA), pp. 207–235.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (**2016**). "WaveNet: A generative model for raw audio," arXiv:1609.03499.

Winn, M. (**2019**). "Make speech-shaped noise [Praat script]," http://www.mattwinn.com/praat/Make_SSN_from_LTAS_selected_sounds.txt (Last viewed December 3, 2021).

Wood, T., and Merritt, T. (**2018**). "Varying speaking styles with neural text-to-speech," Amazon Science.

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., and Chandrasekaran, B. (**2013**). "Reduced efficiency of audiovisual integration for nonnative speech," J. Acoust. Soc. Am. **134**(5), EL387–EL393.

Zellou, G., Cohn, M., and Block, A. (**2021**). "Partial compensation for coarticulatory vowel nasalization across concatenative and neural text-to-speech," J. Acoust. Soc. Am. **149**(5), 3424–3436.