Title

Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners

Authors

Aoki, Nicholas B

Zellou, Georgia

Publication Date

2024-05-01

DOI

10.1016/j.wocn.2024.101328

Journal of Phonetics 104 (2024) 101328

Contents lists available at ScienceDirect

Journal of Phonetics

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

Being *clear* about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1- and L2-English listeners

Nicholas B. Aoki *, Georgia Zellou

Department of Linguistics, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616, USA

ARTICLE INFO

Article history:
Received 29 August 2023
Received in revised form 1 April 2024
Accepted 2 April 2024

Keywords:
Speech intelligibility
Clear speech
Hard-of-hearing-directed speech
Foreigner-directed speech
Non-native listeners
Hyper-articulation
Phonetic variation

ABSTRACT

Relative to one's default (casual) speech, clear speech contains acoustic modifications that are often perceptually beneficial. Clear speech encompasses many different styles, yet most work only compares clear and casual speech as a binary. Furthermore, the term "clear speech" is often *unclear* – despite variation in elicitation instructions across studies (e.g., speak clearly, imagine an L2-listener or someone with hearing loss, etc.), the generic term "clear speech" is used when interpreting results, under the tacit assumption that clear speech is monolithic. The current study examined the acoustics and intelligibility of casual speech and two clear styles (hard-of-hearing-directed and non-native-directed speech). We find: (1) the clear styles are acoustically distinct (non-native-directed speech is slower with lower mean intensity and f0); (2) the clear styles are perceptually distinct (only hard-of-hearing-directed speech enhances intelligibility); (3) no differences in intelligibility benefits are observed between L1 and L2-listeners. These results underscore the importance of considering the intended interlocutor in speaking style elicitation, leading to a discussion about the issues that arise when reference to "clear speech" lacks clarity. It is suggested that to be more *clear* about clear speech, greater caution should be taken when interpreting results about speaking style variation.

1. Introduction

When listeners appear to encounter comprehension difficulties, speakers tend to shift from their default speaking mode (termed "casual", "plain", or "conversational" speech) to "clear speech" (Smiljanić & Bradlow, 2009). Clear speech encompasses a range of speaking styles that are hyper-articulated[1] (i.e., produced with exaggerated acoustic–phonetic[2] modifications, such as greater intensity, higher f0, and a slower speaking rate) with the goal of facilitating listener understanding (Uchanski, 2005). Clarity-intended acoustic enhancements are largely successful at boosting perception – despite some exceptions (Ferguson & Kewley-Port, 2002), clear speech intelligibility enhancements are often found in speech-transcription-in-noise tasks (Payton et al., 1994; Aoki & Zellou, 2023a). In general, the existence of casual and clear speech is often accounted for by Hypo- and Hyper-articulation (H&H) Theory (Lindblom, 1990), which views speaking style variation as a balance between speaker-oriented goals (hypo-articulating to minimize articulatory effort) and listener-oriented goals (hyper-articulating to increase the likelihood of being understood).

Many different subtypes of clear speech have been documented and investigated, such as hard-of-hearing-directed (Scarborough & Zellou, 2013), non-native-directed[3] (Rothermich et al., 2019), and device-directed speech (Cohn

* Corresponding author at: Department of Linguistics, University of California, Davis, 469 Kerr Hall, One Shields Avenue, Davis, CA 95616, USA.
E-mail address: nbaoki@ucdavis.edu (N.B. Aoki).

[1] We consider hyper-articulation to only refer to the acoustic properties of speech, potentially including a wide variety of styles with exaggerated modifications (hard-of-hearing-directed speech, infant-directed speech, emotional speech, etc.). Clear speech is a more specific category within hyper-articulation, only comprising styles that are assumed to have an intention of enhancing clarity (i.e., hard-of-hearing-directed speech, but not emotional speech).

[2] Our definition of clear speech is intentionally speaker-focused, not listener-focused (i.e., clear speech is based on whether the speaker intends to be clear, not whether the listener actually receives a perceptual benefit). Having a speaker-focused, rather than a listener-focused interpretation, makes the definition of clear speech more stable and thus, easier to discuss. If clear speech is predicated on the listener, then whether a particular set of productions are considered to be "clear speech" could vary from one study to the next (as an extreme example, if the signal-to-noise ratio is too low and listeners cannot perceive anything, then according to the listener-based definition, nothing would be considered to be clear speech). However, if clear speech is based on the speaker, then as long as the speaker makes exaggerated acoustic modifications and is intending to be clear, productions can always be considered as "clear speech" regardless of listener effects.

[3] Recent work has problematized the terms "native speaker" and "non-native speaker" to refer to first-language and second-language speakers (Cheng et al., 2021). The current study employs the terms "L1" and "L2" to refer to the actual participants in the experiments, consistent with other studies (e.g., McLaughlin & Van Engen, 2020; Aoki & Zellou, 2023c). However, we continue to use the term "non-native-directed speech" to describe the speaking style because this label is more faithful to the instructions given to speakers in this study (Imagine talking to a listener who is a "native speaker of Mandarin and is learning English").

https://doi.org/10.1016/j.wocn.2024.101328
0095-4470/

et al., 2022), with talkers making targeted acoustic enhancements based on interlocutor identity. However, these subtypes are rarely compared directly (especially in the intelligibility literature) with little to no research examining the acoustic and perceptual properties of multiple clear styles (cf., Lam & Tjaden, 2013; Scarborough & Zellou, 2013). The vast majority of studies (including some of the authors' own work; e.g., Aoki et al., 2022; Aoki & Zellou, 2023a) instead look at clear and casual speech as a binary, presumably inspired by H&H Theory.

Among work that examines speaking style as a clear-casual dichotomy, the term "clear speech" is often *unclear*. Specifically, although studies vary widely in elicitation instructions (e.g., speak clearly, imagine a listener with hearing loss, imagine an L2-listener, etc.), the same, blanket term "clear speech" is almost always used when interpreting results. The phrase "clear speech" is thus usually ambiguous, potentially referring to either hard-of-hearing-directed speech, non-native-directed speech, or another speaking style. This lack of clarity is illustrated in Table 1, which categorizes examples of clear speech conditions in the literature based on the imagined interlocutor, elicitation instructions, and interpretation of results.

The current study fills a gap in the literature by examining the acoustic properties of casual speech and multiple clear styles (hard-of-hearing-directed and non-native-directed speech), and then assessing their intelligibility for L1- and L2-English listeners. This work ultimately advocates for greater clarity about "clear speech".

In the rest of the introduction, we first problematize the unclear nature of "clear speech" in Section 1.1. Section 1.2 discusses our secondary goal: addressing theories of how listener language background affects speaking style intelligibility. The introduction concludes with Section 1.3, which states the design and hypotheses of this study.

1.1. Problematizing the lack of clarity about "clear speech"

Using the blanket, non-specific term "clear speech" as a substitute for a specific elicited style (see Table 1) portrays clear speech as monolithic, even though clear speech actually denotes a constellation of acoustically distinct styles that happen to share similar properties (i.e., acoustic exaggerations with the goal of enhancing listener comprehension). The issue with this portrayal is that it could lead readers to assume that findings for one type of clear speech will generalize to all types of clear speech even when this has not been explicitly tested.

This scenario is not merely hypothetical. A case in point is the comparison between Kang and Guion (2008) and Jung and Dmitrieva (2023a), which are both quoted in Table 1. Kang and Guion (2008) elicited three speaking styles (conversational, citation-form, and clear) from older and younger L1-Korean participants, finding that in clear speech, younger speakers primarily enhanced the 3-way Korean stop contrast with onset f0, as opposed to voice onset time (VOT).

In a later study, Jung and Dmitrieva (2023a) recorded two speaking styles (casual, clear) from young, Korean-accented English speakers. Under the implicit assumption that their experiment is directly comparable to Kang and Guion (2008) because both elicited the "same" speaking style ("clear speech"), Jung and Dmitrieva (2023a) make the following prediction: "in Korean-accented English clear speech onset f0 differences between voiced and voiceless consonants could be over-enhanced in a non-native manner, possibly at the expense of enhancing the VOT contrast" (p. 3). However, Jung and Dmitrieva (2023a) do not find support for this hypothesis: "contrary to our expectations, the [Korean-accented English] group did not enhance [the] English voicing contrast in clear speech in terms of onset f0. . .It is an intriguing finding, and it fits into the overall pattern of non-native speakers not reaching or exceeding the magnitude of acoustic modifications of native clear speech." (p. 8). In short, Jung and Dmitrieva (2023a) attribute an acoustic difference (presence versus absence of onset f0 enhancement in "clear speech") between their study and Kang and Guion (2008) to the language being spoken (L1-Korean versus Korean-accented English).

These two studies, however, elicited distinct types of clear speech with different imagined interlocutors. Whereas Kang and Guion (2008) recorded non-native-directed speech, Jung and Dmitrieva (2023a) asked speakers to produce hard-of-hearing-directed speech (see Table 1). It is thus *unclear* whether acoustic differences between the studies are due to the language being spoken (as claimed by Jung and Dmitrieva) or to differences in the instructions given to speakers. A broader issue is that the authors in both papers use the same label of "clear speech", potentially misleading readers into thinking that the same clear speaking style has been elicited.

The tendency to replace a specific style with the generic term "clear speech" is common throughout the literature (including some of the authors' own prior work), and is not just a characteristic of Jung and Dmitrieva (2023) and Kang and Guion (2008). Since so few studies examine speaking style variation beyond the clear-casual binary (cf. Scarborough & Zellou, 2013), it is not well-understood whether the intended interlocutor of clear speech is important. On the one hand, perhaps using the non-specific term "clear speech" is justified because all clear styles generally overlap in acoustic and perceptual properties. However, if clear styles differ greatly, then using the same term ("clear speech"), regardless of the intended interlocutor, could mislead readers into thinking that clear speech is monolithic. This could subsequently result in confusion when comparing studies with different clear speech elicitation instructions (e.g., Kang and Guion (2008) and Jung and Dmitrieva (2023)). The current study tests whether this concern is justified by directly comparing the acoustic and perceptual properties of two clear styles: hard-of-hearing-directed and non-native-directed speech.

1.2. Effects of listener language background on speaking style intelligibility

Prior work has suggested that the clear speech intelligibility benefit is reduced for L2 listeners relative to L1 listeners (Bradlow & Bent, 2002; Bradlow & Alexander, 2007). Bradlow and Bent (2002) account for these results by claiming that clear speech is "native-listener oriented" (p. 272). Certain featural enhancements, like a slower speaking rate (Smiljanić & Bradlow, 2005) and vowel space expansion (Bradlow, 2002), are language-general and presumably leveraged equally by both L1 and L2 listeners when transcribing speech in noise. However, other modifications are language-specific –

Table 1
Non-exhaustive summary of the speaking style literature showing (a) variation in the type of clear speech elicited (Columns 2 and 3), but (b) uniformity in the label ("clear speech") used for describing the results (Column 4). In Column 4, "clear speech" is bolded for emphasis.

Sample Studies	Imagined Interlocutor	Sample Elicitation Instructions	Sample Interpretation of Results
Ferguson & Kewley-Port, 2002; Kato & Baese-Berk, 2023; Jung & Dmitrieva, 2023a	Hard-of-Hearing Listener	"pronounce the words clearly, as if [you] were talking to a hearing-impaired or elderly person" (Jung & Dmitrieva, 2023a, p. 3).	"L2 speakers implemented less vowel space expansion, less increase of mean f0, and less positive and negative VOT lengthening in **clear speech** than native speakers" (Jung & Dmitrieva, 2023a, p. 1)
Kang & Guion, 2008; Smiljanić et al., 2021	L2 ("Non-Native") Listener	"read in a 'clear' way, as if speaking to a 'foreigner' audience who needs greater linguistic-phonetic resources to have full access to the linguistic information" (Kang & Guion, 2008, p. 3915)	"Results indicated that the older group solely used VOT to enhance the contrast in **clear speech**, whereas the younger group primarily used F0 but also demonstrated small VOT enhancement" (Kang & Guion, 2008, p. 3909).
Bradlow & Bent, 2002; Bradlow & Alexander, 2007; Van Engen et al., 2014	Hard-of-Hearing or L2 ("Non-Native") Listener	"read the sentences as if speaking to a listener with a hearing loss or from a different language background" (Bradlow & Bent, 2002, p. 275).	"Results showed that while native listeners derived a substantial benefit from naturally produced **clear speech**. . .non-native listeners exhibited only a small **clear speech** effect" (Bradlow & Bent, 2002, p. 272).
Cohn et al., 2021; Aoki et al., 2022; Zellou et al., 2022	No Interlocutor Specified	"speak clearly to someone who may have trouble understanding you" (Aoki et al., 2022, p. 2).	"Although using a **clear speech** style improved intelligibility for both human and TTS voices. . .the **clear speech** effect was stronger for TTS voices" (Aoki et al., 2022, p. 1).

for example, English speakers modulate the voice onset time of voiceless stops more than voiced stops, while Croatian speakers show the reverse pattern (Smiljanić & Bradlow, 2008). Given their reduced experience with the target language, L2 listeners are, in theory, not as equipped to take advantage of language-specific acoustic enhancements, leading to a diminished clear speech intelligibility benefit.

However, an alternative explanation for the findings of Bradlow and Bent (2002) is that there is a mismatch between the intended and actual listener. In many intelligibility experiments, elicitation instructions are vague, with speakers instructed to talk clearly to *either* an imagined hard-of-hearing-listener or an L2 listener (e.g., Bradlow & Bent, 2002; Bradlow & Alexander, 2007; refer to Table 1). The lack of specificity in elicitation instructions means that it is not known whether speakers are envisioning a hard-of-hearing or an L2 listener. If, for instance, in Bradlow and Bent (2002) and Bradlow and Alexander (2007), the speakers chose to imagine hard-of-hearing listeners in the clear speech condition, reduced transcription performance for L2 listeners may have resulted from the presentation of an inappropriate speaking style, as opposed to a diminished capacity to leverage native-listener-oriented acoustic properties.

If speech adaptations are targeted to benefit the interlocutor, then aligning the intended and actual listeners should produce similar clear speech intelligibility benefits for L1 and L2 participants. To test this hypothesis, the intelligibility of two clear styles are compared: hard-of-hearing-directed speech (incongruent with L2 listeners) and non-native-directed speech (congruent with L2 listeners).

1.3. The current study

The current study consists of two experiments. Experiment 1 recruits L1-English speakers and investigates the acoustic properties of their casual speech and of their two clear styles: speech to an imagined hard-of-hearing listener and speech to an imagined non-native (specifically, Mandarin-accented English) listener. Experiment 2 then assesses the intelligibility of all three styles for L1- and L2-English listeners through a speech-perception-in-noise task.

We ask two research questions. First, what are the acoustic properties of hard-of-hearing-directed and non-native-directed speech? If acoustic differences between these two clear styles are observed, then: i) they might also differ in intelligibility; and (ii) clear speech should not be considered monolithic, thus questioning its portrayal in the literature.

Second, do hard-of-hearing- and non-native-directed speech differ in intelligibility, and if so, how is the effect modulated by the listeners' language background? If the particular acoustic properties of hyper-articulated speech are targeted to benefit the intended listeners, then non-native-directed speech should be more advantageous for L2-listeners than hard-of-hearing-directed speech. Alternatively, if all forms of hyper-articulated speech are native-listener oriented (Bradlow & Bent, 2002), then because of their inexperience, L2 listeners might have a lower benefit for any kind of clear speech, irrespective of the intended listener.

2. Experiment 1: Speech production

2.1. Materials and methods

2.1.1. Participants

48 L1 speakers of California English (34 women, 13 men, 1 non-binary; mean age = 19.23 years, standard deviation (sd) = 1.55; Asian = 12, Black = 2, Hispanic/Latino = 5, Multiracial = 8, White = 15, No information provided = 6) were recruited from the University of California, Davis (UC Davis) Psychology Subjects Pool. The study was approved by the UC Davis Institutional Review Board (IRB). All participants provided informed consent and received course credit for their participation.

2.1.2. Stimuli

The stimuli consisted of 78 low-predictability sentences (taken from Kalikow et al., 1977) containing a phrase-final keyword (e.g., "Peter should speak about the mugs."). The sentence recordings were made in a sound-attenuated booth using a Shure WH20XLR head-mounted microphone and digitally sampled at a 44.1-kHz rate.

Note that a subset of these recordings was used to examine intelligibility in Experiment 2, and thus, recording low-

Table 2
The instructions and attention check questions for each speaking style in Experiment 1. The bolding reflects how the text was emphasized for participants during the experiment. Although the term "native speaker" has been problematized (Cheng et al., 2021), it was written in the instructions for participants because alternative terms, such as "L1 speaker", may not be familiar to participants.

Style	Instructions	Attention Check Questions
HOH-DS	Produce the sentence below as if you were talking to a listener who is a **native speaker of English and is hard-of-hearing**.	1. Is the listener you are imagining a native speaker of English or a non-native speaker of English? 2. Is the listener you are imagining hard-of-hearing or not hard-of-hearing?
NN-DS[4]	Produce the sentence below as if you were talking to a listener who is a **native speaker of Mandarin and is learning English**.	1. Is the listener you are imagining a native speaker of English or a non-native speaker of English?2. What is the first language of the listener you are imagining? (Options: Mandarin, English, Spanish)3. What language is the listener learning? (Options: Mandarin, English, Spanish)
Casual	Produce the sentence below **casually** as if you were talking to a listener who is a **native speaker of English**.	1. Is the listener you are imagining a native speaker of English or a non-native speaker of English? 2. Are you asked to speak casually or not casually?

predictability sentences, as opposed to high-predictability sentences, was a critical methodological choice. Prior work has suggested that speaking style intelligibility is modulated by semantic-contextual cues, with L2 listeners showing greater clear speech benefits for high-predictability stimuli than low-predictability stimuli (Bradlow & Alexander, 2007). Presenting high-predictability sentences to listeners would make it unclear whether any perceptual benefits were due to greater contextual cues or to acoustic enhancements.

2.1.3. Procedure

The participants in Experiment 1 were seated in front of a computer screen and completed the study via a self-paced Qualtrics survey. There were 3 blocks, with subjects producing 78 sentences in each block (234 total sentences). On each trial, a single sentence was displayed on the screen along with instructions for how to produce the sentence. A demographic questionnaire was administered following the recording session.

Speakers were given different speaking style instructions in each block, either producing hard-of-hearing-directed, non-native-directed, or casual speech (the exact prompts are provided in Table 2). All speaking styles were directed to imagined listeners. In the figures and tables below, hard-of-hearing-directed speech and non-native-directed speech are referred to as "HOH-DS" and "NN-DS", respectively.

The casual speech block always came last. Note that there is variation in whether casual speech is elicited first (Jung & Dmitrieva, 2023a) or last (Zellou et al., 2022; Aoki & Zellou, 2023b). The effect of block order on speech production is beyond the scope of the current study, although it has been suggested that task or block changes induce a speaking style reset, such that block order does not affect production (Lee & Baese-Berk, 2020).

Block order for the two clear styles was counterbalanced across participants. Half of the speakers recorded the hard-of-hearing-directed block before the non-native-directed block, while the other half completed the blocks in the opposite order. Subjects were thus randomly assigned to one of two possible block orders: (1) HOH-DS, NN-DS, Casual, or (2) NN-DS, HOH-DS, Casual.

Sentence order was also counterbalanced by randomly assigning participants to one of two lists. In List 1, the sentences were placed in a particular (arbitrarily decided) order, while in List 2, the sentences were placed in the reverse order (i.e., if "Peter should speak about the mugs" was the first sentence in List 1, then it was the last sentence in List 2). List assignment was consistent throughout the entire recording session, meaning that if a speaker was initially assigned to List 1, they recorded the sentences in the List 1 order for all 3 blocks.

Immediately prior to each block, subjects first read the speaking style instructions and then answered various multiple-choice questions to check their understanding, which are included in Table 2. All of the questions needed to be answered correctly in order to start producing sentences in a particular block.

2.1.4. Analysis

There were 11,232 total sentences (78 sentences x 3 styles x 48 speakers). For each speaker, every unique sentence was manually selected in Praat (Boersma & Weenink, 2021) and saved as an individual.wav file. The first author listened to each file and removed 6 sentences containing artifacts, such as yawning or coughing.

A custom-made Praat script was used to measure three acoustic variables over the duration of each sentence: mean intensity (decibels), mean fundamental frequency or f0 (Hz), and speaking rate (number of syllables divided by sentence duration; Cohn et al., 2021). These specific variables were selected because they are commonly measured in work on speaking style variation (clear speech tends to be associated with higher mean intensity, higher mean f0, and slower speaking rate than casual speech). Note that this is not an exhaustive list – the goal of the acoustic analysis is solely to establish whether the elicited speaking styles differ acoustically, using a limited number of well-known variables.

[4] Note that the instructions in the "non-native-directed" condition specifically ask speakers to imagine talking to a "native speaker of Mandarin [who] is learning English", rather than just a "non-native English speaker". All of the speakers in Experiment 1 were UC Davis (L1-English) students from California, a state with high linguistic diversity (43.9% of the state population report speaking a language other than English at home, with many languages having more than 100,000 speakers, such as "Chinese" (Mandarin or Cantonese), and Punjabi (Migration Policy Institute, 2021)). The authors initially hypothesized that the UC Davis speakers may have developed different speaking styles for different L2-listeners, and that matching the language background of the intended and actual interlocutor might enhance intelligibility (e.g., for L1-Mandarin listeners, L1-Mandarin-directed speech could be more beneficial than L1-Punjabi-directed speech, and vice versa for L1-Punjabi listeners). However, it was not possible to comprehensively address this question as the participants in Experiment 2 (the intelligibility task) were recruited from the UC Davis Psychology Subjects Pool, where the first language of the vast majority of L2-English listeners happens to be Mandarin. Acoustic and perceptual analyses of subtypes of non-native-directed speech are left for future work.

Given perceptual evidence that f0 is better measured with a logarithmic scale (Nolan, 2003), mean f0 was converted from Hertz to cents and normalized within speakers using Eq. (1) below (Jones & Munhall, 2000). In Eq. (1), F refers to the mean f0 in Hertz of a unique stimulus. B refers to the "baseline", or the grand mean f0 in Hertz of the speaker who produced F. The grand mean refers to the average f0 across all 234 stimuli (78 sentences x 3 styles).

$$\text{Cents} = 1200 * \log_2(F/B) \quad (1)$$

The data were analyzed with Bayesian mixed-effects linear regressions in R (R Core Team, 2021). All statistical models were fitted using the *brms* package (Bürkner, 2017) and *Stan* (Stan Development Team, 2023). As noted by Zellou et al. (2020): "[r]ather than dichotomous hypothesis testing based on *p*-values, Bayesian inference relies on estimating the magnitude and uncertainty of different effects estimates" (p. EL274). In other words, instead of using statistical significance and *p*-values to evaluate the models, each parameter will be interpreted with a 95% credible, or "highest density", interval. Note that "the 95% credible interval is analogous to the 95% confidence interval but with one important distinction: whereas a 95% credible interval is an interval that has a 0.95 probability of containing the value of the parameter, 95% confidence intervals are expected to contain the value of the true parameter in 95% of replications" (Gwizdzinski et al., 2023, p. 4). Effects are interpreted as meaningful or consistent if their corresponding intervals do not contain zero.

Each acoustic variable (mean intensity, mean f0, speaking rate) was assessed with a separate model. The three models contained a treatment-coded fixed effect of Style (within-subjects; casual [reference level], hard-of-hearing-directed, non-native-directed), by-speaker and by-sentence random intercepts, as well as by-speaker random slopes for Style. The model structure in R syntax is shown in Eq. (2):

$$\text{Variable} \sim \text{Style} + (1 + \text{Style} \mid \text{Speaker}) + (1 \mid \text{Sentence}) \quad (2)$$

To directly test whether hard-of-hearing-directed and non-native-directed speech differ from each other, three additional post-hoc models were also fit (one for each acoustic variable). The post-hoc models had the same structure as in Eq. (2), except that casual speech was removed from the analysis. This meant that the fixed effect of Style only had 2 levels (within-subjects; hard-of-hearing-directed [reference level], non-native-directed).

The prior distributions in R syntax for the non-intercept fixed effects ("b"), the standard deviation of the random intercepts ("sd"), the correlation between random parameters ("cor"), nu (a parameter of the Student's t distribution), and the random error ("sigma") are as follows: student_t(3, 0, 12), student_t (3, 0, 12), lkj_corr_cholesky(2), gamma(2, 0.1), and student_t (3, 0, 12). Note that "student_t", "lkj_corr_cholesky", and "gamma" refer to the Student's t, Cholesky LKJ Correlation, and gamma distributions, respectively. The prior distributions for the intercepts of the mean intensity, mean f0, and speaking rate models were: student_t(3, 50, 12), student_t(3, 0, 12), student_t(3, 3.8, 1). Note that the mean of the prior distribution for mean f0 is set to 0 due to normalization in cents.

2.2. Results

The marginal posterior distributions of the fixed effects are shown for all acoustic variables in Fig. 1. The comprehensive specifications of the models and the full output can be found in the supplementary material (see the Data Statement).

First, compared to casual speech, hard-of-hearing-directed speech has greater mean intensity [β: 3.00, sd: 0.46, 95% highest density interval (HDI) = (2.08, 3.88)], marginally (but not meaningfully) higher mean f0 [β: 16.22, sd: 10.46, 95% HDI = (−2.55, 38.09)], and a consistently slower speaking rate [β: −1.11, sd: 0.09, 95% HDI = (−1.28, −0.94)].

Second, relative to casual speech, non-native-directed speech has numerically (but not meaningfully) greater mean intensity [β: 0.45, sd: 0.39, HDI = (−0.31, 1.26)], consistently lower mean f0 [β: –33.72, sd: 9.46, HDI = (−51.84, −14.89)], and a slower speaking rate [β: −1.45, sd: 0.11, HDI = (−1.65, −1.23)] compared to casual speech.

Finally, the post-hoc models indicate that hard-of-hearing-directed speech and non-native-directed speech are consistently different from each other on every acoustic measure. Non-native-directed speech has lower mean intensity [β: −2.48, sd: 0.47, HDI = (−3.39, −1.57)], lower mean f0 [β: −54.12, sd: 12.14, HDI = (−76.94, −29.49)], and a slower speaking rate [β: −0.34, sd: 0.08, HDI = (−0.50, −0.17)] relative to hard-of-hearing-directed speech.

2.3. Interim discussion

There are two main takeaways from Experiment 1. First, the acoustic modifications of speech to an imagined hard-of-hearing listener and speech to an imagined non-native (Mandarin-accented English) listener are, for the most part, consistent with previous work on clear speech. Compared to casual speech, hard-of-hearing-directed speech has greater mean intensity, marginally greater mean f0, and a slower speaking rate, while non-native-directed speech has marginally greater mean intensity and a slower speaking rate (but also lower mean f0, which is inconsistent with the clear speech literature). Taken together, these results generally fall in line with H&H Theory, which predicts that, to maximize the likelihood of being understood, speakers will hyper-articulate when talking to listeners who may have difficulty understanding them.

However, we also find that clear speech is not monolithic. Non-native-directed speech has lower mean intensity, lower mean f0, and a slower speech rate than hard-of-hearing-directed speech. Thus, the Experiment 1 speakers made targeted acoustic enhancements based on the specific communicative context.

3. Experiment 2: Intelligibility

Given that hard-of-hearing and non-native-directed speech are acoustically different from each other, an empirical question is whether they differ in intelligibility. Experiment 2 addresses this question using a speech-transcription-in-noise task. Critically, if the particular acoustic modifications of non-native-directed speech are helpful for the intended interlocu-

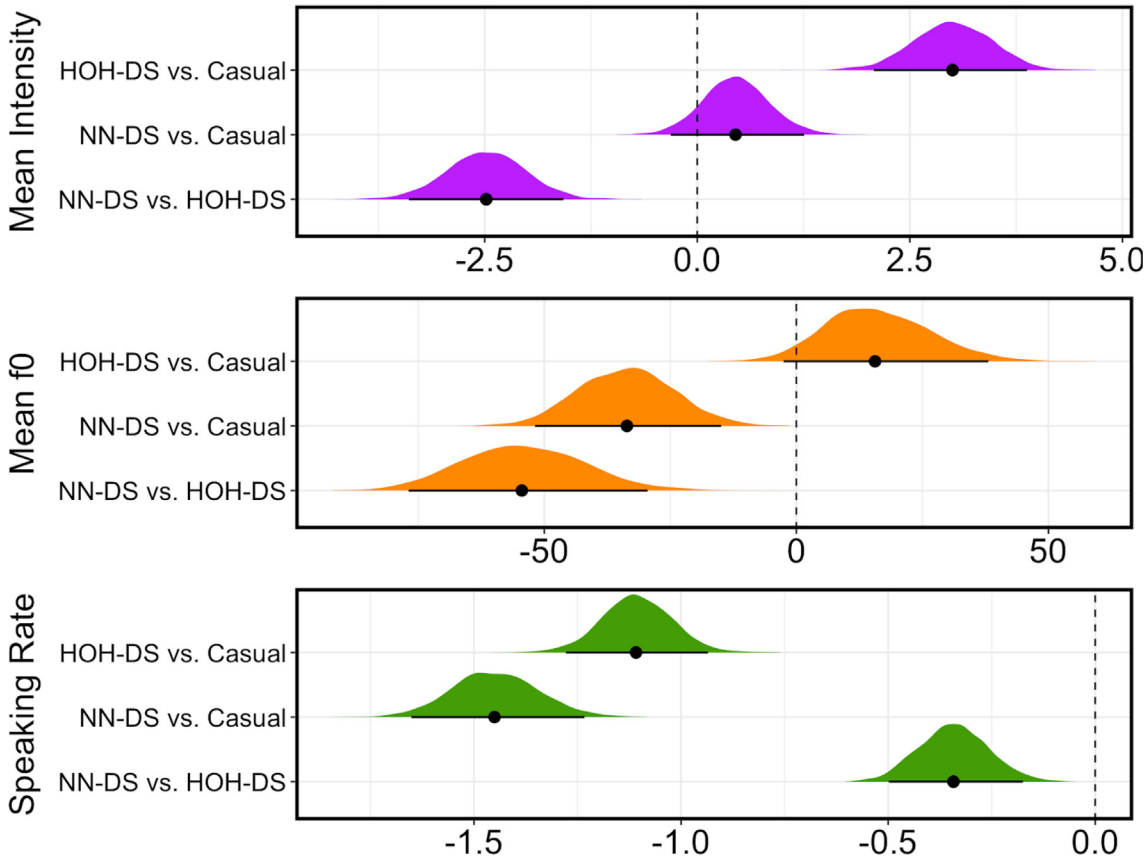

Fig. 1. Marginal posterior distributions for the fixed effects in the Experiment 1 models (upper: mean intensity (dB); middle: mean f0 (cents); lower: speaking rate (syllables/second)). Within each plot, the top distribution (hard-of-hearing-directed versus casual speech) and the center distribution (non-native-directed versus casual speech) are derived from the main models. The bottom distribution (non-native-directed versus hard-of-hearing-directed speech) comes from the post-hoc models. The line segments below each distribution reflect the 95% highest density intervals, with circles at the mean. The dotted, vertical lines are placed at zero. Greater hyper-articulation is indicated by more positive values for mean intensity, more positive values for mean f0, and more negative values for speaking rate.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tors, then L2 listeners should benefit more from non-native-directed than hard-of-hearing-directed speech.

3.1. Materials and methods

3.1.1. Stimuli

Stimuli from 4 talkers in Experiment 1 (out of 48 total) were selected for the intelligibility experiment, meaning that 936 sound files were used in total (78 sentences x 3 styles x 4 talkers). The criteria and rationale for selecting these specific talkers is detailed in the Appendix.

All recordings were set to a presentation level of 65 dB SPL using the "Scale intensity…" function in Praat (Boersma & Weenink, 2021). This rescaling procedure follows prior intelligibility experiments (e.g., Bradlow & Bent, 2002), and neutralizes mean intensity differences between speaking styles (but without affecting the intensity contours). Any intelligibility differences between hard-of-hearing and non-native-directed speech should thus be viewed as conservative. Given its greater mean intensity (see Fig. 1), hard-of-hearing-directed speech could potentially be even more intelligible than non-native-directed speech without intensity rescaling.

After rescaling, speech-shaped noise was created (Winn, 2019). The stimuli were mixed with noise at a + 2 dB signal-to-noise ratio (McCloy, 2015), with noise commencing 500 ms prior to sentence onset and ending 500 ms after sentence offset. Note that signal-to-noise ratio (SNR) is inconsistent across intelligibility studies (e.g., 0 dB in Jung & Dmitrieva (2023b)), and that some experiments present stimuli at varying SNRs for distinct listener groups (e.g., −4 dB for L2-English listeners and −8 dB for L1-English listeners in Bradlow & Bent (2002)). The specific SNR of + 2 dB in the current experiment was selected through pilot testing, which showed that for the initial listeners (5 L1-English, 5 L2-English), transcription accuracy was neither at ceiling nor at floor. Furthermore, L1 and L2 listeners were exposed to stimuli with the same SNR in the current study. As noted by Kato and Baese-Berk (2023), presenting different SNRs for different listener groups can introduce a "potential confound" (p. 20), because it can become unclear whether any observed differences between L1 and L2 listeners result from group identity or from SNR. The role of SNR on intelligibility is beyond the scope of the current study and can be explored in future work.

3.1.2. Participants

101 participants (35 L1-English, 66 L2-English) completed the experiment, none of whom had participated in Experiment 1. All were undergraduates recruited from the UC Davis Psychology Subjects Pool, provided informed consent, and

received course credit for their participation. The study was approved by the UC Davis IRB.

To maximize the difference between the two listener groups, any L1 listener who self-reported exposure to a non-English language during childhood (n = 5) and any L2 listener who self-reported exposure to English during childhood (n = 31) was removed from the analysis. Additionally, 4 subjects were removed for self-reporting a hearing difficulty, 1 subject was omitted for responding with the same word on nearly every trial, and 1 subject was taken out for being significantly older than the typical undergraduate population (38-years-old).

The final analysis, which approximately matched the sample size of Bradlow and Bent (2002), included 59 listeners, all of whom self-reported typical hearing. There were 29 L1 listeners (22 women, 7 men; mean age = 20.17 years, sd = 2.10; Asian = 8, Black = 2, Multiracial = 6, Native Hawaiian or Pacific Islander = 2, White = 11) and 30 L2 listeners (23 women, 7 men; mean age = 20 years, sd = 1.44; Asian = 27, Hispanic/Latino = 3). All of the L1-English subjects were specifically California-English speakers. Among the L2 listeners, most reported their first language as Mandarin or "Chinese" (n = 24), and several others stated either Spanish (n = 3) or Vietnamese (n = 3). The self-reported number of years learning English was collected as a measure of English experience, and this varied widely across participants (mean = 9.6 years, sd = 3.6, range = between 3 and 20 years).

3.1.3. Procedure

Participants completed the self-paced experiment online via a Qualtrics survey. Similar to Experiment 1, a demographic questionnaire was administered following the main task.

Each participant completed 78 trials, one trial for each unique sentence. On a given trial, a single sentence was presented auditorily, and subjects were instructed to type the last word that they heard into a text box (e.g., "Peter should speak about the mugs."). Participants heard each stimulus once, and the stimulus presentation order was randomized. Each listener only heard stimuli from one randomly assigned speaker.

To counterbalance sentence content and speaking style, listeners were randomly assigned to one of three possible lists: List 1 (sentences 1–26 presented in HOH-DS; sentences 27–52 presented in NN-DS; and sentences 53–78 in Casual speech), List 2 (27–52 in HOH-DS; 53–78 in NN-DS; 1–26 in Casual) or List 3 (53–78 in HOH-DS; 1–26 in NN-DS for; 27–52 in Casual). Across the lists, the sentences were evenly divided across casual, hard-of-hearing-directed, and non-native-directed speech. As noted earlier, stimulus presentation order was fully randomized across all 78 sentences (i.e., not presented in blocks for each style).

3.1.4. Analysis

Final keyword transcription accuracy was coded binarily as correct (1) or incorrect (0). Any responses transcribed without the correct affixes were coded as incorrect (e.g., "mug" was coded as incorrect if the right answer was "mugs"). Spelling mistakes were manually corrected ("sleaves" for "sleeves", "broose" for "bruise", "weet" for "wheat"), and homonyms were scored as correct ("brews" for "bruise", "heard" for "herd", "hey" for "hay", "greece" for "grease").

Similar to Experiment 1, the data were modeled with Bayesian mixed-effects logistic regressions in R through the *brms* package and *Stan*. The main model included fixed effects of Style (within-subjects; casual, hard-of-hearing-directed, non-native-directed), and Listener Group (between-subjects; L1, L2), as well as their interaction. Since this model contained an interaction term, all fixed effects were sum-coded "to allow the interpretation of lower order effects in the models as main effects rather than simple effects" (McGowan, 2015, p. 511). By-listener, by-speaker, and by-sentence random intercepts were added, along with by-listener random slopes for Style. The model structure in R syntax is shown in Eq. (4):

$$\begin{aligned}\text{Accuracy} \sim\ & \text{Style} * \text{Listener\_Group} + (1 + \text{Style} \mid \text{Listener}) \\ & + (1 \mid \text{Sentence}) + (1 \mid \text{Speaker}) \end{aligned} \quad (3)$$

Speaking style intelligibility for L2 listeners is potentially affected by listener proficiency (Jung & Dmitrieva, 2023b), and as mentioned in Section 3.1.2, there is high variability across the L2 participants in the reported number of years learning English, a correlate of L2 proficiency (Piske et al., 2001). To assess the effect of English experience on the results, a post-hoc Bayesian mixed-effects logistic regression model was fitted that only included the responses from the L2 listeners. The post-hoc model contained a fixed effect of Style (within-subjects; casual, hard-of-hearing-directed, non-native-directed), a scaled and centered fixed effect of English Experience (the self-reported number of years learning English), and their interaction. All fixed effects were sum-coded. The random effects structure was the same as the main model, except that by-Listener random slopes for English Experience and for the interaction between Style and English Experience were additionally included. The post-hoc model structure in R syntax is shown in Equation 5:

$$\begin{aligned}\text{Accuracy} \sim\ & \text{Style} * \text{English\_Experience} + (1 + \text{Style} \\ & * \text{English\_Experience} \mid \text{Listener}) + (1 \mid \text{Sentence}) \\ & + (1 \mid \text{Speaker}) \end{aligned} \quad (4)$$

Following the logistic regression models in Barreda and Silbert (2023), the prior distributions in R syntax for the Intercept, non-Intercept fixed effects ("b"), the standard deviation of the random intercepts ("sd"), and the correlation between random parameters ("cor") are as follows: student_t(3, 0, 3), student_t(3, 0, 3), student_t(3, 0, 3), and lkj_corr_cholesky(2).

3.2. Results

Fig. 2 shows the marginal posterior distributions of the fixed effects in the main model. The comprehensive specifications of the models and the full output can be found in the supplementary material (see the Data Statement section).

The main model first revealed a meaningful main effect of Listener Group [β: 0.63, sd: 0.07, 95% HDI = (0.49, 0.78)], where L1 listeners (45.8% correct) have higher transcription accuracy than L2 listeners (27.6% correct).

Aggregating across listener groups, hard-of-hearing-directed speech had the highest accuracy (40.2%), followed by non-native-directed speech (36.2%), and casual speech (33.2%). According to the main model (specifically, the sum-coded main effects of Style), the intelligibility benefit of

hard-of-hearing-directed speech was meaningful [β: 0.22, sd: 0.06, 95% HDI = (0.11, 0.33)], in contrast to non-native-directed speech, which did not have a meaningful effect compared to the mean [β: −0.02, sd: 0.05, 95% HDI = (−0.13, 0.08)]. The lack of consistent interactions indicated that the effect of hard-of-hearing-directed speech [β: 0.05, sd: 0.06, 95% HDI = (−0.06, 0.16)] and of non-native-directed speech [β: 0.0004, sd: 0.06, 95% HDI = (−0.11, 0.11)] was the same for both L1- and L2-listeners.

In the post-hoc model focusing solely on the results for L2 listeners, the main effect of Style has the same interpretation as in the main model. Hard-of-hearing-directed speech results in a meaningful intelligibility benefit [β: 0.19, sd: 0.09, 95% HDI = (0.004, 0.37)], unlike non-native-directed speech [β: −0.0006, sd: 0.09, 95% HDI = (−0.17, 0.18)]. Critically, overall task performance does not vary by English experience [β: 0.06, sd: 0.16, 95% HDI = (−0.26, 0.35)]. There was also no evidence of any interactions, meaning that the effects of hard-of-hearing-directed speech [β: 0.002, sd: 0.11, 95% HDI = (−0.21, 0.22)] and non-native-directed speech [β: 0.01, sd: 0.10, 95% HDI = (−0.19, 0.21)] are not modulated by English experience.

As noted in Section 1.3, a central hypothesis of the current study is that aligning the intended and actual interlocutor (i.e., presenting L2-listeners with non-native-directed speech) should enhance intelligibility. However, in the current study, non-native-directed speech could potentially have represented a mismatch for certain L2-listeners. Speakers in Experiment 1 were asked to imagine talking to a "native speaker of Mandarin", even though in Experiment 2, certain L2-listeners were not L1-Mandarin listeners (Vietnamese: n = 3, Spanish: n = 3). To determine whether L2-listener language background affected the results, both the main and post-hoc models were fitted with only the L1-Mandarin listeners (n = 24). The interpretation of all fixed effects remained the same (see the Data Statement for more information).

4. General discussion

The current study investigated the acoustic properties of casual speech and two clear styles (hard-of-hearing-directed and non-native-directed speech) and evaluated their intelligibility in noise for L1 and L2 listeners. There were three primary findings. First, although the two clear styles are both hyperarticulated compared to casual speech, they are also markedly distinct acoustically, with non-native-directed speech having lower mean intensity, lower mean f0, and a slower speaking rate than hard-of-hearing-directed speech. Second, for both L1 and L2 listeners, the two clear styles are also perceptually different – in contrast to non-native-directed speech, only hard-of-hearing-directed speech enhances intelligibility compared to casual speech. Note that the benefit of hard-of-hearing-directed speech cannot be due to its greater mean intensity, given that all of the speech stimuli were rescaled to the same mean intensity in Experiment 2. Third, contrary to prior work (Bradlow & Bent, 2002; Bradlow & Alexander, 2007), L1 and L2 listeners did not differ in the effect of speaking style on intelligibility.

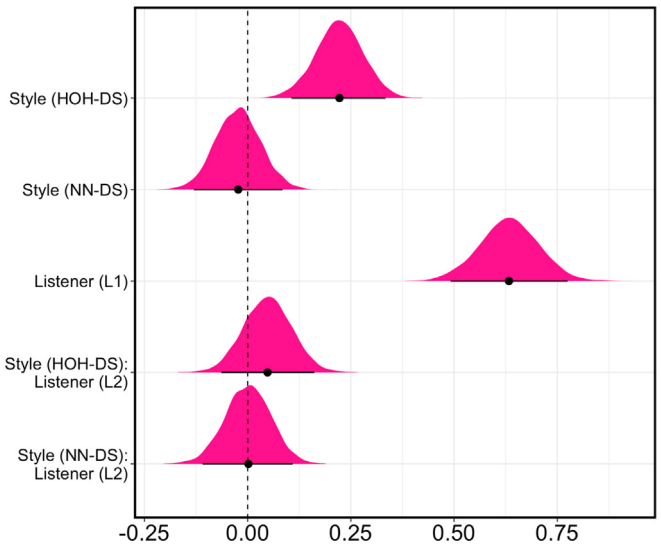

Fig. 2. Marginal posterior distributions for the fixed effects in the main model of Experiment 2. The line segments below each distribution reflect the 95% highest density intervals, with circles at the mean. The dotted, vertical line is placed at zero (note that the statistical model for Experiment 2 is sum-coded, meaning that zero reflects the mean, not a particular reference level). More positive values indicate greater transcription accuracy, while more negative values reflect lower transcription accuracy.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.1. Being clear about clear speech

The results in the current study build upon prior research by underscoring that clear speech is not a singular speaking style (Scarborough & Zellou, 2013). Rather, clear speech consists of a wide variety of distinct styles (e.g., hard-of-hearing-directed and non-native-directed speech) that happen to be acoustically exaggerated with the presumed intention of increasing listener comprehension.

The acoustic and perceptual divergence of clear speech subtypes leads to an important implication: that results found for one type of clear speech might not replicate if another type is elicited. If two studies have a clear speech condition, but have different elicitation instructions (e.g., hard-of-hearing-directed speech in Kang & Guion (2008) and non-native-directed speech in Jung & Dmitrieva (2023a)), then they are not directly comparable. Even if both elicit "clear speech", the differences in elicitation method could result in acoustic or perceptual disparities.

In general, substituting specific styles with the generic term "clear speech" (a common practice throughout the literature, including the authors' own prior work; see Table 1) erases variation. This practice potentially amplifies a misconception that clear speech is a singular speaking style, rather than a series of distinct styles. Furthermore, it can lead to an assumption that findings from one style will transfer to all types of clear speech, which can then result in confusion within the literature (e.g., the comparison between Kang & Guion (2008) and Jung & Dmitrieva (2023a), as discussed in Section 1.1).

To avoid misconceptions, we advocate for clarity when discussing "clear speech" and propose that greater caution should be taken when interpreting results on speaking style experiments. For example, the current study cannot address

any sweeping claims about clear speech in general – we only shed light on the properties of three specific speaking styles to imagined listeners based on the particular instructions given in Table 2. The results could have been different had other clear styles been recorded (e.g., device-directed speech; Cohn et al., 2022), had the speech been directed to real listeners (Scarborough & Zellou, 2013), or had different instructions been given. Being more *clear* when discussing clear speech and constraining the interpretation of results could help to mitigate confusion and emphasize that clear speech is not monolithic.

To be clear, however, we do not believe that the phrase "clear speech" should be abandoned. Speaking styles like hard-of-hearing-directed and non-native-directed speech do share some similar acoustic properties (e.g., both are slower than casual speech) and ostensibly have overlapping goals (e.g., facilitating communication with listeners who may have trouble understanding the speech signal). "Clear speech" can therefore be useful as an umbrella term when summarizing the literature. Referring to "clear speech" could also be appropriate if a generic version has been elicited and there is no other suitable term for describing the elicited speaking style (e.g., "Speak clearly to someone who is having a hard time understanding you"; Zellou et al., 2022, p. 3433). Nevertheless, regardless of the specificity of elicitation instructions, we suggest that the interpretation of speaking style experiments should still be delimited, acknowledging that even results for a generic version of clear speech cannot necessarily be directly compared to other types of clear speech.

If clear speech is not monolithic, then a natural follow-up question is whether there are also distinct subtypes of casual, hard-of-hearing-directed, or non-native-directed speech. For example, in our non-native-directed condition, speakers in Experiment 1 were asked to imagine "a native speaker of Mandarin [who] is learning English", but perhaps the acoustic and perceptual effects would have been altered had the talkers imagined a native speaker of Spanish who is proficient in English. As mentioned in the introduction, many works within the literature only focus on the clear-casual dichotomy, ostensibly because of the dominance of H&H Theory. While we acknowledge the usefulness of H&H Theory, it is also a theoretical heuristic that drastically simplifies speaking style variation as a continuum between two articulatory settings. Although we do not claim that every study should examine more than two speaking styles, a key step for future work is to expand beyond the binary of clear and casual speech, which could help achieve a more nuanced understanding of speaking style variation.

4.2. Do L2 listeners benefit less from clear speech?

Even though the current study did not find a difference between L1 and L2 listeners in the effect of speaking style on intelligibility, this does not necessarily contradict the proposal by Bradlow and Bent (2002) that L1 listeners are better at leveraging language-specific acoustic modifications than L2 listeners. A general limitation of many perceptual experiments on speaking style is that acoustic features are often not controlled. When talkers are asked to produce different styles, many variables are simultaneously modified (e.g., higher mean intensity, higher mean f0, vowel space expansion, etc.), and it is not known precisely which acoustic variables are responsible for perceptual enhancements. Individual differences in the acoustic properties of speaking styles are well-documented (Ferguson & Kewley-Port, 2007), and in the current study, perhaps the specific talkers in the intelligibility experiment happened to enhance language-general features much more than language-specific variables, resulting in a similar benefit for L1 and L2 listeners.

Another way of addressing Bradlow and Bent (2002) is to more carefully manipulate stimulus acoustics. A potential speech-transcription-in-noise task could present four types of words to L1 and L2 listeners: (1) unmodified [control]; (2) only modified with a language-general hyper-articulation strategy; (3) only modified with a language-specific hyper-articulation strategy; (4) modified with both a language-general and a language-specific hyper-articulation strategy. The language-general strategy could be any acoustic property that speakers modulate regardless of their language background (e.g., lengthened duration; Smiljanić & Bradlow, 2005). The language-specific strategy would depend on which listeners are recruited. If, inspired by Smiljanić and Bradlow (2008), L1-English and L1-Croatian/L2-English listeners were selected, then the English-specific strategy could be lengthened voice-onset-time in voiceless stop consonants. According to Bradlow and Bent (2002), L1-listeners should show enhanced transcription accuracy for all modified words compared to the unmodified control words, whereas L2-listeners should not benefit (or should have a reduced benefit) from Condition (3), where words are only modified with a language-specific hyper-articulation strategy.

The current study also found, contrary to our expectations, that L2 listeners benefited more from hard-of-hearing-directed speech (where there was a mismatch between the actual and intended interlocutor) than from non-native-directed speech. However, non-native-directed speech also did not enhance intelligibility for L1-listeners. This suggests that the acoustic properties of the non-native-directed speech that we elicited are generally unhelpful, and does not rule out the hypothesis that aligning the intended and actual interlocutor could facilitate perception.

All speech productions in the current study were directed to imagined listeners, so one possibility is that speech directed toward real L2 listeners might be more perceptually beneficial. Talkers tend to shift their speech based on listener feedback (Buz et al., 2016), so in the absence of feedback, the speakers in the current study had no opportunity to evaluate whether their productions were helpful for the listener. However, if participants are asked to interact with a real listener, they might be able to more effectively modulate their speaking style, as suggested by Scarborough & Zellou (2013). More broadly, additional work comparing the intelligibility of real-listener-directed and imagined-listener-directed speech would help to further our understanding of speaking style perception beyond the binary of clear and casual speech.

4.3. Limitations, future directions & implications for text-to-speech development

There are several additional avenues for further research. First, the intelligibility experiment in the current study only

presented stimuli from 4 speakers, out of the 48 total talkers in Experiment 1. Prior work has shown that talkers can vary greatly both in the acoustic properties (Wright et al., 2023) and the relative intelligibility of their speaking styles (e.g., in Payton et al., (1994), the clear speech intelligibility benefit was much larger than the benefit of hard-of-hearing-directed speech in this study). Future work should explore how intelligibility is affected by between-speaker differences in speaking style acoustic properties.

Similar to the vast majority of prior work, the current study only examines productions from L1-English speakers, so another future direction is to examine the acoustic and perceptual properties of different clear styles for L2 speakers and non-English speakers. Although there is some extant work that investigates speaking styles in L2-speech (Kato & Baese-Berk, 2023; Jung & Dmitrieva, 2023a) and in non-English languages (Kang & Guion, 2008; Zellou et al., 2022), all of these studies only compare clear speech to casual speech as a binary. More work is needed to evaluate different types of clear styles across accents and languages.

Finally, the current study has potential implications for the development of text-to-speech (TTS) voices. TTS voices are often less intelligible than naturally produced voices (Aoki et al., 2022), and researchers have developed hyper-articulated TTS speaking styles to help resolve this issue (e.g., Cohn & Zellou, 2020; Xiao et al., 2022). Work on TTS style development often draws upon research on the intelligibility of naturally produced speaking styles to determine which acoustic features to manipulate (Raitio et al., 2022). If certain varieties of naturally produced clear speech are not useful for listeners, such as non-native-directed speech, understanding which acoustic properties underlie this effect can help TTS developers create styles that are the most accessible and user-friendly.

5. Conclusion

Extensive research has found that, relative to one's default speaking style, clear speech contains exaggerated acoustic modifications and is often perceptually beneficial. However, clear speech encompasses a wide variety of styles, and in the literature (including the authors' own prior work), the term "clear speech" is often employed in an ambiguous and *unclear* way. Specifically, there is a tendency to use the generic term "clear speech" when interpreting results, regardless of what subtype of clear speech has been elicited (e.g., hard-of-hearing-directed speech, non-native-directed speech). This practice tacitly assumes that the intended interlocutor is inconsequential, and has led to a portrayal in certain studies of clear speech as homogeneous, with findings found for one type of clear speech assumed to generalize to any subtype.

The current study highlights that clear speech is not monolithic and emphasizes the importance of the intended interlocutor in speaking style elicitation. Casual speech and two clear styles (hard-of-hearing-directed speech and non-native-directed speech) were compared through an acoustic analysis and a speech-perception-in-noise task. Not only are the two clear styles acoustically different (non-native-directed speech has lower mean intensity, lower mean f0, and a slower speaking rate), but they are also perceptually different (for both L1 and L2 listeners, only hard-of-hearing-directed speech enhances intelligibility). To avoid potential confusion, the authors advocate for being more *clear* about clear speech – we suggest that in any speaking style experiment, it should be clarified that the results only speak to the specific variety of clear speech that was elicited and may not necessarily generalize to all forms of clear speech.

Declaration of interest

None.

Author statement

Nicholas B. Aoki and Georgia Zellou are the sole authors of this study and are responsible for completing the entire study.

CRediT authorship contribution statement

Nicholas B. Aoki: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Georgia Zellou:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Statement

The data and scripts used to perform the analyses and generate the graphs are available on the Open Science Framework (OSF) at the following link: https://doi.org/10.17605/osf.io/f5pdb.

Appendix

As noted in Section 3.1.1, stimuli from 4 talkers (out of 48 total) from Experiment 1 were presented to listeners in Experiment 2. The goal of the appendix is to provide details about how these 4 talkers were chosen.

Following Scarborough and Zellou (2013), individual variation among the Experiment 1 speakers was examined, with the goal of finding talkers that mirrored the broad acoustic patterns in Section 2.2. Note that no speakers exactly paralleled the aggregated acoustic results, and thus, the talkers in Experiment 2 were chosen as the most optimal representatives.

To examine individual acoustic variation, a separate Bayesian mixed-effects linear regression model was fitted for each acoustic variable (mean intensity, mean f0, speaking rate) for each subject in Experiment 1. The analysis was conducted in R (R Core Team, 2021) through the *brms* package (Bürkner, 2017) and *Stan* (Stan Development Team, 2023). All models had a treatment-coded fixed effect of Style and a by-Sentence random intercept, as shown in R syntax in Equation (A.1). Unlike the Experiment 1 models (see Section 2.1.4, Equation (2), by-Speaker random effects could not be included because each model only contained data for one speaker.

$$\text{Variable} \sim \text{Style} + (1 \mid \text{Sentence}) \quad \text{(A.1)}$$

There were two sets of models, akin to the Experiment 1 analysis. In Set 1 (48 speakers x 3 acoustic variables = 144 models), each model contained 234 observations (78 sentences x 3 styles), and Style had 3 levels (within-subjects; casual [reference], hard-of-hearing-directed, non-native-directed). The Set 2 models (48 speakers x 3 acoustic variables = 144 models) contained 156 observations (78 sentences x 2 styles), and Style had 2 levels (within-subjects; hard-of-hearing-directed [reference], non-native-directed). Set 1 thus compared casual speech to hard-of-hearing-directed speech and to non-native-directed speech, while Set 2 directly compared hard-of-hearing-directed to non-native-directed speech.

Consistent with prior work (Ferguson & Kewley-Port, 2007), high between-speaker acoustic variability was observed in the magnitude of the effects. Critically, however, the majority aligned with the direction of the results in Section 2.2 (e.g., all speakers had a meaningfully slower speaking rate in hard-of-hearing-directed and non-native-directed speech compared to casual speech). All of the individual-level models can be found in the supplementary material (see the Data Statement).

Using the output from these individual-level models, the final 4 speakers in Experiment 2 were found via a 10-step, backward-selection process, which is summarized in Table A.1 for clarity. The first 9 steps cover each of the main findings in Section 2.2, but note that the criteria in Steps 1–3 (the comparisons of hard-of-hearing-directed and non-native-directed speech) are more stringent, since the current study is primarily interested in comparing the intelligibility of the two clear styles. Steps 4–6 (comparisons of hard-of-hearing-directed and casual speech) and Steps 7–9 (comparisons of non-native-directed and casual speech) were necessarily more lenient, since otherwise, it would have been impossible to select any speakers. Among the 10 remaining speakers after Step 9, 4 talkers were randomly selected so that the number of listeners hearing each talker in Experiment 2 matched Jung and Dmitrieva (2023b) (approximately 8 L1 listeners and 8 L2 listeners).

Table A.1

The backward-selection process used to select the final 4 speakers presented in Experiment 2. The process started with 48 total speakers (Step 0) and narrowed down the pool based on the selection criteria (i.e., out of the 48 initial speakers, 33 complied with the Step 1 requirement; then among the 33 remaining speakers, 26 complied with Step 2, and so on). Within the selection criteria, "meaningful" implies that the highest-density intervals did not contain zero. "Numerically higher" implies that the mean estimate was higher than zero and that the highest-density intervals may or may not have contained zero ("numerically lower" is defined the same way, except that the mean estimate was lower than zero).

Step	Selection Criterion	Speakers Left
0	----------	48
1	Meaningfully lower mean intensity in NN-DS than HOH-DS	33
2	Meaningfully lower mean f0 in NN-DS than HOH-DS	26
3	Meaningfully slower speaking rate in NN-DS than HOH-DS	19
4	Numerically higher mean intensity in HOH-DS than casual speech	18
5	Numerically higher mean f0 in HOH-DS than casual speech	16
6	Numerically slower speaking rate in HOH-DS than casual speech	16
7	Numerically higher mean intensity in NN-DS than casual speech	9
8	Numerically lower mean f0 in NN-DS than casual speech	7
9	Numerically slower speaking rate in NN-DS than casual speech	7
10	Random selection of 4 speakers	4

The acoustic analysis for the final 4 speakers is shown in Fig. 3 and contrasted with the full model (containing data for all speakers) reported in Section 2.2. For all 4 speakers: (1) hard-of-hearing-directed speech has meaningfully higher mean intensity, higher mean f0, and slower speaking rate than casual speech; (2) non-native-directed speech has meaningfully higher mean intensity, either numerically or meaningfully lower mean f0, and slower speaking rate than casual speech; (3) non-native-directed speech has meaningfully lower mean intensity, lower mean f0, and slower speaking rate than hard-of-hearing-directed speech. Although the magnitude of the effects differ across the individual talkers, they are all in the same direction as the full model effects, and were deemed as optimal speakers to present to listeners in Experiment 2.

Fig. A1. 95% highest density intervals (with circles at the mean) for the 4 speakers presented in Experiment 2 (in blue; initials on the y-axis) and for the full models combining data from all speakers (in red; taken directly from Fig. 1 in Section 2.2). The 9 plots cross the 3 acoustic variables (mean intensity: top row, dB; mean f0: middle row, cents; speaking rate: bottom row, syllables/second) and the 3 style comparisons (hard-of-hearing-directed versus casual speech: left column; non-native-directed versus casual speech: middle column; non-native-directed versus hard-of-hearing-directed speech: right column). The dotted, vertical line is placed at zero, and effects are considered "meaningful" if intervals do not cross the dotted line. Greater hyper-articulation is indicated by more positive values for mean intensity, more positive values for mean f0, and more negative values for speaking rate.

(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

References

Aoki, N. B., Cohn, M., & Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise. *JASA Express Letters, 2*(4). https://doi.org/10.1121/10.0010274 045204.

Aoki, N. B., & Zellou, G. (2023a). When clear speech does not enhance memory: Effects of speaking style, voice naturalness, and listener age. *Proceedings of Meetings on Acoustics, 51*(1). https://doi.org/10.1121/2.0001766 060002.

Aoki, N., & Zellou, G. (2023). Speakers talk more clearly when they see an East Asian face: Effects of visual guise on speech production. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 2294–2298). Guarant International.

Aoki, N. B., & Zellou, G. (2023b). Visual information affects adaptation to novel talkers: Ethnicity-specific and -independent learning of L2-accented speech. *The Journal of the Acoustical Society of America, 154*(4), 2290–2304. https://doi.org/10.1121/10.0021289.

Barreda, S., & Silbert, N. (2023). *Bayesian Multilevel Models for Repeated Measures Data: A Conceptual and Practical Introduction in R*. Routledge.

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer (Version 6.1.40) [Computer program]. https://www.fon.hum.uva.nl/praat/.

Bradlow, A. R. (2002). Confluent talker- and listener-related forces in clear speech production. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 237–274). Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197105.1.241.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272–284. https://doi.org/10.1121/1.1487837.

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America, 121*(4), 2339–2349. https://doi.org/10.1121/1.2642103.

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28 10.18637/jss.v080.i01.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language, 89*, 68–86. https://doi.org/10.1016/j.jml.2015.12.009.

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and Harmful terminology with inclusive and accurate measures. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.715843 715843.

Cohn, M., & 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* (pp. 1733–1737).

Cohn, M., Pycha, A., & Zellou, G. (2021). Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech. *Cognition, 210*. https://doi.org/10.1016/j.cognition.2020.104570 104570.

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics, 90*. https://doi.org/10.1016/j.wocn.2021.101123 101123.

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 112*(1), 259–271. https://doi.org/10.1121/1.1482078.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: Acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research, 50*(5), 1241–1255. https://doi.org/10.1044/1092-4388(2007/087).

Gwizdzinski, J., Barreda, S., Carignan, C., & Zellou, G. (2023). Perceptual identification of oral and nasalized vowels across American English and British English listeners and TTS voices. *Frontiers in Communication, 8*, 1307547. https://doi.org/10.3389/fcomm.2023.1307547.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America, 108*(3), 1246–1251. https://doi.org/10.1121/1.1288414.

Jung, Y.-J., & Dmitrieva, O. (2023a). Acoustic properties of non-native clear speech: Korean speakers of English. *Speech Communication, 154*. https://doi.org/10.1016/j.specom.2023.102982 102982.

Jung, Y.-J., & Dmitrieva, O. (2023b). Non-native talkers and listeners and the perceptual benefits of clear speech. *The Journal of the Acoustical Society of America, 153*(1), 137–148. https://doi.org/10.1121/10.0016820.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351. https://doi.org/10.1121/1.381436.

Kang, K.-H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *The Journal of the Acoustical Society of America, 124*(6), 3909–3917. https://doi.org/10.1121/1.2988292.

Kato, M., & Baese-Berk, M. (2023). The Effects of Acoustic and Semantic Enhancements on Perception of Native and Non-Native Speech. https://doi.org/10.1177/00238309231156615.

Lam, J., & Tjaden, K. (2013). Intelligibility of Clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research, 56*(5), 1429–1440. https://doi.org/10.1044/1092-4388(2013/12-0335).

Lee, D.-Y., & Baese-Berk, M. M. (2020). The maintenance of clear speech in naturalistic conversations. *The Journal of the Acoustical Society of America, 147*(5), 3702–3711. https://doi.org/10.1121/10.0001315.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Springer. https://doi.org/10.1007/978-94-009-2037-8_16.

McCloy, D. (2015). Mix Speech with Noise [Praat script]. https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502–521. https://doi.org/10.1177/0023830914565191.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America, 147*(2), EL151–EL156. https://doi.org/10.1121/10.0000718.

Migration Policy Institute (2021). *California*. Available online at: https://www.migrationpolicy.org/data/state-profiles/state/language/CA (accessed January 3, 2024).

Nolan, F. (2003). Intonational equivalence: An experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 771–774).

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America, 95*(3), 1581–1592. https://doi.org/10.1121/1.408545.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134.

R Core Team. (2021). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

Raitio, T., Petkov, P., Li, J., Shifas, M., Davis, A., & Stylianou, Y. (2022). Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise. https://doi.org/10.48550/arXiv.2203.10637.

Rothermich, K., Harris, H. L., Sewell, K., & Bobb, S. C. (2019). Listener impressions of foreigner-directed speech: A systematic review. *Speech Communication, 112*, 22–29. https://doi.org/10.1016/j.specom.2019.07.002.

Scarborough, R., & Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America, 134*(5), 3793–3807. https://doi.org/10.1121/1.4824120.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in croatian and English. *The Journal of the Acoustical Society of America, 118*(3), 1677–1688. https://doi.org/10.1121/1.2000788.

Smiljanić, R., & Bradlow, A. R. (2008). Stability of temporal contrasts across speaking styles in English and croatian. *Journal of Phonetics, 36*(1), 91–113. https://doi.org/10.1016/j.wocn.2007.02.002.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly. Talker and listener factors in speaking style changes. language and linguistics. *Compass, 3*(1), 236–264. https://doi.org/10.1111/j.1749-818X.2008.00112.x.

Smiljanić, R., Keerstock, S., Meemann, K., & Ransom, S. M. (2021). Face masks and speaking style affect audio-visual word recognition and memory of native and non-native speech. *The Journal of the Acoustical Society of America, 149*(6), 4013–4023. https://doi.org/10.1121/10.0005191.

Stan Development Team (2023). *Stan Modeling Language Users Guide and Reference Manual, Version*. Available online at: https://mc-stan.org (accessed January 3, 2024).

Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 207–235). Blackwell.

Van Engen, K. J., Phelps, J. E. B., Smiljanić, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research, 57*(5), 1908–1918. https://doi.org/10.1044/JSLHR-H-13-0076.

Winn, M. (2019). Make speech-shaped noise [Praat script]. http://www.mattwinn.com/praat/Make_SSN_from_LTAS_selected_sounds.txt.

Wright, R., Tucker, B. V., & Kelley, M. C. (2023). The Effect of Speaker on Speech Intelligibility. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 516-520). Guarant International.

Xiao, J., Liu, J., Li, D., Zhao, L., & Wang, Q. (2022). Speech intelligibility enhancement by non-parallel speech style conversion using CWT and iMetricGAN based CycleGAN. In *International Conference on Multimedia Modeling* (pp. 544-556). Springer. https://doi.org/10.1007/978-3-030-98358-1_43.

Zellou, G., Barreda, S., & Segedin, B. F. (2020). Partial perceptual compensation for nasal coarticulation is robust to fundamental frequency variation. *The Journal of the Acoustical Society of America, 147*(3), EL271-EL276. https://doi.org/10.1121/10.0000951.

Zellou, G., Lahrouchi, M., & Bensoukas, K. (2022). Clear speech in tashlhiyt Berber: The perception of typologically uncommon word-initial contrasts by native and naive listeners. *The Journal of the Acoustical Society of America, 152*(6), 3429–3443. https://doi.org/10.1121/10.0016579.