Journal of Phonetics 81 (2020) 100980

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information

Charles Redmon [a,*], Keith Leung [b,*], Yue Wang [b], Bob McMurray [c,d,e], Allard Jongman [a], Joan A. Sereno [a]

[a] Department of Linguistics, University of Kansas, Lawrence, KS 66045, USA
[b] Department of Linguistics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada
[c] Department of Psychology, University of Iowa, Iowa City, IA 52242, USA
[d] Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242, USA
[e] Department of Linguistics, University of Iowa, Iowa City, IA 52242, USA

ARTICLE INFO

Article history:
Received 20 July 2019
Received in revised form 21 April 2020
Accepted 26 April 2020
Available online 21 May 2020

Keywords:
Clear speech
Audio-visual perception
Ideal perceiver models
L2
English
Mandarin
C-CuRE

ABSTRACT

The effect of clear speech on the integration of auditory and visual cues to the tense-lax vowel distinction in English was investigated in native and non-native (Mandarin) perceivers. Clear speech benefits for tense vowels /i, ɑ, u/ were found for both groups across modalities, while lax vowels /ɪ, ʌ, ʊ/ showed a clear speech disadvantage for both groups when presented in the visual-only modality, with Mandarin perceivers showing a further disadvantage for lax vowels presented audio-visually, and no difference in speech styles auditorily. English perceiver responses were then simulated in an ideal perceiver model which both identified auditory (F1, F2, spectral change, duration) and visual (horizontal lip stretch, duration) cues predictive of the clear speech advantage for tense vowels, and indicated which dimensions presented the greatest conflict between cues to tensity and modifications from clear speech (F2 and duration acoustically, duration visually). Altogether, by combining clear speech acoustics, articulation, and perception into a single integrated framework we are able to identify some of the signal properties responsible for both beneficial and detrimental speech style modifications.

1. Introduction

Face-to-face speech communication may adopt different forms and styles depending on speaking environments or communicative needs. In auditorily or visually challenging contexts, talkers often alter speech production using a clarified, hyper-articulated speech style to enhance intelligibility. This results in both articulatory and acoustic modifications (Gagné, Rochette, & Charest, 2002; Helfer, 1997; Moon & Lindblom, 1994; Payton, Uchanski, & Braida, 1994; Picheny, Durlach, & Braida, 1985; Uchanski, Choi, Braida, Reed, & Durlach, 1996). This well attested style of speech raises important questions as to whether and how these articulatory and acoustic changes are utilized by the perceiver to improve intelligibility. While the question of perceiver benefits has been addressed by several prior studies (Bradlow & Bent, 2002; Ferguson & Kewley-Port, 2002; Krause & Braida, 2002; Picheny et al., 1985; Uchanski et al., 1996), fully addressing these questions requires us to simultaneously understand the (implicit) motivation of the talker to modify their articulation, the specific articulatory changes of the talker, and the resultant effects on perception. This has not been attempted by prior studies. Thus, the present study investigates the entire speech chain, by examining the effects of clear (relative to plain) speech on auditory-visual (AV) perception of English tense and lax vowels by native (English) and non-native (Mandarin) perceivers, as well as the association between articulatory-acoustic clear-speech modifications and intelligibility.

1.1. Theoretical framework

Clear speech, a type of hyper-articulation, has been explained within the framework of the H & H (hyper- and hypo-articulation) theory (Lindblom, 1990). Under this view, hyper-articulated speech is typically produced with the intention to enhance sound category discriminability in response to challenging listening situations. Clear speech has been claimed to arise from two levels of modifications: signal-based and code-based (Bradlow & Bent, 2002).

[*] Corresponding authors.
E-mail addresses: redmon@ku.edu (C. Redmon), kwl23@sfu.ca (K. Leung).

https://doi.org/10.1016/j.wocn.2020.100980
0095-4470/© 2020 Elsevier Ltd. All rights reserved.

First, talkers could globally modify the signal to enhance general acoustic clarity or saliency (*signal-based modifications*). For example, they could raise the pitch or change the dynamic pitch range, decrease speaking rate and insert more pauses, or they could increase the amplitude to help separate speech and noise. Such modifications would presumably be uniformly beneficial to all listeners, both native (L1) and non-native (L2).

Second, talkers could also engage what Bradlow and Bent term *code-based modifications*. Such modifications could enhance acoustic distance between phonemic categories, for example, by altering the formants to make two vowels more phonetically distinct (e.g., Leung, Jongman, Wang, & Sereno, 2016), by non-uniformly modifying segment durations (e.g., lengthening typically longer tense vowels more than lax) (Leung et al., 2016), by producing less vowel reduction (Ferguson & Kewley-Port, 2007), or by just maintaining pronunciation norms (coarticulation, voice onset time) in speech (Ohala, 1995).

Both of these modifications must retain segmental cues and keep those cue values within the intended category, so that phonemic categorical distinctions can be maintained (Moon & Lindblom, 1994; Ohala, 1995). Thus, clear-speech effects must involve coordination of signal- and code-based strategies to enhance as well as preserve phonemic distinctions (Moon & Lindblom, 1994; Ohala, 1995; Smiljanić & Bradlow, 2009). This may be more challenging in cases where signal-based cues like duration or pitch also serve code-based functions.

In considering the interaction of clear-speech effects on various cues on perception, it is clear that cues and their influences cannot be examined individually. McMurray and Jongman (2011), for example, examined 24 distinct cues to fricatives (and see Cole, Linebaugh, Munson, & McMurray, 2010, for applications to vowels). Individually, most, if not all, of these cues were highly variable and were insufficient to distinguish the fricatives, and even optimally weighting and combining them could not lead to listener-like levels of performance. However, when the same cues were subjected to a simple model that accounted for various causal factors (e.g., talker differences, coarticulation), they were able to predict listener performance fairly accurately. This suggests that to properly understand the way a given factor (like clear speech) affects perception, one must determine (1) how its effects on multiple cues are weighted and combined to lead to the percept; and (2) how the effect of the factor of interest (e.g., clear speech) fits into the context of other known influences on the acoustics (e.g., talker differences). We accomplish this here by using the Computing Cues Relative to Expectations (C-CuRE) framework (McMurray & Jongman, 2011), which relativizes cues to speaker means and then combines them in a statistical learning model (typically within the logistic family of models) meant to approximate the decision problem presented to listeners in a perception experiment. We use this framework for the following: (1) to weight and combine cues; (2) to understand the variety of factors (clear speech and beyond) that influence the acoustics and articulation; and (3) to link acoustic and visible articulatory modifications to response patterns in perception.

1.2. Clear speech in auditory and visual perception

1.2.1. Clear-speech benefit

Clear speech has been shown to be more intelligible than plain, conversational speech. This is particularly so when listening conditions are challenging, such as in background noise (Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Krause & Braida, 2002; Payton et al., 1994; Uchanski et al., 1996), or when listeners are hearing-impaired (Bradlow, Kraus, & Hayes, 2003; Liu, Del Rio, Bradlow, & Zeng, 2004; Picheny et al., 1985) or are non-native listeners (Bradlow & Bent, 2002). Clear speech typically results in a gain of about 7–38% of tokens recognized in clear speech relative to plain speech (Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Maniwa, Jongman, & Wade, 2009; Payton, Uchanski, & Braida, 1994; Uchanski, Choi, Braida, Reed, & Durlach, 1996). This clear-speech advantage has been observed at different linguistic levels, for sentences (Bradlow & Bent, 2002; Gagné, Querengesser, Folkeard, Munhall, & Masterson, 1995; Krause & Braida, 2002; Payton et al., 1994), words (Gagné, Masterson, Munhall, Bilida, & Querengesser, 1994; Uchanski et al., 1996), and segments (Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Gagné et al., 2002).

Specifically relevant for the current study is research on vowel intelligibility in English. Ferguson (2004) tested the intelligibility of ten English vowels (/i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u/ in a /bVd/ context) in plain and clear speech styles by 7 young healthy adult native English-speaking listeners. The stimuli were presented auditorily in multi-talker babble noise (−10 dB SNR). The results show that clear speech was 8.5% more intelligible on average than plain speech. Results for individual vowels as shown in Figure 1 of Ferguson (2004) suggest a significant clear-speech advantage for /æ, ɑ, ʌ/. Detailed analyses of the acoustics of the stimuli in Ferguson (2004), as well as the relation between the acoustics and intelligibility, were provided in a subsequent study by Ferguson and Quené (2014). We will refer to those results in Section 1.3.2 below.

Clear speech can also improve intelligibility in visual (facial) speech perception (Gagné et al., 1994, 2002; Helfer, 1997; Lander & Capek, 2013; Van Engen, Phelps, Smiljanić, & Chandrasekaran, 2014). For example, Gagné and colleagues (1994, 2002) examined the perception of clear and plain French CV syllables (/b, d, g, v, z, ʒ/ + /i, y, a/) and found significant clear-speech gains in the intelligibility of AV, visual-only, as well as auditory-only presentations. These findings demonstrate the existence of a clear-speech advantage across input modalities, suggesting that clear speech affects not only acoustic cues, but also visual cues.

1.2.2. Weighting cues across modalities

Gagné et al. (2002) suggest the magnitude of the clear-speech benefit in visual speech may be less than in the auditory modality. Moreover, while either speaking clearly or providing visual speech information can be beneficial, the combination of the two can result in greater intelligibility gains than each domain alone (Helfer, 1997). Thus, speech style and modality may interact to affect speech intelligibility. This raises the question of what factors give rise to this interaction. However, research has not systematically explored under what circumstances clear-speech benefits may differ in auditory versus visual conditions.

A critical issue in understanding the mechanisms of these variable intelligibility gains is the question of the degree to which perceivers weight (or use) inputs from different modalities (or different cues within a modality). In AV speech perception, the weight granted to auditory versus visual cues can be affected by the relative quality of the information in each channel (Gagné et al., 2002; Hazan, Kim, & Chen, 2010). For example, a compensatory modality weighting effect has been found where perceivers utilize information from an alternate modality (e.g., visual) when the other (auditory) was degraded (Hazan et al., 2010; Van Engen et al., 2014). Similarly, perceivers rely more on the auditory modality for low vowels, as the acoustic cue to vowel height (F1) is more salient, whereas they put more weight on the visual input to perceive rounded vowels since the visual cue (lip-rounding) is more salient (Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Traunmüller & Öhrström, 2007). Likewise, higher visual perceptual accuracy was found for identification of the visually more salient labial/labio-dental consonantal contrasts compared to visually less salient alveolar/post-alveolar contrasts (Hazan et al., 2006; Wang, Behne, & Jiang, 2008).

These patterns of AV weighting raise questions regarding the role of clear speech in AV perception: Does clear speech enhance code-based cues only, making them more salient as category-distinctive cues? Or, does clear speech involve global signal-based enhancement, resulting in increased salience of information across modalities? Or do these enhancements vary across modality?

1.2.3. Clear-speech effects as a function of listener linguistic experience

Although clear speech consistently benefits typical native language adult listeners, research on non-native perception suggests clear speech may be less helpful or even detrimental for L2 listeners (Bradlow & Bent, 2002; Fenwick, Davis, & Best, 2015; Granlund, Hazan, & Baker, 2012; Smiljanić & Bradlow, 2011). For example, Bradlow and Bent (2002) found substantially smaller clear-speech benefits for non-native listeners as compared to native listeners in the intelligibility of clearly produced English sentences.

What can account for such differences? Bradlow and Bent (2002) suggest that both groups are able to take advantage of signal-based modifications, which are largely language-independent, accounting for the small benefit in L2 listeners. However, these groups may differ in their ability to use code-based modifications. Native speakers have extensive experience with the language and are knowledgeable about the particular phonetic realizations of segments in their language, as well as the higher-level contextual structures. This enables them to make use of code-based modifications. In contrast, non-native speakers have less experience with these aspects of the code (in their L2) and may not have been able to perceive or utilize code-based clear-speech cue enhancements specific to the L2.

Research has shown evidence supporting a code-based component to the small clear-speech intelligibility gains in non-native listeners. For example, in contrast to non-proficient L2 listeners (Bradlow & Bent, 2002), fluent L2 listeners showed significantly larger clear-speech intelligibility gains in the perception of English sentences (Smiljanić & Bradlow, 2011). Indeed, further research at the segmental level has shown that the degree and direction of clear-speech effects on non-native speech intelligibility may depend on the relation between L1 and L2 phonetic inventories. Fenwick et al. (2015) tested AV perception of Sindhi consonants in consonant–vowel syllables in clear and plain speech by Australian-English perceivers. The consonants contrasted both in place of articulation (POA) and voicing, and in their proximity to the perceivers' L1 (English), with phonologically "two-category" contrasts (/ɓ-ɗ/ [POA] and /f-v/ [voicing]) and phonetic-level "category-goodness" differences (/d̪-ɖ/ [POA] and /t̪-d̪/ [voicing]). While the results showed no clear-speech effects for the stimuli with POA contrasts, a clear-speech benefit was found for voicing only for the phonetic-level category goodness differences but not for the two-category contrasts. The results show that clear speech can benefit non-native perception when the contrasts are perceived as differing in phonetic "category-goodness", indicating benefits from within-category enhancement may be at the "signal" rather than "code" level (cf. Bradlow & Bent, 2002) for the non-native listeners.

These non-native patterns in clear speech reflect the influence of linguistic experience. Clear-speech benefits may be less robust when non-native listeners are less knowledgeable about the sounds in the L2, or about the specific cues to phonetic contrasts in the L2 (Smiljanić & Bradlow, 2009), or when they are less proficient in the L2 (Smiljanić & Bradlow, 2011). Such findings underscore the possibility of code-based modifications that are specific to the phonetics and phonology of the language. On the other hand, non-native listeners may also benefit from clear speech in the L2 when the modifications are perceived as signal-enhancing cues in their L1 (cf. Fenwick et al., 2015), supporting additional, more general signal-based modifications for the clear-speech effect.

Together, results from native and non-native clear-speech perception across AV modalities demonstrate differences in clear-speech benefits that may be triggered by saliency-enhancing (signal-based) and category-enhancing (code-based) cues. However, intelligibility data alone cannot disentangle whether any observed perceptual patterns are directly attributable to signal-based or code-based modifications in production.

1.3. Linking clear-speech perception and production

1.3.1. Acoustic and visual/articulatory clear-speech features

Isolating code-based from signal-based effects in clear speech is challenging with intelligibility data alone, particularly in L1 speakers where variation in linguistic knowledge cannot be brought into play. However, in L2 speakers this can be difficult as well, given overlap between the languages and variation in the degree of L2 experience.

In contrast, phonetic studies may be able to isolate code-based changes by examining the specific acoustic and articulatory modifications to aspects of the signal that indicate speech categories. Understanding the details of what is changing acoustically and articulatorily/visually will shed light on differentiating code-based and signal-based effects in clear speech.

Research on acoustic and articulatory correlates of clear speech (Ferguson & Quené, 2014; Ferguson and Kewley-Port, 2002, 2007; Leung et al., 2016; Tang et al., 2015; Tasko & Greilick, 2010; Yehia, Kuratate, & Vatikiotis-Bateson, 2002) has shown that clear speech involves more extreme articulatory configurations and correspondingly, more exaggerated acoustic properties than are seen in plain speech. In the acoustic domain, studies examining English vowels produced in controlled segmental contexts (Ferguson & Quené, 2014; Ferguson and Kewley-Port, 2002, 2007; Leung et al., 2016) or excised from natural sentential contexts (Hazan & Baker, 2011; Kim & Davis, 2014; Lam, Tjaden, & Wilding, 2012; Picheny, Durlach, & Braida, 1985; Smiljanić & Bradlow, 2008) consistently reveal that vowel duration increases in clear speech relative to plain speech. Given that this is a global lengthening across all vowels, it is assumed to be a signal-based effect.

However, vowel length is a useful phonetic cue for distinguishing tense and lax vowels. In Leung et al. (2016), measures of both absolute and relative vowel duration showed a greater lengthening in clear speech for tense vowels than for lax vowels. This data suggests that clear-speech modifications differentially enhance the properties of vowels (tense vowels being intrinsically longer than lax vowels), suggesting instead a code-based modification.

In this same vein, clear and plain vowels also differ in the spectral domain. Clearly produced vowels are characterized by a larger vowel space (F1 × F2 space) than plain vowels (Cooke & Lu, 2010; Ferguson & Kewley-Port, 2007; Ferguson & Quené, 2014; Leung et al., 2016; Smiljanić & Bradlow, 2005), suggesting a code-based modification. Moreover, F1 modifications may also reflect signal-based properties: plain-to-clear-speech modifications generally involve a global increase in F1 regardless of the height of the vowel (Ferguson & Kewley-Port, 2002; Ferguson & Quené, 2014; Huber, Stathopoulos, Curione, Ash, & Johnson, 1999; Lu & Cooke, 2008). Furthermore, clearly produced vowels are globally found to be more dynamic than plain vowels, as indicated by relative formant changes along the formant trajectories (Ferguson and Kewley-Port, 2002, 2007; Leung et al., 2016; Moon & Lindblom, 1994), all suggesting signal-based modification.

However, the degree of vowel dynamicity varies among individual vowels, suggesting a more code-based component. In particular, the more dynamic lax vowels show greater spectral change in clear speech than the intrinsically less dynamic tense vowels (Assmann & Katz, 2005; Ferguson & Kewley-Port, 2007; Hillenbrand & Nearey, 1999; Leung et al., 2016).

Articulatory studies have also revealed both code- and signal-based clear-speech modifications. For example, Tang et al. (2015), which examined visible articulatory movements in English vowel production using computational image analysis, has shown that talkers modify their speaking style to produce clear speech with exaggerated visual cues corresponding to code-based articulatory features of different vowels. In particular, in clear compared to plain speech, the results show greater horizontal lip stretch for front unrounded vowels and greater degree of lip rounding and protrusion for rounded vowels. On the other hand, signal-based modifications are shown by a larger jaw opening across vowels in clear relative to plain speech, which is probably a consequence of increased articulatory effort in general, as also claimed previously (Kim & Davis, 2014).

In sum, these production studies have documented both signal-based and code-based changes in clear speech. Yet the question remains as to how the effects seen in these acoustic and articulatory measurements are linked to intelligibility. In particular, no acoustic or articulatory analysis has yet adopted the more comprehensive approach, as in the C-CuRE framework of McMurray and Jongman (2011), to ask how specific acoustic cues (as opposed to broad measures of clarity like vowel space area) contribute to perception, or how this may be impacted by other sources of variation.

1.3.2. Linking clear-speech features to intelligibility

Research relating clear-speech acoustic patterns to perception could be crucial in identifying the locus of the clear-speech advantage as it can reveal which modifications most predict intelligibility gains. Such work is scarce.

Lam et al. (2012) used regression analyses to directly relate acoustic features in clear speech to sentence intelligibility. In clear speech, increases in intelligibility were related to greater increases in the area of the tense vowel space, greater dynamic spectral changes for lax vowels, along with greater reduction in speaking rate and greater increases in intensity. Although not specifically targeting segment-level intelligibility, these findings indicate that enhanced intelligibility in clear speech may be associated with different acoustic cues depending on the features of different sound categories.

Ferguson and Quené (2014) used Generalized Linear Mixed Modeling to relate their acoustic measurements to the intelligibility data reported in Ferguson (2004). Their results are generally in good agreement with those of Lam et al. (2012) in that a decrease in speaking rate, increase in F1 (due to greater mouth opening in an effort to increase intensity), and increase in the vowel space area all contributed to a clear-speech intelligibility benefit. In addition, greater F1 and F2 movement over the vowel nucleus in the clear production of the vowels /e, o, ʊ, u/ was also seen to enhance their intelligibility. Thus, like Lam and colleagues, this suggests both signal and code-based modifications are important.

In terms of articulation, studies using kinematic measures have shown positive correlations of articulation and acoustics with clear-speech effects on intelligibility (Kim & Davis, 2014; Kim, Sironic, & Davis, 2011; Tasko & Greilick, 2010). For example, Kim et al. (2011) tracked the motion of facial markers during clear speech produced in quiet or in the presence of background noise (Lombard speech), and coupled this with tests of the audio-visual intelligibility of these productions in noise. Motion tracking results revealed a greater degree of articulatory movement in speech in noise (clear speech) than in quiet (plain speech), with the differences correlated with speech acoustics. Moreover, increased movement of the jaw and mouth (greater degree of opening) during clear speech translated to increased intelligibility, indicating that clear speech is also more visually distinct than plain speech.

With the exception of sentence-level intelligibility (e.g., Kim et al., 2011), research has not examined the degree to which specific articulatory cues contribute to enhanced intelligibility in clear-speech segments, nor is there robust evidence identifying signal- and code-based modifications in acoustic cues that lead to intelligibility gains. The gap in this area of work reveals the need for research to establish the link between specific articulatory and acoustic features used in clear-speech segmental productions and the impact of these features on the intelligibility of clear-speech segments. Critically, here by adopting an explicit computational model of perception (McMurray & Jongman, 2011), we can examine the impact of clear speech on the way in which multiple cues combine to yield perception.

1.4. The present study

The above-reviewed findings on AV clear-speech intelligibility indicate that the perception of clear-speech effects may depend on factors such as the saliency of the source of modifications (acoustic and articulatory), perceptual weighting in auditory and visual modalities, and perceivers' linguistic experience. However, research has not systematically examined the extent to which these inter-related factors collectively affect intelligibility, nor is it clear whether these modifications are global signal-based changes, or more phonetically specific, code-based changes. Thus, the current study addressed how speech style interacts with AV input modality and perceiver experience in the intelligibility of clear-speech segments, and what acoustic and articulatory modifications are responsible for these interactions.

Specifically, the present study investigates AV perception of English tense and lax vowels in clear speech by native English and Mandarin (L2) perceivers. This study aims to isolate the effects of signal- and code-based acoustic and articulatory clear-speech modifications on the intelligibility of these vowels in two ways. First, we compare the patterns by native and non-native listeners who may interpret signal- and code-level cues differently based on their native language experience. Second, we relate differences in identification to differences in both signal- and code-based cues measured from the acoustic and visual input.

Tense and lax vowels were chosen as target stimuli due to their unique articulatory and acoustic characteristics in relation to clear-speech features. As noted previously (Leung et al., 2016), features that mark plain-to-clear speech modifications and lax-to-tense vowel contrasts are similar, both involving increased duration, fundamental frequency (f0) and intensity, and more peripheral formant frequencies (associated with an expanded vowel space), as well as increased dynamic temporal and spectral changes (Cooke & Lu, 2010; Ferguson & Kewley-Port, 2002, 2007; Ferguson & Quené, 2014; Hazan & Baker, 2011; Kim & Davis, 2014; Krause & Braida, 2002; Lu & Cooke, 2008; Picheny, Durlach, & Braida, 1985). These similarities provide a unique test case to unravel the underlying mechanisms governing clear-speech production and perception based on how the same physical features may be utilized differently depending on different priorities needed for efficient communication.

In terms of the interactive effects of speech style and input modality, first, we hypothesize greater overall intelligibility for vowels produced in clear speech relative to plain speech. This should be seen across tensity (tense vs. lax vowel stimuli) and modality (A vs. V) conditions. This is based on our previous findings of greater articulatory (jaw, lip) movements (Tang et al., 2015) as well as greater acoustic (temporal, spectral) changes (Leung et al., 2016) in plain-to-clear modifications for both tense and lax vowels. However, based on our findings of greater acoustic distinctions between tense and lax vowels in clear (relative to plain) speech, but similar articulatory plain-to-clear modifications for both tensity categories, we predict that the Speech Style × Input Modality interaction would be reflected in perception as well. In particular, code-based acoustic modifications that result in greater tense-lax differences may enhance auditory intelligibility in clear speech, whereas articulatory modifications that do not differentiate tense and lax vowels should not provide a comparable benefit in the visual domain.

Regarding the effects of linguistic experience, we recruited native Mandarin perceivers as the non-native group in order to test the signal- versus code-based hypothesis for clear speech, since unlike English, Mandarin does not have lax counterparts to its tense vowels and this difference poses difficulties for Mandarin native speakers in perceiving the tense and lax vowel distinctions in English (Jia, Strange, Wu, Collado, & Guan, 2006; Wang & Munro, 2004). Based on the previous findings of language-specific, code-based clear-speech effects in the auditory domain (Bradlow & Bent, 2002; Smiljanić & Bradlow, 2011), we predict greater clear-speech benefits for native English than for Mandarin perceivers, particularly for perception of the lax vowels that are unfamiliar to the Mandarin perceivers. However, in the visual domain, on the basis of the previous findings that non-native perceivers may utilize signal-based clear-speech enhancements (Fenwick et al., 2015) and that non-native perceivers generally rely more on the visual domain than native perceivers (Hazan et al., 2006; Wang et al., 2008), we expect Mandarin perception in the current study to be more affected by clear than plain speech (although the effects may be skewed if attention was paid to incorrect visual cues, Hazan et al., 2006; Kirchhoff & Schimmel, 2005; Wang et al., 2008).

Finally, we relate articulatory, acoustic, and perception data to determine the relative weight of each articulatory and acoustic cue in predicting perceiver performance. Extending the previous findings of positive correlations between specific articulatory and acoustic clear-speech modifications and improved overall sentence intelligibility (Ferguson & Kewley-Port, 2002; Kim et al., 2011), we predict similar positive correlations in segmental intelligibility. Furthermore, we expect enhanced clear-speech intelligibility to correlate with those articulatory and acoustic features used to make quantitative modifications, whereas we expect the features used to characterize phonemic categorical contrasts to correlate with identification of different vowels across speech styles.

2. Methods

2.1. Perceivers

Twenty-one (19 female) native perceivers of Western Canadian English (aged 19–27, mean: 22) and 30 (18 female) non-native perceivers (aged 18–26, mean: 22) who had Mandarin as their first language (L1) were recruited from the undergraduate and graduate population at Simon Fraser University, Canada. The perceivers reported normal hearing, normal or corrected vision, and no history of speech or language disorders.

The Mandarin perceivers were late, intermediate-level learners of English. According to a self-reported questionnaire, they initially started learning English as a second language (L2) at a mean age of 10 (SD: 3.7) in a classroom setting. They arrived in Canada at a mean age of 19 (SD: 2.7) and had been residing in an English-speaking environment for 3.4 years on average (SD: 2.0) at the time of testing. The Mandarin perceivers reported that their daily use of English was 41% on average (SD: 20.5), and their standard English test scores were: 5.5–7.5 (IELTS) or 96–103 (TOEFL). In order to establish that the Mandarin participants did have difficulty with the English vowel tensity distinctions (thus allowing the test of interactive effects of speech style and tensity), a screening procedure was included prior to the perception experiment, where participants were asked to produce the six target English words containing tense and lax vowels. Their productions in terms of the degree of tense-lax vowel distinction were assessed by a phonetically-trained native English listener on a scale of 1 to 5, with 1 meaning "no distinction at all" and 5 meaning "perfect, native-like distinction". The mean rating was 2.9 (SD: 0.9).

2.2. Items

The stimuli included six English words: "keyed, kid, cod, cud, cooed," and "could" spoken in plain and clear speech styles. These words carry three pairs of English tense and lax vowels (/i-ɪ/, /ɑ-ʌ/ and /u-ʊ/) based on previously established tense-lax categorization (e.g., Gopal, 1990; Lam, Tjaden, & Wilding, 2012; Leung, Jongman, Wang, & Sereno, 2016).

2.2.1. Talkers

Eighteen (10 female) native speakers of Western Canadian English provided the audio-visual stimuli. From this pool, six talkers (3 female) whose productions contained the most contrastive visible articulatory features in plain versus clear speech, based on our previous articulatory analysis (Tang et al., 2015), were chosen for the current study.

The talkers (aged 17–30, mean: 22) were recruited from the undergraduate and graduate population at Simon Fraser University. Their English dialect exhibits the /ɑ/ and /ɔ/ merger (Clopper, Pisoni, & De Jong, 2005); thus they produced the vowel in "cod" as the target vowel /ɑ/ of this study. The talkers indicated no history of speech or language impairment.

2.2.2. Elicitation of plain and clear stimuli

The plain and clear stimuli were elicited using a simulated interactive computer speech recognition program established previously (Leung et al., 2016; Maniwa, Jongman, & Wade, 2009; Tang et al., 2015). On each trial, one of the six English words was displayed on a computer screen. The talker was instructed to produce the word naturally, eliciting a neutral, 'plain' speech style. Then, the program would "guess" and indicate on the screen what they produced. If the guess was incorrect, the program would instruct the talker to repeat the stimulus as clearly as possible (thus eliciting clear style productions).

Audio/video recordings were acquired in a sound-attenuated booth in the Language and Brain Lab at Simon Fraser University. Front-view videos were captured with a Canon Vixia HF30 camera at a recording rate of 29 fps. Audio recordings were made simultaneously using Sonic Foundry Sound Forge 6.4 at a sampling rate of 48 kHz, with a Shure KSM microphone placed at a 45-degree angle, about 20 cm away from the talker's mouth.

Each word was presented three times in a random order, resulting in three elicitations of each plain-clear pair of productions. Further, all audio and video stimuli were evaluated by two phonetically trained native speakers of Canadian English to ensure the accuracy of pronunciation and quality of recordings. All productions were judged as correct productions of the target words.

2.2.3. Editing of stimuli

Three sets of stimuli for the perceptual experiments were created. Stimuli varied in the three presentation conditions: audio-only (AO), audio-visual (AV) and visual-only (VO). The AO stimuli were excised from the microphone audio recordings as individual word clips of two seconds each, using Audacity 2.1. The AV stimuli were created by replacing the on-camera audio track with the high-quality audio recordings from the microphone, and the VO stimuli were created by removing the audio track from the video recordings, both using Adobe Premier Pro CC 2014. Each AV or VO word clip lasts four seconds to ensure that both mouth opening and closing are captured. To increase the difficulty level for the AO and AV conditions (thus inducing sufficient errors for comparisons between plain and clear speech), the audio stimuli were normalized at 60 dB and were embedded in one of three stretches of cafeteria noise at a level of 75 dB (i.e., −15 dB signal-to-noise ratio or SNR).

Pilot experiment. This SNR level was empirically established by a separate pilot study, with the target error rate set at 30% based on similar previous studies (Gagné et al., 2002; Wang et al., 2008). In this pilot, the six target words were each embedded in six noise levels (SNRs of −5, −10, −13, −15, −17 and −19 dB). In total, this pilot experiment tested 144 audio stimuli (6 talkers × 2 styles × 2 words × 6 SNRs). Within a given cell (talker × style × word), the specific audio recording was randomly selected from the three elicitations in the production task described earlier. These stimuli were then excluded from the main perception experiment to ensure that any idiosyncrasies of these particular items (and the SNR calibration done on them) did not bias the results of the primary experiment. This left two elicitations in each that were available for the main experiment.

To keep the pilot experiment short (30 min), each listener responded to two different words that were selected from each talker, but all six target words were used across talkers. Eleven native Canadian English listeners (9 female) who did not participate in the subsequent perception experiment took part in this pilot study. On each trial, participants indicated which word they had heard from among the six alternatives displayed on the screen. The target error rate was obtained at a −15 dB SNR.

2.3. Procedures

The perception experiment was presented using Paradigm (Tagliaferri, 2005). Perceivers were tested in a sound-attenuated room. Each perceiver was seated in front of a flat-screen monitor. Auditory stimuli were normalized to 70 dB, and presented binaurally over headphones at a comfortable listening level. Visual stimuli were presented in full color video showing a front view of the speaker's face, at an image size of 15 × 22 cm (width by height), where the viewing distance was approximately 70 cm. For AO and VO conditions, only one stimulus channel (audio and video, respectively) was presented to participants, while for AV conditions the above two channels were presented simultaneously. A six-alternative forced-choice identification task was used. On each trial, a stimulus was presented, and participants were asked to identify what they had perceived from among the six alternatives ("keyed, kid, cod, cud, cooed, could") displayed on the screen. They were given up to four seconds on each trial to indicate their response, which was registered via mouse click on one of the six words shown on the screen.

A familiarization session was administered prior to the main testing session. First, each target word was presented auditorily without noise to ensure that the perceivers could recognize the target words. Then, perceivers went through the three stimulus modalities to be familiar with the task, each of which included two example trials containing stimuli that were not used in the testing sessions.

The main perception experiment contained stimuli from one of the elicitations not used in the SNR pilot study. Each plain and clear production from each talker was presented three times in a random order in each A/V modality, resulting in a total of 648 stimuli (6 talkers × 2 styles × 3 stimulus modalities × 6 words × 3 repetitions). Stimuli were presented over two testing sessions on two consecutive days, each lasting 60–80 minutes, including the main testing session as well as task instructions, practice trials, and breaks. Each testing session contained three blocks and each block had one of the three stimulus modalities so that the perceivers went through all three stimulus modalities in one testing session. All blocks had an equal number of trials (108). The order of presentation was randomized and the order of blocks (stimulus modalities) was counter-balanced across participants.

2.4. Acoustic and articulatory analyses

In order to relate perceptual patterns to specific speech parameters from audio and visual input, acoustic and visible articulatory analyses were performed on the audio and video recordings. The acoustic measurements (i.e., acoustic cues) include vowel duration, the frequencies of the first three formants at vowel midpoint, spectral change and spectral angle (Leung et al., 2016). The articulatory measurements (i.e., visual cues), conducted on videos of talkers' faces using computer-vision and image processing techniques, include peak of horizontal and vertical lip stretch, vertical jaw displacement and eccentricity of lip rounding (Tang et al., 2015). See Sections 3.3.1-3.3.2 for a review of the above parameters, originally presented in Leung et al. (2016) and Tang et al. (2015), as a function of vowel and speech style.

3. Results

First, English and Mandarin perceiver identifications of tense and lax vowels were analyzed separately for overall accuracy as a function of speech style and stimulus modality. Accompanying the overall accuracy analysis, we also analyzed the accuracy of identifying specific features (among *tensity*, *height*, *backness*, and *rounding* distinctions) to understand what specific cue enhancements or distortions underlie the overall effects of clear speech on tense/lax vowel perception. Finally, acoustic and visual parameters from two prior studies (Leung et al., 2016; Tang et al., 2015, respectively) were used in both inferential and predictive models of English perceiver behavior to determine the relative contribution of information from auditory and visual modalities to clear and plain speech perception.

3.1. English perceivers

Fig. 1 shows the accuracy of English perceivers as a function of modality, speech style, and stimulus vowel tensity. Overall, this shows a clear-speech advantage in all conditions except when lax vowels are presented in the visual-only modality. That is, when only visual information is available, clear speech presents a disadvantage to lax vowel identification, a disadvantage which will later be shown to derive from the conflict between speech style modifications of articulations and those used to cue the tense-lax distinction in English.

This pattern of results was analyzed numerically in a logistic mixed-effects model predicting accuracy (correct = 1; incorrect, including non-responses, = 0). Fixed effects were dummy coded and included Modality (AV [reference], AO, VO), Style (plain [reference], clear), and Stimulus Vowel Tensity (tense [reference], lax). Because effects were dummy codes, the significance of the individual regression tests the hypothesis that a given level of one of the factors (e.g., AO of the modality factor) significantly differs from the reference level. This allowed us to conduct a number of what would normally be post-hoc tests without the need for separate models.

Random effects were chosen by forward model selection to find the most complex random slope structure necessary to fit the data (c.f., Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017; Bates, Maechler, Bolker, & Walker, 2015). The resulting model included a random intercept for Subject, a random slope for Modality on Subject, and a random intercept for Item (see Eq. (1) for model in lme4 syntax).[1] Finally, while a random effect for Talker was tested, it was not included in the final model as it did not yield any substantive changes in the results, and is an effect which is already reflected in the Item random effect that refers to the exact items (talker and SNR-specific) presented in the experiment.

$$\begin{aligned}\text{Accuracy} &\sim \text{Modality} * \text{Style} * \text{Tensity} + (\text{Modality}|\text{Subject})\\ &\quad + (1|\text{Item})\end{aligned} \tag{1}$$

[1] Attempts at introducing additional complexity in the random effects structure, such as including additive effects of Style and Tensity, did not improve model fit, and random slopes for interaction effects resulted in model convergence failures, due in part to the fact that the model could completely predict performance in certain cells from the fixed and random effects (a marker of overfitting).

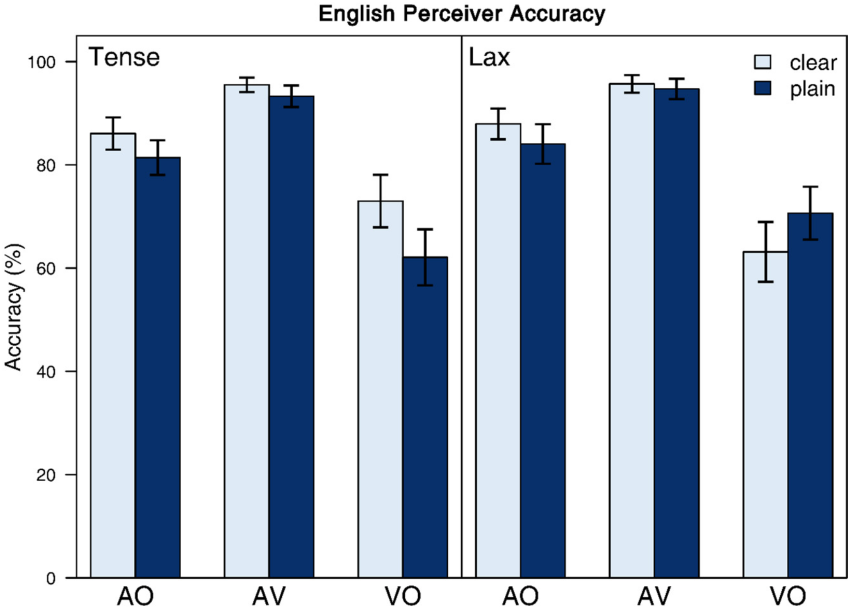

Fig. 1. Identification accuracy of English perceivers by speech style (clear, plain) and stimulus vowel tensity (tense, lax) in audio-only (AO), audio-visual (AV), and visual-only (VO) modalities. Error bars represent ±2 standard errors about the mean.

As several factors included more than two levels, the significance of main effects and interactions were assessed by comparing models with and without the relevant factor.

This model showed a three-way interaction between Modality, Style, and Tensity ($\chi^2(2) = 18.1$, $p < 0.001$). This interaction derives from a significant clear-speech advantage in audio-only for tense vowels ($\beta = -0.508$, CI = [−0.8, −0.2], $z = -3.851$, $p < 0.001$) and lax vowels ($\beta = -0.445$, CI = [−0.7, −0.2], $z = -3.264$, $p = 0.001$), while for audio-visual stimuli only tense vowels showed a clear-speech advantage (tense: $\beta = -0.522$, CI = [−0.9, −0.1], $z = -2.622$, $p = 0.009$; lax: $\beta = -0.280$, CI = [−0.7, 0.1], $z = -1.347$, $p = 0.178$), and for visual-only stimuli the clear-speech advantage for tense vowels ($\beta = -0.629$, CI = [−0.8, −0.4], $z = -6.210$, $p < 0.001$) is inverted for lax vowels, which show a significant *disadvantage* for clear speech ($\beta = 0.398$, CI = [0.2, 0.6], $z = 4.120$, $p < 0.001$). See Table A1 in the Appendix for the full regression table.

To further dissect the more phonologically specific advantages and disadvantages of clear speech, we recoded accuracy in terms of the features [tense], [back], [high], and [round]. For example, in measuring height accuracy, any response of /i, ɪ, u, ʊ/ was considered accurate if the stimulus was a high vowel. In contrast, if the listener responded /ɑ, ʌ/ for a high vowel, this was inaccurate. In addition to providing greater granularity to the analysis of perceiver responses, this decomposition serves to verify that the assumed source of the overall accuracy pattern in a conflict between articulatory modifications in clear speech and those distinguishing tense and lax vowels is in fact due to vowel tensity misperceptions, and not due to misperceptions between other vowel pairs representing non-tensity contrasts.

Fig. 2 plots feature accuracy by modality, stimulus vowel tensity, and speech style. The figure confirms the assumption that clear-speech modifications affected the perception of tensity, as the primary Tensity × Style interaction that emerges visually in Fig. 2 is in the bottom panel of column 1 (tensity perception in the VO modality). Here we find a substantial disadvantage for clearly spoken lax vowels, meaning that clear speech causes lax vowels to be misperceived as tense vowels, whereas lax vowels are accurately perceived in plain speech.

We next conducted four statistical analyses, one for each feature. This used a model similar to the overall accuracy model: a logistic mixed-effects regression predicting feature accuracy from Modality, Style, and Stimulus Vowel Tensity. This model used random intercepts for Subject and random slopes were included according to the forward model selection procedure described above. For all four analyses this yielded the same random slope for Modality, but not for Style or Tensity. Thus, all feature accuracy models are formulated according to Eq. (1).

We started by examining the [tense] feature (the primary feature of interest). As in the overall accuracy model, there was a significant interaction between Modality, Style, and Stimulus Vowel Tensity ($\chi^2(2) = 22.7$, $p < 0.001$). This effect derives primarily from the clear-speech disadvantage for lax vowels presented in VO ($\beta = 1.180$, CI = [1.0, 1.4], $z = 10.18$, $p < 0.001$), which runs counter to the significant advantage for clearly spoken tense vowels in the visual-only modality ($\beta = -0.646$, CI = [−0.8, −0.4], $z = -6.338$, $p < 0.001$). For the audio-only modality, while clear speech does not yield a disadvantage for lax vowels as it does in VO, it yields only a modest advantage ($\beta = -0.381$, CI = [−0.8, 0.0], $z = -1.993$, $p = 0.046$) as compared with that for tense vowels ($\beta = -0.883$, CI = [−1.2, −0.6], $z = -5.243$, $p < 0.001$). Finally, in AV, clear speech provides an advantage for tense vowels ($\beta = -0.548$, CI = [−1.0, −0.1], $z = -2.664$, $p = 0.008$), but not for lax ($\beta = 0.181$, CI = [−0.3, 0.7], $z = 0.717$, $p > 0.1$). In other words, clear speech appears to induce a bias to perceive more tense vowels overall. This bias results in greater tensity errors on lax vowels and fewer tensity errors on tense vowels.

Next, we examined the [back] feature. Here, no significant interaction between Modality, Style, and Tensity was observed

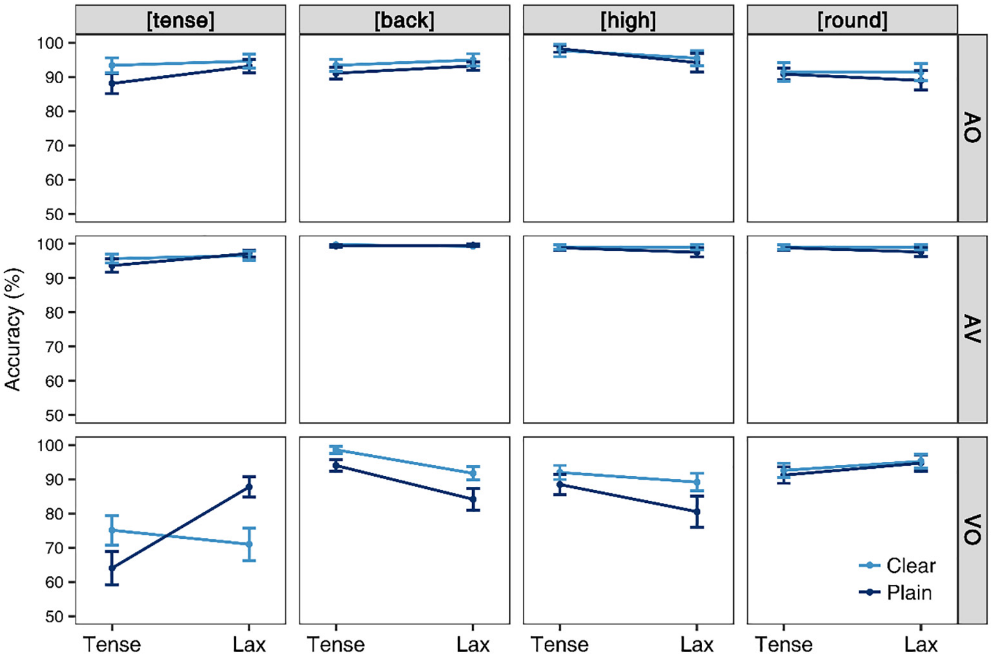

Fig. 2. Identification accuracy of English perceivers by feature ([tense], [back], [high], [round]), speech style (clear, plain) and stimulus vowel tensity (tense, lax) in audio-only (AO), audio-visual (AV), and visual-only (VO) modalities. Error bars represent ±2 standard errors about the mean.

($\chi^2(2) = 4.4$, $p > 0.1$), though there were significant two-way interactions between Modality and Style ($\chi^2(4) = 21.3$, $p < 0.001$), Modality and Tensity ($\chi^2(4) = 20.5$, $p < 0.001$), and Style and Tensity ($\chi^2(3) = 8.2$, $p = 0.042$). From Fig. 2 it can be seen that these interactions primarily derive from the visual-only modality, where both clear-speech effects and vowel tensity effects are more pronounced than in AO or AV. Specifically, in audio-only there was a clear-speech advantage for both tense ($\beta = -0.471$, CI = [−0.8, −0.1], $z = -2.576$, $p = 0.010$) and lax ($\beta = -0.452$, CI = [−0.9, 0.0], $z = -2.117$, $p = 0.034$) vowels. In AV, backness accuracy was at ceiling, so no significant clear-speech effects could be measured (ps > 0.2). Finally, in VO, accuracy on the [back] feature was significantly greater for both clearly spoken tense ($\beta = -1.707$, CI = [−2.3, −1.1], $z = -5.713$, $p < 0.001$) and lax ($\beta = -0.922$, CI = [−1.2, −0.6], $z = -6.083$, $p < 0.001$) vowels.

Our third analysis examined [height]. Height errors were generally uncommon, though when perceivers were presented with only visual information there were predictable effects of clear speech (namely, that the baseline lower accuracy on lax vowels in plain speech, relative to tense vowels, is largely remedied in clear speech). Overall, model comparisons revealed no significant Modality × Style × Tensity interaction ($\chi^2(2) < 1$), nor a significant interaction between Modality and Tensity ($\chi^2(4) = 5.2$, $p = 0.263$). However, there were marginal interactions between Modality and Style ($\chi^2(4) = 8.7$, $p = 0.070$) and Style and Tensity ($\chi^2(3) = 6.9$, $p = 0.077$). The source of this result is evident in Fig. 2 in the bottom panel of column 3, and in conditional effects of clear speech implied by the model. For instance, while there is no significant effect of clear speech on accurate perception of vowel height in AO (tense: $\beta = 0.218$, $p > 0.1$; lax: $\beta = -0.334$, CI = [−0.7, 0.1], $z = -1.650$, $p = 0.099$) and no reliable effects in AV due to perceivers' ceiling performance, in VO clear speech provides a significant advantage for both lax vowels ($\beta = -0.819$, CI = [−1.1, −0.6], $z = -6.303$, $p < 0.001$) and tense vowels ($\beta = -0.487$, CI = [−0.8, −0.2], $z = -3.184$, $p = 0.002$), though the effect on lax vowel height perception is more pronounced.

Finally, we examined perception of the [round] feature. Rounding perception was robust across modalities, with no significant Modality × Style × Tensity interaction ($\chi^2(2) = 2.7$, $p > 0.1$), nor any significant two-way interactions (ps > 0.3). The only relevant effect for the present study was an overall significant effect of speech style ($\chi^2(6) = 16.0$, $p = 0.014$). However, from Fig. 2 and model estimates (see Table A1 in the Appendix) this effect appears to be largely due to restricted clear-speech advantages for lax vowels in AO ($\beta = -0.349$, CI = [−0.65, −0.05], $z = -2.263$, $p = 0.024$) and AV ($\beta = -1.02$, CI = [−1.7, −0.3], $z = -2.882$, $p = 0.004$).

In summary, clear speech generally showed benefits for vowel perception along several featural dimensions, particularly for lax vowels in backness and height. However, accurate perception of vowel tensity is dependent on both modality and speech style, with clear-speech modifications confounding cues to lax vowels such that either benefits of clear speech disappear for lax vowels (as in AO and AV), or clear-speech effects reverse and become directly disadvantageous to the perceiver (lower accuracies relative to plain speech), as seen in visual-only presentation.

3.2. Mandarin perceivers

Fig. 3 shows results for Mandarin perceivers. The pattern of accuracy for tense vowel stimuli largely mirrored that for English perceivers. However, for lax vowels there were two main discrepancies. First, Mandarin perceivers appeared to rely more on visual information than English perceivers, as both VO and AV modalities show disadvantages for clear speech. Second, the clear-speech benefit in audio-only that was present for English perceivers disappears.

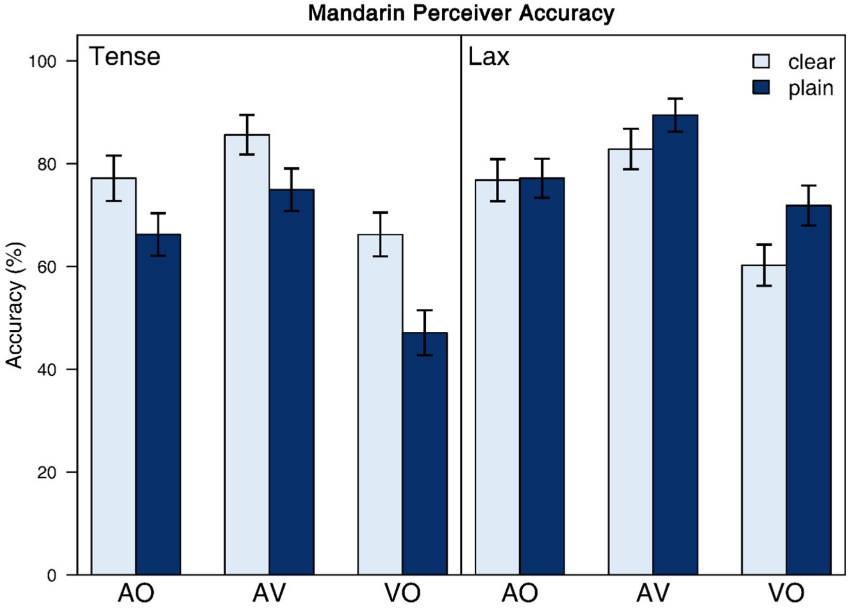

Fig. 3. Identification accuracy of Mandarin perceivers by speech style (clear, plain) and stimulus vowel tensity (tense, lax) in audio-only (AO), audio-visual (AV), and visual-only (VO) modalities. Error bars represent ±2 standard errors about the mean.

Statistical model analyses of Mandarin perceiver accuracy followed those for English perceivers and used a similar mixed effects model as Eq. (1), but with the addition of a random slope for Stimulus Vowel Tensity on Subject (see Table A2 in the Appendix for the full model specification).

As with English perceivers, the Mandarin group exhibited a significant Modality × Style × Tensity interaction ($\chi^2(2) = 27.4$, $p < 0.001$). Clear speech yielded significant advantages for tense vowels across modalities (AO: $\beta = -0.666$, CI = [−0.8, −0.5], $z = -7.637$, $p < 0.001$; AV: $\beta = -0.813$, CI = [−1.0, −0.6], $z = -8.157$, $p < 0.001$; VO: $\beta = -0.929$, CI = [−1.1, −0.8], $z = -11.76$, $p < 0.001$), while lax vowels were more accurately perceived in plain speech in AV ($\beta = 0.655$, CI = [0.4, 0.9], $z = 5.830$, $p < 0.001$) and VO ($\beta = 0.585$, CI = [0.4, 0.7], $z = 7.369$, $p < 0.001$). No significant difference between clear and plain speech was found for lax vowels in audio-only ($\beta = 0.019$, $p > 0.1$). Thus, for Mandarin perceivers the information from visual cues to clearly spoken lax vowels appears to have a greater effect on their vowel identification, as both AV and VO modalities showed the disadvantage that English listeners only showed in visual-only.

Regarding featural accuracy, Fig. 4 shows Mandarin perceiver accuracy on [tense], [back], [high], and [round] features by modality, speech style, and stimulus vowel tensity. The pattern of results is consistent with the general conclusion from the English perceivers; namely, the interaction between speech style and stimulus vowel tensity is primarily driven by direct tensity errors, though unlike the English perceivers this interaction is evident in all three modalities. Again, we conducted separate analyses for each feature.

Overall, the [tense] accuracy model showed a significant interaction between Modality, Style, and Tensity ($\chi^2(2) = 36.8$, $p < 0.001$). This was driven by a significant clear-speech advantage for tense vowels in AO ($\beta = -0.859$, CI = [−1.1, −0.7], $z = -8.765$, $p < 0.001$), AV ($\beta = -0.932$, CI = [−1.1, −0.7], $z = -8.941$, $p < 0.001$), and VO ($\beta = -0.976$, CI = [−1.1, −0.8], $z = -12.12$, $p < 0.001$), but a clear-speech *disadvantage* for lax vowels in AV ($\beta = 0.784$, CI = [0.5, 1.0], $z = 6.445$, $p < 0.001$) and VO ($\beta = 1.187$, CI = [1.0, 1.4], $z = 12.82$, $p < 0.001$), and no effect of speech style for lax vowels in audio-only ($\beta = 0.148$, CI = [−0.1, 0.4], $z = 1.408$, $p > 0.1$). Therefore, the general pattern of conflict between clear speech and lax vowel production in the overall accuracy model and in Fig. 3 is directly linked to clear-speech-induced confusions between tense and lax vowel pairs. All other featural errors, as reviewed below are minimal and hardly modulated by speech style.

Mandarin perceiver accuracy on vowel backness showed a significant Modality × Style × Tensity interaction ($\chi^2(2) = 6.5$, $p = 0.038$), with a clear-speech advantage for both tense and lax vowels in AO (tense: $\beta = -0.394$, CI = [−0.7, −0.1], $z = -2.953$, $p = 0.003$; lax: $\beta = -0.942$, CI = [−1.3, −0.6], $z = -5.302$, $p < 0.001$) and VO (tense: $\beta = -0.894$, CI = [−1.3, −0.5], $z = -4.240$, $p < 0.001$; lax: $\beta = -0.638$, CI = [−0.9, −0.4], $z = -5.311$, $p < 0.001$). Accuracy on the [back] feature was at ceiling in AV.

Vowel height was near-ceiling in audio-only and audio-visual modalities, with the only significant effect of speech style arising for lax vowels in VO ($\beta = -0.530$, CI = [−0.7, −0.3], $z = -5.098$, $p < 0.001$). And thus, while there are significant interactions between Style and Modality ($\chi^2(4) = 10.5$, $p = 0.032$) and Style and Tensity ($\chi^2(3) = 11.0$, $p = 0.012$), these effects are largely driven by the single clear-speech benefit for lax vowels in visual-only.

Finally, Mandarin perceiver accuracy on vowel rounding was unaffected by speech style when visual information was present (i.e., in AV and VO), but in audio-only there was a clear-speech advantage for both tense ($\beta = -0.288$, CI = [−0.5, −0.1], $z = -2.455$, $p = 0.014$) and lax ($\beta = -0.257$, CI = [−0.5, 0.0], $z = -2.052$, $p = 0.040$) vowels, resulting in a minor overall effect of speech style ($\chi^2(6) = 12.7$, $p = 0.048$), though Fig. 4 and the above numerical analysis confirm that this effect is restricted to AO.

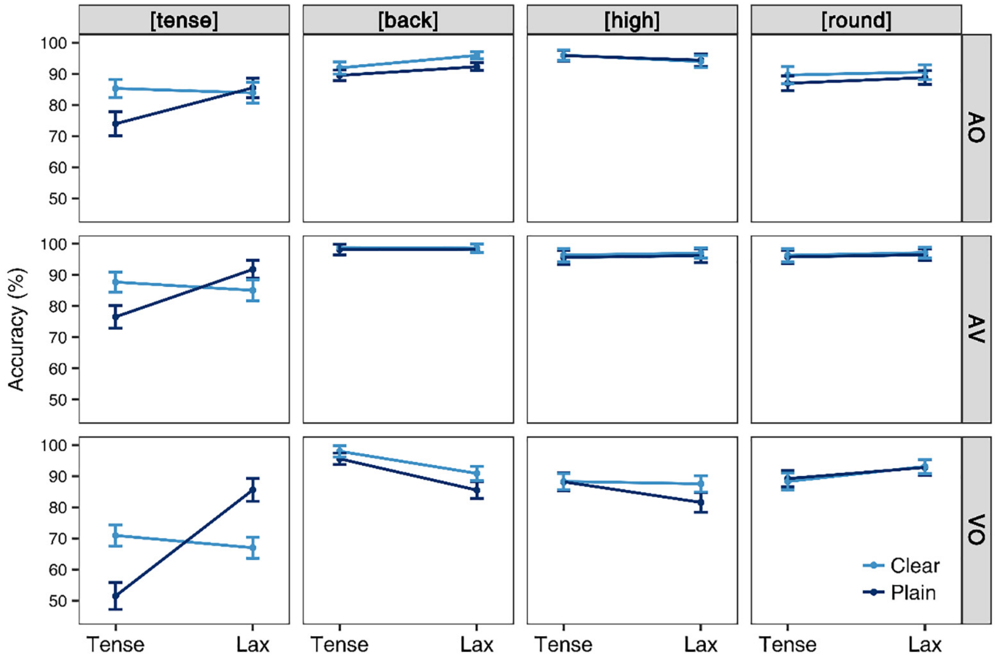

Fig. 4. Identification accuracy of Mandarin perceivers by feature ([tense], [back], [high], [round]), speech style (clear, plain) and stimulus vowel tensity (tense, lax) in audio-only (AO), audio-visual (AV), and visual-only (VO) modalities. Error bars represent ± 2 standard errors about the mean.

In summary, Mandarin perceivers exhibit many of the same vowel perception patterns as English listeners, particularly with respect to vowel height and backness, and while they show a parallel clear-speech disadvantage for lax vowels in VO, the extension of this interaction to AV and AO modalities indicates that the cues Mandarin perceivers rely on (particularly in the context of visual information) are less robust to changes in speech style, and thus may be more closely utilizing the same dimensions as those manipulated in clear speech.

3.3. Predicting perceptual patterns from acoustic and visual cues

The next set of analyses aimed to link the acoustic and visual measures of individual vowels to the pattern of perceptual data, and to provide a more comprehensive window on the effect of clear speech on the acoustics by looking across multiple cues, and by considering the impact of clear speech relative to other factors such as talker differences. The stimuli used here were part of a corpus of vowels that were collected, measured, and analyzed in two prior studies (Leung et al., 2016; Tang et al., 2015). In Section 3.3.1 we model audio-only perception, and in Section 3.3.2 visual-only. Audio-visual perception cannot be modelled in the present study because the listener results, which serve as a benchmark against which model predictions are evaluated, were at ceiling and therefore errors were too sparse to discern any reliable patterns.

Analyses were conducted in four stages. These followed different aspects of the C-CuRE framework to address three distinct questions. First, we started with a review of the effect of clear speech on each cue found in our prior studies. Second, we sought to understand the overall contributions of different factors to the variance in each cue (*Variance* models). For this, we conducted a new analysis of this data to determine how much variance speech mode contributes over and above other factors.

Third, we asked what cues listeners were using, and the relative cue weight of each cue. In these *inferential* models, listener responses were predicted from a multinomial logistic regression relating the measured cue-values to the likelihood of choosing each of the six vowel categories. We then assessed the relative weighting of each cue by examining changes in model performance when a given cue was included or excluded. Further, we allowed each cue to interact with speech style to determine if cues were used differently depending on the speech mode.

Fourth, we used a similar model to predict perceptual performance from the combination of cues (*predictive* model). In this stage we constructed *predictive* models. These sought to mirror as closely as possible the listener's task in the experiment, but without access to any of the listeners' results. The predictive models, as in McMurray and Jongman (2011), were meant as a form of ideal observer analysis to determine how much information is in the signal (across cues) to potentially support categorization, and to determine to what extent errors shown by the listeners may reflect ambiguity in the statistical structure of the categories in the language. Here, models are evaluated by the degree to which they exhibit the same pattern of accuracy across conditions. In particular, adequate predictive models should mirror listener response patterns in showing a clear-speech advantage for both tense and lax vowels in audio-only, while showing a tensity-dependent reversal in visual-only with model accuracy being better in plain than in clear speech for lax vowels, in contrast with tense vowels, which should retain the clear-speech advantage.

It was unclear a priori whether listeners are best modelled as if they are using raw cue values, or if they may be using cues whose values have compensated for talker differences. Thus, both inferential and predictive models were conducted separately with two classes of cues as predictors of the vowel (either the perceived vowel in the inferential models or the produced vowel in the predictive models). First, we examined the raw cues as predictors of identification. Second, we examined the same cues after *compensating* for talker (following the C-CuRE approach of McMurray & Jongman, 2011). By examining different cue types, we aimed to extend the McMurray and Jongman (2011) findings on fricative perception to several new conditions: a different phonetic system (tense/lax vowels), different modalities, and different speech styles.

3.3.1. Audio-only models

Summary of phonetic analyses. Our prior phonetic study on the acoustic characteristics of these tokens revealed significant overall expansions of the vowel space in clear speech. There was also significant vowel lengthening, though tense and lax vowels were differentially impacted by such influences of speech style (Leung et al., 2016). For instance, tense vowels displayed greater lengthening than lax vowels, with this set further distinguished according to height (high vowels exhibiting greater lengthening in clear speech).

Regarding formant frequencies, clear vowels were produced with more peripheral formant patterns than plain vowels. The vowels showing the greatest influence of speech style were the following. The high front lax vowel /ɪ/ had a higher F2 and F3 in clear speech than in plain speech, indicating that clear /ɪ/ was more fronted and less rounded than plain /ɪ/. Clear low vowels showed higher F1 than their plain counterparts; further, the low tense vowel /ɑ/ had a lower F2 in clear than in plain speech, the above two results indicating greater tongue body retraction and lowering in clear speech. Clear rounded vowels /u, ʊ/ had a lower F2 than their plain counterparts, reflecting front cavity expansion in clear speech consistent with either tongue retraction, lip protrusion, or both.

Finally, a measure of vowel-inherent production dynamics, *spectral change*, showed a significant decrease in the high tense vowel /i/, while lax back vowels /ʌ/ and /ʊ/ became more dynamic in clear relative to plain speech.

More relevant to the impact of speech style on accuracy of identification in auditory perception, however, is the relative separation between tense-lax vowel pairs in clear versus plain speech. Leung et al. (2016) found that while the relative acoustic difference between tense and lax vowels increased in clear speech, for a number of cues and vowel pairs the opposite pattern was obtained, consistent with the potential conflict between global clear-speech modifications and gestural distinctions between tense and lax vowels.

The relative difference in F1 between tense and lax vowels reduced in clear speech relative to plain for the pairs /i, ɪ/ and /ɑ, ʌ/, in both cases due primarily to changes in lax vowel height in the direction of the tense counterpart (i.e., lowering for /ʌ/, raising for /ɪ/). F2 and F3 only showed a clear-speech reduction in tensity distinctions for the high front vowels, again due to movement of /ɪ/ toward /i/ in clear speech. Leung and colleagues posited such effects for the high front vowels could be due to /i/'s extreme peripheral position not permitting much further raising or fronting in clear speech (see Granlund et al., 2012, for further discussion of this argument).

Regarding dynamic formant measures, differences in spectral change in tense-lax pairs consistently increased in clear speech, with such increases primarily due to greater spectral change observed in clearly spoken lax vowels. Finally, vowel duration generally increased for tense vowels and decreased for lax vowels in clear speech, resulting in greater separation between tense-lax pairs and therefore greater predicted accuracy on tensity perception (due to duration) in clear speech. No significant changes in fundamental frequency due to speech style were observed, while clearly spoken vowels were generally louder than their plain counterparts, though only by an average of 1 dB. For summary statistics on the acoustic characteristics of clearly and plainly spoken tense and lax vowel stimuli in this study, see Table B1 in the Appendix.

Variance models. We first sought to understand the relative contribution of clear speech to each acoustic cue in the context of other factors that may play a role. To this, we extended the analyses of Leung and colleagues by conducting a series of hierarchical regressions to identify the variance in each cue that was associated with several factors. These models used eight acoustic cues measured in Leung et al. (2016)—F1, F2, F3, spectral change (SC), spectral angle (SA), intensity (Amp), fundamental frequency (f0), and word duration (Dur).[2]

In these regressions (c.f., McMurray & Jongman, 2011; Jongman & McMurray, 2017), the cue value of each token was the dependent variable, and predictors included the talker, vowel tensity, and speech style. This was a standard linear regression, and the predictor values were dummy coded. Each set of predictors was entered into the model as a whole (e.g., all the dummy codes for talker), and each level of the model added a new set of codes (e.g., all the codes for vowel tensity), along with those of the prior model. Talker was added first, followed by tensity, followed by mode. After each model, the unique variance was extracted as the difference in R^2 between consecutive models.

Fig. 5 presents the unique variance in that parameter attributable to each factor, and demonstrates that variation in raw cue values is primarily due to differences between talkers and vowel tensity, with speech style differences only accounting for substantial additional variance in duration. While this pattern may appear to be in conflict with Leung et al. (2016) who reported effects of speech style (both main effects and as interactions with vowel tensity), we note that Fig. 5 addresses a different question. Namely, it asks what proportion of the total variation in vowel tokens can be attributed to different sources in a hierarchical manner, first accounting for talker differences, then vowel tensity, then speech style. This approach was taken because it is more relevant to the problem perceivers are tasked with. Finally, we should note that while the variance analysis is informative as to the general structure of variance in acoustic cues, it remains distinct from the analysis of cue contributions in the inferential and predictive models below, as it considers each cue independently, with no optimization of cue weights in the prediction of listener responses (in the case of the inferential model) and vowel categories (in the case of the predictive model).

[2] Of the cues investigated, intensity is potentially compromised in the present study due to the fact that all stimuli were amplitude-normalized prior to their presentation to listeners. However, this does not mean that all effects of intensity differences by tensity and style have been eliminated, as more intense vowels should also show an amplification of formant frequencies that is less affected by global amplitude normalization. Nevertheless, this manipulation should be kept in mind when interpreting the intensity cue in the model.

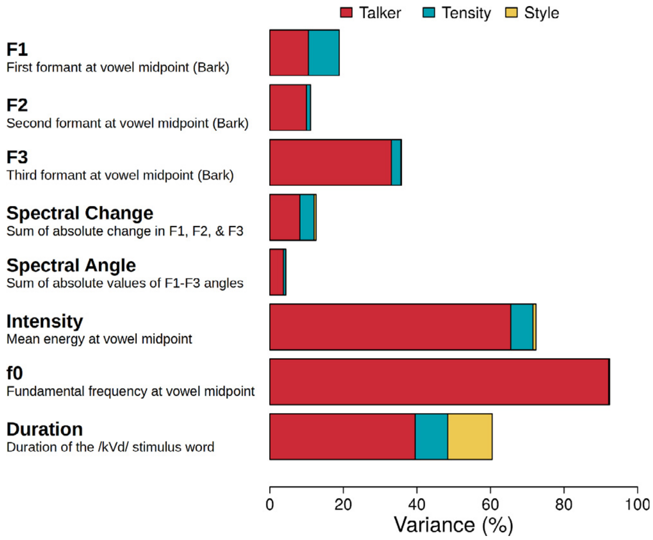

Fig. 5. Acoustic parameter definitions and relative variance accounted for by Talker, Tensity, and Style, where tensity effects are considered in addition to variance already captured by talker differences, and style effects derive from the further variance explained by an interaction between Style and Tensity.

Inferential models. The goal of the inferential models was to identify the relative contribution of each cue, and the impact of speech style on their utility.

These were done by training a series of multinomial logistic regression models to predict listener vowel choices (6AFC) from the acoustic cues described previously. We identified the unique contribution of individual cues by measuring changes in model performance when a given cue is excluded from the model. We assessed the relative impact of speech style on the predictive power of each cue by examining changes in model performance when an interaction between a given cue and speech style is added to the baseline model with only main effects of each cue. Separate models were trained on both raw and talker-compensated cues.

We started by fitting a baseline model. Here, the listener's response on each trial was the dependent variable and the reference category was /i/. The predictors included the eight cues as main effects, and a dummy-coded effect of listener. Predictors were *z*-scored to improve model convergence, though we should note that this transformation has no substantive effect on the outcome. We further note that multinomial logistic models have no way to account for repeated measures by listener (a mixed effects implementation is not yet available); our approach of including listener as a fixed effect is merely to capture overall listener biases in category choice. Thus, these models are not intended to make strong inferential claims. Rather, we use them descriptively to evaluate the relative contribution of each cue.

From the baseline model we computed model fit as the model's accuracy on the data on which it was trained, both as a whole and separately for the tense/lax contrast within each of the three modalities studied. Next, we refit the model, dropping each cue in turn to examine the change in accuracy when that cue was lost. This was meant to address our first question about the relative importance of each cue. Next, we added interaction terms for the interaction of speech style with each cue. These were added one by one and compared to the baseline model to determine whether listeners used each cue differently in clear speech.

Finally, this whole procedure was repeated using talker-compensated cues as the inputs to both models. To perform talker compensation, we ran a series of linear regressions that were similar to the variance models. Here, the cue value of each token was predicted from the talker (a set of dummy codes; vowel tensity and mode were not included as these were factors we wanted to investigate). We then saved the residual (the difference from prediction) as the new "compensated" cue value, after removing the effect of talker. These were then used as independent variables in the baseline and second-order models.

Results of the inferential models are shown in Table 1, which is organized as follows. In the first line, baseline model results are shown for raw cues (left half) and talker-compensated cues (right half). In this model, the eight cues are modelled additively, without any potential interactions. The next block of rows presents changes in model performance (relative to baseline) when each cue was excluded. Here, the utility of a cue should be indicated as a decrease in accuracy, or an increase in the Deviance statistic (D). Finally, the third block shows changes in model performance due to inclusion of an interaction of speech style with each cue. For example, in the F1 row in the final block, values represent differences between a model including an F1 × Style interaction alongside the other seven cues (without an interaction), and the baseline model that has no interactions. Thus, if speech style interacts with a cue, models in the third block should show an increase in model performance relative to baseline (simply due to changes in the size of the predictor set), or a decrease in the deviance statistic.

Table 1 suggests that listener responses are primarily predicted by the first two formants, both in raw cue and talker-compensated cue models. While F1 contributes to the pattern of identification for all three vowel pairs, the unique contribution of F2 is primarily limited to the low vowels /ɑ, ʌ/. The next two consistent predictors are the dynamic spectral parameters (SC and SA), which contribute to overall model accuracy in both cue models (raw and talker-compensated), though their impact on different vowel pairs is more variable than for F1 and F2. Loss of information about SC/SA is detrimental to the prediction of listener responses to high front vowels, but actually results in improved performance on high back vowels relative to when they are included in the model, though the size of this negative effect is smaller than the positive contribution for high front vowels. Fundamental frequency is primarily informative as a raw cue, which could be capturing general differences in talker accuracies, while duration, intensity, and the third formant frequency exhibit relatively minor unique contributions to model fit.

Thus, in general, whether models were trained with raw cues, or after compensating for talker mean differences, the relative contributions of each cue in predicting listener responses in the inferential model were similar. However, we will return to this comparison in the predictive models, which more directly replicate the way models were evaluated as ideal perceivers in McMurray and Jongman (2011).

As for the effect of speech style on individual cue utilization, very little change can be observed in model fit for any specific

Table 1

Results of inferential models predicting listener responses from both raw and talker-compensated acoustic cues in a multinomial logistic regression. The first row shows the Deviance statistic (D), which equals −2 times the log-likelihood of the model, the overall classification accuracy, and the accuracy on each tense-lax pair (/i, ɪ/, /ɑ, ʌ/, /u, ʊ/). Relative changes in these statistics from models excluding each parameter are shown in the next block, while changes due to the addition of an interaction term with a given cue are shown in the final block.

	Raw Cues					Talker-Compensated Cues				
Baseline	*D*	Acc.	/i, ɪ/	/ɑ, ʌ/	/u, ʊ/	*D*	Acc.	/i, ɪ/	/ɑ, ʌ/	/u, ʊ/
	4974	73.5	94.0	89.2	82.9	5006	73.2	94.6	88.7	82.0
Change relative to baseline due to dropped cues										
F1	1916	−12.4	−11.7	−0.8	−4.9	1222	−7.2	−3.6	−2.1	−1.9
F2	1050	−7.5	0.2	−20.6	0.1	778	−6.1	−0.2	−15.5	0.0
F3	108	0.0	−0.4	0.2	0.1	41	−0.6	−0.5	−0.7	−0.2
SC	313	−1.7	−1.8	−0.4	0.0	292	−1.5	−2.7	0.1	0.6
SA	214	−1.3	−2.3	−0.6	1.0	287	−2.2	−2.4	−0.4	0.0
Int.	206	−1.1	−3.3	−0.4	0.1	34	−0.2	−0.5	0.0	0.1
f0	813	−5.9	−8.3	−4.3	−0.5	78	−0.3	0.2	−1.0	−0.3
Dur.	132	−0.4	−0.7	−0.3	−0.1	65	−0.6	−0.2	−0.2	−0.4
Change relative to baseline due to added interaction of cue with style										
F1	−56	0.6	0.8	0.1	0.4	−102	0.6	0.0	0.3	1.3
F2	−34	0.6	0.7	0.2	0.5	−89	0.7	0.3	0.3	1.1
F3	−41	0.4	0.6	−0.1	0.6	−95	0.5	0.2	0.2	0.9
SC	−33	0.4	0.6	−0.1	0.6	−102	0.7	0.1	0.4	1.1
SA	−22	0.4	0.5	−0.2	0.4	−130	0.8	0.2	0.4	1.1
Int.	−30	0.4	0.4	−0.1	0.6	−117	0.6	0.1	0.2	1.2
f0	−48	0.5	0.6	0.1	0.4	−101	0.6	0.1	0.1	1.0
Dur.	−54	0.8	0.7	−0.1	0.8	−110	0.3	0.1	0.1	1.0

cue interaction. All increases in model accuracy were between 0.3% and 0.8%, with high front vowels exhibiting the greatest change, followed by the low vowel pair, and lastly the high back vowels. And while no particular cues emerge as uniquely modulated by speech style, the relative contribution of speech style is significant in all of the above models (as reflected in deviance changes, ΔD, greater than the critical value of 18.3 on 10 degrees of freedom). This suggests that listeners are not reweighting or recalibrating their use of any particular cue as a function of speech style, because the model improves similarly for each cue that is allowed to vary according to style (i.e., when the ideal perceiver is assumed to apply independent weights to a given cue in clear and plain speech).[3]

Predictive models. Next, we fit predictive acoustic models in a manner designed to replicate the listener task as closely as possible. Here, models had no access to the listeners' behavior, but were trained to predict the talker's intended vowel. The goal was to characterize the degree to which the listener pattern of accuracy would emerge from the statistical structure of the cues.

Again, multinomial logistic regressions were trained on the 12 speakers' data in the Leung et al. (2016) database. However, this time the dependent variable was the target vowel intended by the talker (i.e., we are no longer directly modelling listeners' responses so only the target category is relevant). Predictive models were first trained on the 12 speakers in the corpus (Leung et al., 2016; Tang et al., 2015) whose data was not presented to listeners. Next, the data from the six speakers that were presented to listeners is used as the test data for model evaluation.

Three separate models were considered initially, each with all eight cues discussed above, but differing in whether the cues were entered into the model in their raw form, as talker-compensated cues (i.e., the same residuals used in the inferential models), or as talker + style-compensated cues. This latter set of cues used the same regression model as the talker-compensated cues, but with a further predictor for speech style. It was intended as a secondary way to test whether cues might be used differently as a function of style. Each model was then tested on the six speakers' data that was used for the perception task and not included in the training set.

To simulate the effect of background noise (to mimic auditory constraints placed on listeners), we randomly perturbed each *z*-scored cue. This was done at two levels: with a SD of 0.25 (roughly corresponding to 75% overall model accuracy, close to, but slightly lower than listener averages), and with an SD of 0.5 (corresponding to 65% overall model accuracy). Mean results for each model over 1000 iterations (sampling different perturbations) are shown in Fig. 6.

Fig. 6 demonstrates that all three models closely approximated listener responses to tense vowel stimuli, particularly the compensated cue models (talker- and talker + style-compensated cues). This replicates McMurray and Jongman (2011) and extends the results to new vowels and to clear speech.

However, the three models were less consistent in replicating the clear-speech advantage for lax vowel stimuli in audio-only. The raw cue model shows the expected pattern at all three noise levels (clean, $\sigma = 0.25$, and $\sigma = 0.5$), but the compensated cue models only show this result when noise is added, with the highest noise level yielding patterns similar to the raw cue model but higher in accuracy. The talker + style-compensated model even shows a clear-speech disadvantage under clean signal simulations.

Finally, Fig. 6 is consistent with the general picture of primarily talker-determined cue variability shown in Fig. 5. The talker + style-compensated model performs nearly identically to the talker-compensated model.

To unpack this inconsistency in tense vs. lax vowel results, we examined the relative patterns in model performance as a function of vowel tensity and speech style when only one cue

[3] Given the additional degree of freedom afforded by adding the interaction term, significant improvements in model fit across cues could also be an artifact of the model specification, and thus we caution against interpreting these results as evidence for some kind of global cue weighting mechanism.

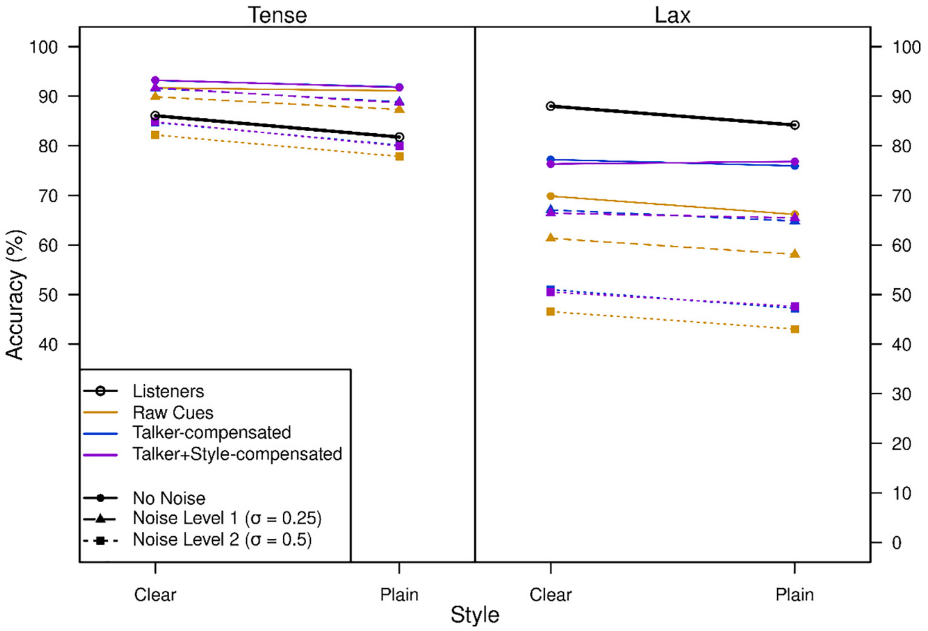

Fig. 6. Predictive audio-only (AO) model results from models fit to raw cues (orange lines), talker-compensated cues (blue lines) and talker + style-compensated cues (violet lines), with no simulated noise on the parameters (solid lines), simulated noise based on random perturbations from a normal distribution of mean 0 and standard deviation 0.25 (dashed lines), and simulated noise of σ = 0.5 (dotted lines). Listener accuracies on AO stimuli (solid black lines) are provided for reference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

at a time is used as a predictor. In other words, we considered the degree to which each cue is able to partition the vowel space on its own, and compared these patterns with overall listener performance. Fig. 7 displays the results of the single-cue predictive models, and for simplicity only considers talker-compensated cues in the absence of simulated noise.

From Fig. 7 we can see that several cues—F1, F2, SC, and Duration—follow the speech style pattern for tense vowel stimuli in showing a clear-speech advantage, while others, like Intensity (Amp), show a pattern contrary to the listener result (i.e., plain speech more accurate than clear). For lax vowel stimuli, only SC and Intensity are consistent with the listener pattern, while Duration, F2, and F3 (particularly F2) models show the opposing plain speech advantage. Thus, the aggregate model results on lax vowel stimuli in Fig. 6 appear to be driven by a handful of cues for which lax vowels are more distinct in clear speech than in plain, despite the fact that certain cues pose a notable disadvantage in clear speech.

This result provides evidence that even when aggregate improvement is shown in listener recognition of clear speech, on some dimensions (notably F2 and duration) clear speech can reduce the amount of information available in the signal to support perception. Further, some apparent code-based modifications, such as the relatively greater lengthening of vowel duration for tense vowels than for lax vowels (Leung et al., 2016), while consistent with the perceptual asymmetry shown in Fig. 7 (i.e., that duration modifications are more informative for tense than for lax vowels), are not completely utilized in perception. This is shown in the result that duration remains relatively uninformative in the inferential models, and is detrimental for clearly spoken lax vowels in the predictive models. One possible explanation for this latter result is that the mixed presentation of clear and plain speech styles caused cues with greater distributional variability under speech style modifications, such as vowel duration,[4] to be less reliable in perception. Listeners may then have shifted attention to more reliable cues, such as spectral change, which consistently index tensity in both clear and plain speech styles.

3.3.2. Visual-only models

We next considered models of visual-only perception. To convert the dynamic visual information in the videos to cues that could be used in a model, we used data acquired from automatic recognition of facial landmarks in Tang et al. (2015) as our independent variables. This was then used to predict both listener responses (the inferential model) and serve as a model of the general six-alternative forced-choice task given to listeners (the predictive model).

Summary of phonetic analysis. Tang et al. (2015) showed that speakers modify their speaking style to produce clear speech with exaggerated visual cues for vertical and horizontal lip stretching (VLS and HLS, respectively), lip rounding (LR), jaw displacement (JD), and duration (Dur.). In particular, in clear speech, speakers showed greater vertical lip stretch and jaw displacement across vowels, greater horizontal lip stretch for front unrounded vowels, and greater degrees of lip rounding for rounded vowels than in plain speech. Further, all clear vowels were longer than their plain counterparts, similar to what was observed in the acoustic analysis (Section 3.3.1), though visually the movement of articulators need not coincide exactly with the onset/offset of audible speech.

Crucially for the current research, the articulatory results also reveal that tense and lax vowels were modified to the same extent, on average, in clear speech. This average equivalence in effects of speech style partly derives from the fact

[4] Clear tokens of lax vowels ($\bar{X}$ = 205 ms) are nearly as similar in duration to plain tense vowels (272 ms) as plain lax vowels (165 ms).

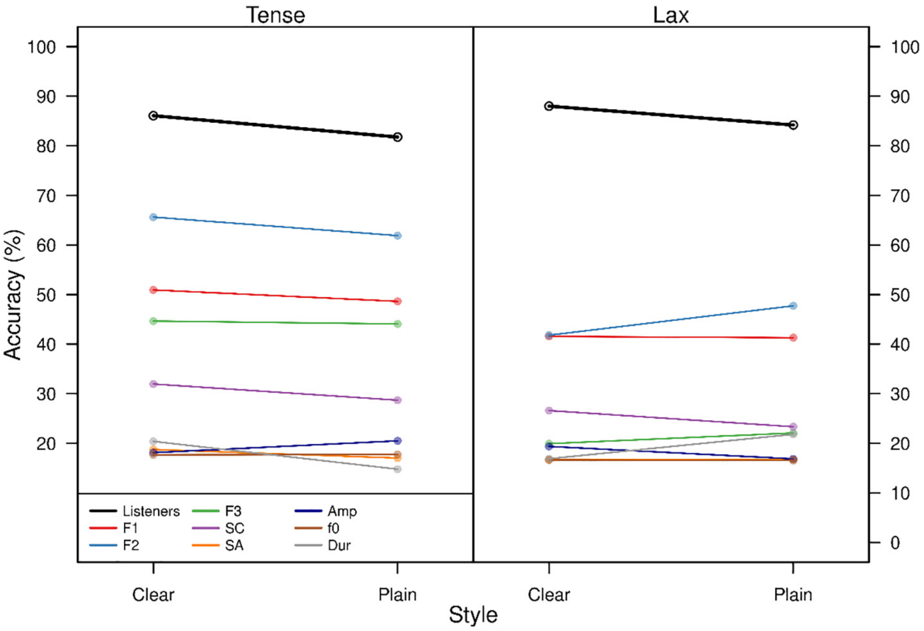

Fig. 7. Predictive audio-only (AO) model results from models fit to single talker-compensated cues. Listener accuracies on AO stimuli (solid black lines) are provided for reference.

that for some cues, such as vertical lip stretch (VLS), clear speech resulted in greater similarity within vowel pairs, while for others, such as jaw displacement (JD), clear-speech modifications resulted in an increase in tense-lax distinctiveness. For summary statistics on the visual characteristics of clearly and plainly spoken tense and lax vowel stimuli in this study, see Table B2 in the Appendix.

Variance models. As with the acoustic analyses, we also conducted a hierarchical regression analysis to assess the overall amount of variance accounted for by each contributing factor. Fig. 8 displays visual parameter definitions alongside the relative variance in that parameter attributable to talker, vowel tensity, and speech style differences, as Fig. 5 did for acoustic cues. Relative to Fig. 5, we can see that speech style accounts for greater variance on top of talker and vowel tensity differences than in the audio-only modality, with notable variation in VLS with speech style. The duration results are comparable to Fig. 5, though lower in talker variance due to the relatively less precise measurement of word onset and offset in the video signal.

Inferential models. The inferential model was conducted similarly to the acoustic model, only with the five visual cues as predictors and the perceiver responses in the visual-only condition as the dependent variable.

Results are presented in Table 2. Several key patterns in cue weight structure are evident in Table 2. First, Duration exhibits the greatest unique contribution to model fit in both raw and talker-compensated cue models; however, both Vertical and Horizontal Lip Stretch parameters are nearly as informative, particularly when compensating for talker. This is notable given that some redundancy in lip position and movement information is expected between the three lip parameters. While raw cue model accuracy was responsive to all the cues (to some extent), for every cue but Jaw Displacement the overall results were primarily driven by the high back vowels /u, ʊ/—other vowels showed little to no difference.

In the talker-compensated model, only HLS and Duration retain the strong link with high back vowel performance, while VLS, Lip Rounding, and Jaw Displacement all play a greater role in discriminating the tense and lax low vowels /ɑ, ʌ/. Generally speaking, compensating for talker, no single cue appears to be closely linked to tensity perception in the high front vowels, though overall model accuracy on the /i, ɪ/ pair is not notably lower than the other two (low vowel accuracy is the lowest in the raw cue model, and low and high-front pairs are equal in the talker-compensated cue model). Finally, in terms of overall accuracy the talker-compensated model outperformed the raw cue model by 6%, contrary to the audio-only condition where the two models were equivalent (73.5% vs. 73.2%, respectively, for raw and compensated cues).

Regarding interactions with speech style, the inclusion of interaction terms had a substantially greater effect in the VO model than in the AO model. This was perhaps expected given the large amount of variance associated with speech style (Fig. 8). In the raw cue models, we saw the greatest impact of cue by speech style interactions for LR and JD (overall accuracy increased by 1.2% and 1.1%, respectively). For jaw displacement this effect was carried by the high back vowels. In contrast, for lip rounding this effect was carried by the low vowels. This result is perhaps surprising given that we expect a greater dependence of the /ɑ, ʌ/ distinction on jaw displacement, and similarly for lip rounding and /u, ʊ/. However, considering the interaction between speech style and vowel tensity in VO perception in Section 3.1 (and replicated below in the predictive model results), this result could be capturing the fact that the conflicting nature of such cues in clear speech makes them poorly modeled as a function of speech style, because their resultant down-weighting in clear speech may ultimately lead to less overall improvement in the predictive power of the model.

This general pattern is retained for talker-compensated cues (i.e., LR × Style improves /ɑ, ʌ/, JD × Style improves

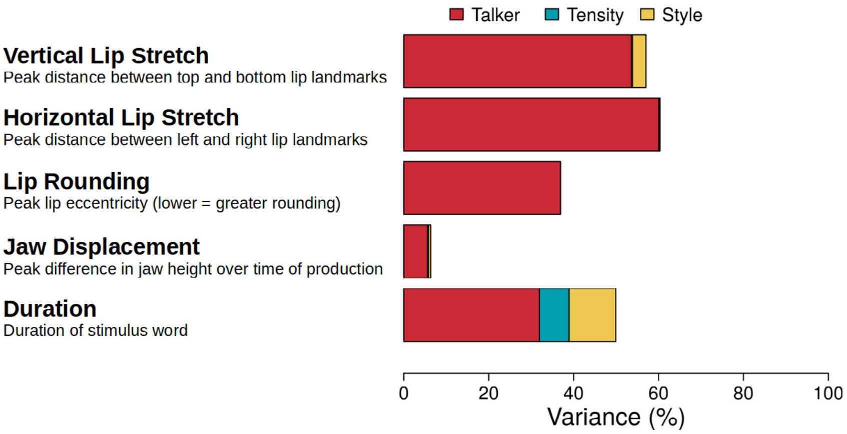

Fig. 8. Visual parameter definitions and relative variance accounted for by Talker, Tensity, and Style, where tensity effects are considered in addition to variance already captured by talker differences, and style effects derive from the further variance explained by an interaction between Style and Tensity.

Table 2
Results of inferential models predicting listener responses from both raw and talker-compensated visual cues in a multinomial logistic regression. The first row shows the Deviance statistic (D), which equals −2 times the log-likelihood of the model, the overall classification accuracy, and the accuracy on each tense-lax pair (/i, ɪ/, /ɑ, ʌ/, /u, ʊ/). Relative changes in these statistics from models excluding each parameter are shown in the next block, while changes due to the addition of an interaction term with a given cue are shown in the final block.

	Raw Cues					Talker-Compensated Cues				
Baseline	*D*	Acc.	/i, ɪ/	/ɑ, ʌ/	/u, ʊ/	*D*	Acc.	/i, ɪ/	/ɑ, ʌ/	/u, ʊ/
	13,963	26.4	62.0	57.1	65.9	12,527	32.4	60.5	59.9	63.8
Change relative to baseline due to dropped cues										
VLS	601	−3.2	0.0	−0.6	−9.9	664	−3.5	0.1	−1.2	−5.9
HLS	338	−1.6	−0.4	−0.4	−2.0	790	−3.4	0.0	−1.4	−0.3
LR	171	−1.1	0.1	−0.5	−2.4	31	−0.2	0.0	−0.5	−0.2
JD	374	−1.7	−0.6	−0.2	0.1	292	−1.5	−0.2	−0.8	−0.2
Dur.	993	−4.8	−3.2	−2.1	−13.9	829	−4.6	−1.2	−0.8	−12.4
Change relative to baseline due to added interaction of cue with style										
VLS	−192	0.8	0.7	−0.2	0.2	−218	1.4	0.3	1.5	1.3
HLS	−192	0.8	0.3	0.0	0.5	−388	1.7	0.1	1.1	1.1
LR	−275	1.2	0.3	1.2	0.1	−256	1.4	0.7	1.6	0.5
JD	−260	1.1	0.1	0.1	3.2	−269	1.4	0.2	1.2	2.5
Dur.	−80	0.4	0.0	0.5	0.8	−235	1.2	0.3	1.7	0.4

/u, ʊ/). However, the lip stretch parameters (VLS, HLS) interact more with speech style relative to raw cue models, with the increase in model accuracy primarily reflected in improved prediction of listener responses to back vowels (/ɑ, ʌ/ and /u, ʊ/ accuracies increased between 1.1 and 1.5%). Finally, as in the models examining the weighting of individual cues, high front vowels show little change.

Predictive models. Fig. 9 displays the results of raw, talker-compensated, and talker + style-compensated cue models alongside perceiver recognition patterns for visually presented tense and lax vowels. All three models capture the clear-speech advantage for tense vowels, though their overall accuracy was substantially lower than the perceivers'. The model based on talker-compensated cues shows the greatest predicted clear-speech advantage, while both raw and talker + style-compensated cue models show a smaller advantage but one that is more consistent with that observed for perceivers. For the critical reversal of speech style effects on lax vowels, however, both the talker-compensated and raw cue models showed the plain speech advantage exhibited by perceivers, though to a slightly lesser degree. However, the fully compensated model did not. Thus, in aggregate both raw and talker-compensated cue models are consistent with perceiver performance.

We next examined the predictive models' fit to single cues to determine whether clear speech had a similar benefit for each cue. That is, as in 3.3.1, this analysis assesses the degree to which partitions of the six-vowel space according to a single cue are consistent with listener accuracies on tense and lax vowels in clear and plain speech. Here we restrict our attention to the talker-compensated models, because they more closely fit the overall listener accuracy patterns, particularly for tense vowels.

Fig. 10 displays the results of independent cue models. Both Duration and Horizontal Lip Stretch show substantial clear-speech advantages for tense vowels, as expected from the listener data, while Vertical Lip Stretch shows a slight clear-speech advantage. The remaining two parameters (Lip Rounding and Jaw Displacement) perform at chance (0.167) and thus do not yield any difference between the two speech styles. In predicting lax vowel perception, however, only Duration is consistent with listener performance in showing a plain speech advantage. HLS and VLS lose the clear-speech advantage shown for tense vowels, but VLS performance is

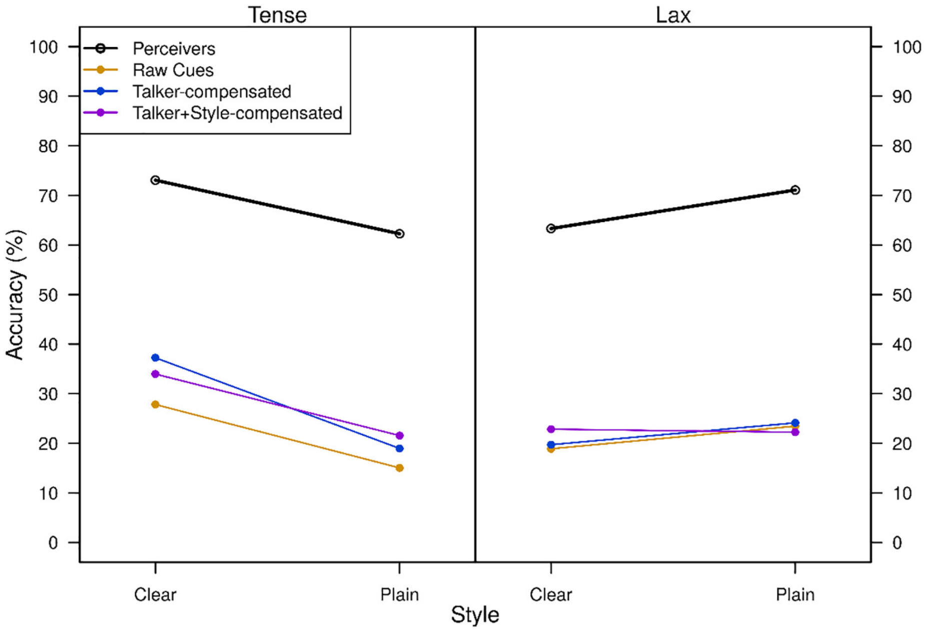

Fig. 9. Predictive visual-only (VO) model results from models fit to raw cues (orange lines), talker-compensated cues (blue lines) and talker + style-compensated cues (violet lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

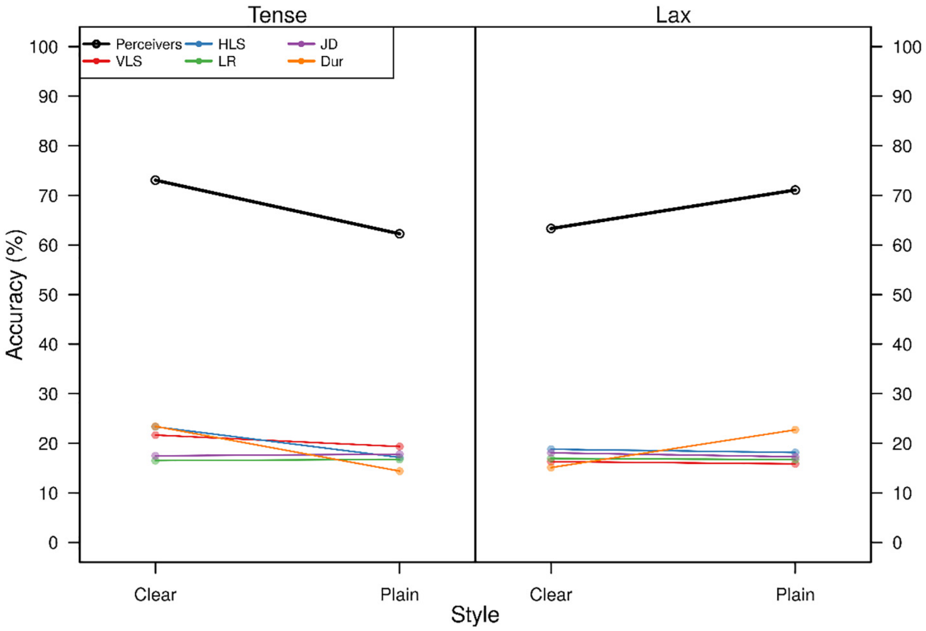

Fig. 10. Predictive visual-only (VO) model results from models fit to single talker-compensated cues. Listener accuracies on VO stimuli (solid black lines) are provided for reference.

around chance and HLS performance is only marginally better at just below 20%. Thus, we can conclude that Duration is clearly important in accounting for the conflicting cues generated by clear speech for vowel tensity perception, with the remaining cues consistent with the conflict (all advantages disappear for lax vowels) but not independently robust.

In summary, in modeling the impact of clear speech on visual cue parsing in the absence of acoustic information, multivariate cue models are consistent with perceiver response patterns in showing a clear-speech advantage for tense vowels, and a clear-speech disadvantage for lax vowels. The single cue that best reflects this reversal, however, is one that is also available in auditory perception: duration. And while the two lip stretch parameters show a clear-speech advantage for tense vowels that disappears for lax vowels, they are not able to independently show the robust advantage for plain speech that perceivers exhibit. Whether this result implies a priority for visual cues consistent with those available in the acoustic signal is unclear at present, and ultimately will require a larger sample of audio-visual perception errors (with noise in the visual display manipulated in addition to the acoustics) to test.

4. Discussion

Previous research indicates that the perception of clear speech depends on several factors, including the saliency of the source of information (acoustic or articulatory) (Maniwa, Jongman, & Wade, 2009; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998), the perceptual weighting of auditory and visual cues (Gagné et al., 2002; Helfer, 1997), and the linguistic experience of the perceivers (Bradlow & Bent, 2002; Fenwick et al., 2015). The goal of the present study was to provide a comprehensive approach to the study of clear speech by integrating acoustic, articulatory (facial), and perceptual data in an attempt to determine the extent to which these factors affect speech intelligibility. Specifically, we tested the perception of English tense and lax vowels by both native English and non-native Mandarin perceivers based on audio-visual, audio-only, and video-only input.

The tense/lax vowel distinction was selected because its acoustic and articulatory correlates are similar to those of the clear/plain speech distinction. For example, all vowels are lengthened in clear speech (Ferguson and Kewley-Port, 2002, 2007; Ferguson & Quené, 2014; Leung et al., 2016; among others), but vowel duration also serves as a cue to the tense-lax distinction. Similarly, clear speech is generally marked by an expansion of the vowel space, where changes in F1 and F2 from plain to clear mirror those from lax to tense. This similarity allowed us to establish whether the use of the same acoustic or articulatory properties differs depending on the communicative goal to be achieved. Further, the manner in which perceivers map variation onto different sources—speech style versus vowel tensity—provided insight into the distinction between global signal-based modifications and local code-based modifications.

Three lines of evidence were used to evaluate this distinction with respect to the present data. First, we evaluated the perceptual consequences of clear speech in AO, VO, and AV modalities in native, L1 listeners. Second, we compared native vs. non-native response patterns. Finally, we evaluated the fit between observed English perceiver responses and those predicted based on acoustic and visual cues.

4.1. Clear-speech enhancement in L1 listeners

Our perception results show that English perceivers generally exhibited a clear-speech advantage for both tense and lax vowels across modalities, consistent with the previous findings for clear-speech intelligibility of segments with both auditory (Ferguson & Kewley-Port, 2002) and visual input modalities (Gagné et al., 2002). However, perception of tensity was affected by both speech style and modality: while a clear-speech gain was observed for both tense and lax vowels in the AO modality, clear-speech modifications confounded visual cues to lax vowels, resulting in the absence of a clear-speech benefit in the AV modality, and a clear-speech disadvantage in the VO condition. A breakdown of perceivers' accuracy in terms of the features [tense], [back], [high], and [round] confirmed that errors consistent with a clear-speech disadvantage were due to tensity misperceptions in the VO modality.

In our prior articulatory study, we found no interaction between style (plain or clear) and vowel tensity, indicating that clear-speech articulatory modifications did not differ between tense and lax vowels (Tang et al., 2015). These clear-speech modifications included greater extent of movement and greater overall duration. In the present study, in the VO modality, these characteristics led perceivers to identify more vowels as tense overall, resulting in the observed worse performance on clear lax than plain lax vowels.

These interactive effects between speech style, input modality and vowel tensity support our prediction that the distinctions between signal- and code-based clear-speech modifications are also reflected in perception to differentially affect intelligibility. Our results indicate that code-based modification in clear speech did not occur universally across vowels and input modalities. In particular, clear speech aided tense vowel perception in the visual modality, but appeared to be detrimental in visual lax vowel perception.

This is similar to what was shown in our articulatory study of these vowels (Tang et al., 2015): clear-speech modifications, which also involve hyperarticulation, are compatible with the inherent features of tense vowels. In the case of vowel tensity, these modifications appear to confound a strong distinction between signal- and code-based modifications. For tense vowels, these modifications enhance the distinctiveness of the visual cues characterizing tense vowel categories, thus facilitating their increased intelligibility in clear speech. In contrast, our articulatory results showed that lax vowels underwent a similar degree of clear-speech modification as tense vowels (Tang et al., 2015). As lax vowels are characterized by less extreme (and shorter) articulation (Hillenbrand, Getty, Clark, & Wheeler, 1995), such clear-speech modifications conflict with the intrinsic features of the lax vowels, making them approximate their tense vowel counterparts and thus blurring category distinctions. As such, these signal-enhancing modifications of lax vowels result in vowels with tense vowel features which distort the phonemic category distinctions and consequently hurt perception. In this case, signal-based modifications (longer and stronger articulation) affect the primary cues to the contrast, and are therefore interpreted as code-based.

4.2. Native vs. non-native response patterns

Inclusion of native English and Mandarin perceivers served to further assess enhancement. While signal-based modifications have been shown to be beneficial to both native and non-native listeners, code-based modifications, which do not affect acoustic properties across the board but instead affect one or more properties to specifically distinguish one phoneme from one or more other phonemes were predicted to be only or more beneficial to native listeners, who have learned to associate specific patterns of cues with phonemic categories. English speakers have tense and lax vowels in their native system, and thus were expected to show a clear-speech advantage for both. Since Mandarin does not have lax vowels, Mandarin perceivers were expected to show no clear-speech benefit for lax vowels in the auditory modality.

Mandarin perceivers performed similarly to English perceivers, with two notable exceptions. First, Mandarin perceivers showed a clear-speech disadvantage for lax vowels not only in the VO modality but also in the AV modality. Thus, compared to English perceivers, Mandarin perceivers showed this disadvantage whenever visual information was available. This could derive from a greater reliance on visual information. This finding is consistent with previous research that has shown that non-native perceivers attend more to visual cues than native perceivers (Hazan et al., 2010).

Second, Mandarin perceivers did not show a clear-speech advantage for lax vowels in the AO modality. This result would seem to be an effect of native language background, given that Mandarin has tense but not lax vowels. Featural analysis showed that errors were almost entirely due to tensity misperceptions in all three modalities. Overall, the cues that Mandarin perceivers seem to rely on for the tense/lax distinction are also the cues that signal the clear/plain distinction.

This raises the possibility that L2 listeners lack the auditory skills necessary to properly parse the acoustic cues for tensity from other sources of variability in the signal (the clear-speech modifications). That is, they were less successful when signal-based modifications affected code-based cues.

However, more broadly Mandarin perceivers benefited from clear-speech in both auditory and visual perception of the tense vowels (which exist in Mandarin) but not for unfamiliar lax vowels. This confirmed our prediction with respect to effects of linguistic experience on clear-speech intelligibility. These results are consistent with the previous findings of the lack of or even detrimental clear-speech effects on L2 perception (Fenwick et al., 2015; Granlund et al., 2012; Smiljanić & Bradlow, 2011). The patterns exhibited by the Mandarin perceivers have further implications for signal versus code-based explanations of clear-speech intelligibility. First, the AO results indicate the effectiveness of code-based clear-speech perception, in that Mandarin perceivers were able to utilize code-based cues in improving tense vowel perception in clear speech, but failed to adopt the critical cues characterizing lax vowels to further improve lax vowel intelligibility. If their perception were based on signal-enhancing cues, we would have expected no difference in clear-speech effects between tense and lax vowel perception. Furthermore, the negative clear-speech effects on the AV and VO perception of lax vowels indicate that signal-enhancing clear-speech modifications (which reduce category distinctiveness) can be even more detrimental to non-native perceivers than to native perceivers, as non-natives appear to rely more on visual cues.

4.3. Direct prediction of English perceiver responses from acoustic and visual cues

Next, we related acoustic and articulatory measurements to the perception results to determine to what extent each cue contributes to perceivers' performance. Results from this inferential approach show that F1 and F2 are the main acoustic cues used by listeners, followed by spectral change and spectral angle. While speech style did not affect any specific cue in particular, it did have a small but consistent effect across all cues. In terms of visual cues, duration emerged as the primary cue, followed by vertical and horizontal lip stretch. In addition, speech style had a stronger influence in the visual than in the auditory modality.

Finally, we used three statistical models to determine which was most accurate at predicting the pattern of responses observed in the perceivers: raw cues, cues compensating for talker, and cues compensating for talker and speech style. These were trained based on either acoustic or visual cues.

Results from this predictive approach for acoustic cues indicate that all three models performed qualitatively similarly to listeners for the tense vowels. However, for the lax vowels, only the raw cue model replicated the perceivers' clear-speech advantage at all simulated noise levels. The compensated cue models only showed this pattern when noise was added. There was little difference in performance between the two compensated cue models, confirming that the variability in acoustic cues was largely determined by the talkers, not by the style. F1, F2, spectral change and duration all contributed to the clear-speech advantage for tense vowels, but only spectral change and amplitude did so for the lax vowels. This suggests that while listeners may track the current talker (and compensate for this variance), they may not do so with "style"—that is, they are not compensating specifically for modifications due to clear or plain speech.

For the visual cues, all three models again captured the clear-speech advantage for tense vowels observed in our perceivers. The reversal of this effect for the lax vowels (a clear-speech disadvantage) was captured by the raw and the talker-compensated cue models. Duration and horizontal lip stretch contributed to the clear-speech advantage for tense vowels while duration was the only cue that predicted the clear-speech disadvantage for lax vowels.

Integrating the results from each modality, the clear-speech disadvantage for lax vowels in video-only was seen to arise from the fact that visually, all vowels, both tense and lax, are lengthened to the same degree. Our modeling showed that duration does appear to drive perceivers' responses in the visual modality. As a result, there is a greater tendency for clear lax vowels to be misperceived as tense. Acoustically, however, clear speech does not affect all vowels in the same way or to the same degree. Clear tense vowels are lengthened much more than clear lax vowels, but clear lax vowels exhibit greater spectral change than clear tense vowels. In the auditory modality, these modifications resulted in a clear-speech benefit for both tense and lax vowels. However, this advantage disappeared for lax vowels in the audio-visual modality, suggesting that the conflicting visual information increased the ambiguity of the audio-visual information.

The results of our inferential and predictive models are consistent with previous results. Our previous analysis (Leung et al., 2016) demonstrated both signal- and code-based acoustic cue modifications. There was a signal-based, global increase in intensity from plain to clear speech, and code-based clear-speech effects were observed in the static formant frequency results, with the direction of plain-to-clear modifications resulting in more peripheral formant patterns in clear speech. Code-based modifications were also shown in the greater tense-lax contrast in clear speech for formant dynamicity and vowel duration. Our current inferential models showed that speech style consistently affected all cues in predicting English perceivers' performance, suggesting a role for signal-based modifications contributing to their performance. Code-based modifications closely aligned with the intrinsic vowel properties. Consistently, our audio-only predictive models showed that code-based modifications predominantly contributed to the clear-speech advantage for tense vowels, demonstrated by the lesser vowel reduction of /ɑ/ and /u/ in terms of their F1 and F2 modifications and the greater tense-lax contrast of formant dynamicity and vowel duration in clear speech. For the clear-speech advantage in lax vowels, the contribution of signal-based modifications came from the global increase in vowel intensity across speech style, whereas the contribution of code-based modifications was the greater spectral change (a critical feature characterizing lax vowels) in clear speech compared to plain speech.

In our previous articulatory analysis (Tang et al., 2015), the findings also showed both signal-based and code-based modifications. Comparing the visual cue modifications across vowel types, all clear vowels involved greater vertical lip stretch and duration than their plain counterparts, demonstrating signal-based modifications. Other visual cues exhibited code-based modifications since the direction of modifications corresponded to the characteristics of specific vowel pairs. Consequently, these clear-speech modifications enhanced the phonemic contrasts in visual speech. Specifically, greater horizontal lip stretch was found only for clear high front vowels (/i-ɪ/); greater lip rounding for clear rounded vowels (/u-ʊ/); and greater vertical jaw displacement for clear rounded and low vowels (/u-ʊ/ and /ɑ-ʌ/). Integrating our inferential model results with the findings from our previous articulatory analysis shows that such code-based modifications influenced English perceivers' responses in the raw cue model based on the effect of speech style, although talker-compensated cues also showed interactions between speech style and cues that involved signal-based modifications (vertical lip stretch and duration). With respect to vowel tensity, our visual-only predictive models showed that both signal-based (duration) and code-based modifications (horizontal lip stretch) predicted substantial clear-speech advantages for tense vowels, whereas only signal-based modifications (duration) accounted for the clear-speech disadvantage.

4.4. Implications for C-CuRE

Our previous research suggests that listeners overcome the ubiquitous variability in the speech signal by engaging in a data-explanatory approach. That is, listeners do not make decisions based on raw cues; instead, they build up expectations about what a segment produced by a specific talker and in a specific context should sound like and then compare these expectations to the observed signal. The model which inspired this, C-CuRE, was developed based on acoustic and perceptual studies of fricatives. It works similarly to the talker-compensated inferential models described above: listeners code cues relative to their expected values given that talker or coarticulatory context. These cues are then weighted and combined and a simple decision-making model (again, logistic regression) chooses the ultimate response.

In our initial evaluation of this model, we found that listener-like fricative identification could only be achieved by a model in which acoustic cues were interpreted relative to the talker and vowel context (McMurray & Jongman, 2011)—as seen here, raw cues were insufficient to drive high levels of accuracy. This suggests that listeners are actively forming expectations about what speech cues should sound like, and using the degree of match or mismatch as information for further processing. Supporting this, subsequent empirical work showed that providing listeners with such contextual information (i.e., an image of the talker) resulted in faster and more accurate fricative identification (McMurray & Jongman, 2015).

The present study extends our approach to vowel tensity and, more importantly, is the first attempt to incorporate visual cues. In general, C-CuRE is successful in that vowel classification based on compensated cues (for talker and style) was better than that based on raw cues. This suggests that the computational approach embodied in C-CuRE can be used to not only investigate the contributions of talker and vowel context but of speech style as well. However, in contrast to the fricative studies, the present results indicate that while compensation generally led to better classification, it did not always result in more perceiver-like categorization. One possibility for this discrepancy could be the nature of the categories under investigation. The English fricatives provided a perfectly balanced set of contrasts (4 places of articulation, each with a voiced and a voiceless member) which perhaps made them eminently suitable for the linear-based compensation of C-CuRE which employs a series of hierarchical linear regressions. Vowels, on the other hand, contrast along several additional dimensions which may require inclusion of non-linear transformations in our model.

A second possibility is that C-CuRE was designed mainly to deal with speech cues—the code. It assumes that each cue can be extracted reliably from the signal. However, this assumption is likely false. In fact, signal-based modifications due to clear speech may primarily serve to make it easier to extract and identify cue values from speech. Thus, future work with C-CuRE should consider modelling cue extraction as a probabilistic (rather than deterministic) and sometimes inaccurate process. This may help extend this model to consider signal-based processes rather than just code-based.

4.5. General implications for signal- vs. code-based explanations

In summary, findings from our perception study suggest that perceivers were affected by both signal- and code-based clear-speech modifications. However, code-based clear-speech cues that are aligned with vowel-intrinsic properties appear to be more effective than signal-based cues in aiding intelligibility. Comparisons of the native and non-native perceptual patterns indicate that perceivers need to be able to identify and utilize language-specific, code-based cues to improve intelligibility. A subset of these modifications contributed to the clear-speech perceptual advantage in both production (audio or visual) and perception domains. However, perceptual patterns can be primarily predicted by code-based cues; when only signal-based clear-speech modifications influence perception, this leads to greater chance of misperception, as exemplified by the clear-speech disadvantage of lax vowels in visual-only.

The alignment of code- and signal-based modifications challenges a strong distinction between signal- and code-based explanations. In fact, under some conditions listeners appear to misinterpret clear-speech-based lengthening and hyperarticulation as indicating that the talker intended a different vowel. This suggests that listeners (particularly non-native listeners) may not have a clear distinction between such changes. This is not to say that code- and signal-based changes are always confounded—in most circumstances these largely affect different cues. However, this particular case (and likely others) is a boundary condition that illustrates the complexity of this distinction.

Taken together, results from the current audio-visual study are in keeping with the auditory-based principles governing clear-speech production and perception, suggesting that clear-speech modifications, be they articulatory or acoustic, need to be balanced between enhancing signal saliency and preserving phonemic distinctions, with the language-specific, category-defining cues being the most effective cues to improve intelligibility.

5. Concluding remarks

The approach advocated in the current study is to carefully examine properties of the signal, both acoustic and visual, at the individual cue level to determine which specific properties define the categories that must be identified in perception, and how those properties are affected by changes in speech style. That is, we know from literature on clear-speech acoustics and articulation that modifications are non-uniform across cues, and therefore we expect perceptual uptake of acoustic/visual information to be similarly complex. By adopting a more nuanced approach to the link between characteristics of clear-speech production and cross-modal perception, we hope to achieve a better understanding of the extent to which clear-speech modifications may be beneficial for communication and why.

Acknowledgements

Portions of this study were presented at the 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan in Honolulu, Hawaii, 2016. We thank Quentin Qin, Lisa Tang, and members of the SFU Language and Brain Lab for their assistance in stimulus development and data collection. This project was supported by a research grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435-2012-1641).

Appendix A

Table A1
English perceiver models.

Dependent variable:	Overall Estimate (S.E.)	[tense] Estimate (S.E.)	[back] Estimate (S.E.)	[high] Estimate (S.E.)	[round] Estimate (S.E.)
Intercept	3.395***	3.435***	6.754***	5.640***	5.848***
	(0.25)	(0.24)	(0.56)	(0.43)	(0.44)
Mode (AO)	−1.476***	−0.856***	−3.384***	−0.461	−2.886***
	(0.16)	(0.17)	(0.49)	(0.38)	(0.36)
Mode (VO)	−2.747***	−2.693***	−2.849***	−2.893***	−2.625***
	(0.30)	(0.30)	(0.62)	(0.48)	(0.48)
Style (clear)	0.523**	0.548**	0.844	0.210	0.134
	(0.20)	(0.21)	(0.70)	(0.39)	(0.40)
Tensity (lax)	0.430	1.270***	0.867	−0.517	−0.628
	(0.29)	(0.33)	(0.71)	(0.41)	(0.42)
Mode × Style (AO, clear)	−0.015	0.335	−0.374	−0.429	0.027
	(0.24)	(0.26)	(0.72)	(0.48)	(0.43)
Mode × Style (VO, clear)	0.107	0.098	0.863	0.277	0.097
	(0.22)	(0.23)	(0.76)	(0.41)	(0.43)
Mode × Tensity (AO, lax)	−0.068	−0.240	−0.269	−0.602	0.534
	(0.23)	(0.28)	(0.64)	(0.41)	(0.37)
Mode × Tensity (VO, lax)	−0.037	0.228	−2.241**	−0.451	0.874
	(0.38)	(0.42)	(0.82)	(0.50)	(0.53)
Style × Tensity (clear, lax)	−0.243	−0.729*	−1.357	0.840	0.889
	(0.29)	(0.33)	(0.91)	(0.52)	(0.53)
Mode × Style × Tensity	0.181	0.227	1.339	−0.287	−0.701
(AO, clear, lax)	(0.34)	(0.41)	(0.95)	(0.63)	(0.57)
Mode × Style × Tensity	−0.785*	−1.097**	0.573	−0.508	−1.002
(VO, clear, lax)	(0.32)	(0.36)	(0.97)	(0.56)	(0.59)
N	13,607	13,607	13,607	13,607	13,607
LL	−4910	−3895	−1958	−2307	−2415
Random effects (σ)					
Item	1.174	1.207	1.719	1.250	1.30
Subject	0.634	0.540	0.881	1.116	1.17
Subject/Mode (AO)	0.154	0.131	0.629	0.226	0.68
Subject/Mode (VO)	0.626	0.487	0.708	1.183	1.06

*$p \le 0.05$, **$p \le 0.01$ and ***$p \le 0.001$.

Table A2
Mandarin Perceiver Models.

Dependent Variable:	Overall	[tense]	[back]	[high]	[round]
	Estimate	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	1.233***	1.359***	5.863***	3.977***	3.992***
	(0.17)	(0.17)	(0.40)	(0.28)	(0.28)
Mode (AO)	−0.528***	−0.188	−3.068***	−0.104	−1.696***
	(0.10)	(0.10)	(0.34)	(0.22)	(0.21)
Mode (VO)	−1.371***	−1.282***	−1.954***	−1.470***	−1.259***
	(0.19)	(0.19)	(0.42)	(0.28)	(0.30)
Style (clear)	0.813***	0.932***	0.603	0.181	0.085
	(0.10)	(0.10)	(0.32)	(0.19)	(0.19)
Tensity (lax)	1.377***	1.581***	0.470	0.279	0.490
	(0.21)	(0.22)	(0.39)	(0.29)	(0.31)
Mode × Style (AO, clear)	−0.147	−0.073	−0.208	−0.173	0.204
	(0.13)	(0.14)	(0.35)	(0.26)	(0.22)
Mode × Style (VO, clear)	0.115	0.044	0.292	−0.168	−0.197
	(0.13)	(0.13)	(0.39)	(0.22)	(0.22)
Mode × Tensity (AO, lax)	−0.591***	−0.579***	0.190	−0.567*	−0.094
	(0.14)	(0.15)	(0.32)	(0.26)	(0.23)
Mode × Tensity (VO, lax)	−0.183	0.340	−1.964***	−1.022**	−0.088
	(0.25)	(0.27)	(0.50)	(0.35)	(0.39)
Style × Tensity (clear, lax)	−1.469***	−1.716***	−0.295	0.122	0.210
	(0.15)	(0.16)	(0.45)	(0.28)	(0.28)
Mode × Style × Tensity	0.784***	0.709***	0.842	−0.209	−0.241
(AO, clear, lax)	(0.20)	(0.21)	(0.50)	(0.36)	(0.33)
Mode × Style × Tensity	−0.045	−0.447*	0.039	0.395	−0.045
(VO, clear, lax)	(0.19)	(0.20)	(0.51)	(0.32)	(0.34)
N	19,440	19,440	19,440	19,440	19,440
LL	−9676	−8497	−3082	−4252	−4477
Random effects (σ)					
Item	0.806	0.849	1.376	0.998	1.122
Subject	0.669	0.604	1.393	1.055	0.998
Subject/Mode (AO)	0.290	0.274	1.207	0.538	0.683
Subject/Mode (VO)	0.445	0.391	0.866	0.624	0.633
Subject/Tensity (lax)	0.543	0.591		0.542	0.544

*$p \le 0.05$, **$p \le 0.01$ and ***$p \le 0.001$.

Table B1
Means and (standard errors) of acoustic parameters in Leung et al. (2016).

		Tense						Lax					
		/i/		/ɑ/		/u/		/ɪ/		/ʌ/		/ʊ/	
		Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain
F1	Female	3.68	3.73	7.17	7.08	3.81	3.89	5.28	5.32	6.83	6.72	5.67	5.69
(Bark)		(0.04)	(0.04)	(0.05)	(0.07)	(0.04)	(0.04)	(0.04)	(0.04)	(0.07)	(0.07)	(0.05)	(0.05)
	Male	2.90	2.89	6.48	6.45	3.31	3.30	4.47	4.52	6.21	6.07	4.79	4.72
		(0.02)	(0.02)	(0.05)	(0.06)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)
F2	Female	15.05	15.03	10.12	10.33	10.29	10.56	13.58	13.51	11.85	11.87	11.76	11.95
(Bark)		(0.03)	(0.03)	(0.03)	(0.04)	(0.09)	(0.10)	(0.03)	(0.04)	(0.05)	(0.04)	(0.06)	(0.06)
	Male	14.07	14.06	8.58	8.64	9.39	9.56	12.96	12.75	10.65	10.62	10.43	10.60
		(0.05)	(0.04)	(0.04)	(0.04)	(0.12)	(0.12)	(0.07)	(0.06)	(0.06)	(0.07)	(0.09)	(0.07)
F3	Female	16.33	16.30	15.29	15.13	14.78	14.74	15.62	15.56	15.38	15.29	15.20	15.13
(Bark)		(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)
	Male	15.86	15.88	14.91	14.82	13.90	13.92	14.96	14.83	14.61	14.60	14.22	14.15
		(0.05)	(0.04)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(0.05)	(0.07)	(0.06)	(0.06)	(0.06)
Spectral	Female	0.86	1.07	2.89	2.85	1.59	1.49	1.09	1.04	3.10	2.77	2.77	2.39
Change		(0.04)	(0.04)	(0.09)	(0.08)	(0.09)	(0.07)	(0.04)	(0.04)	(0.09)	(0.09)	(0.09)	(0.08)
(Bark)	Male	0.66	0.70	2.32	2.26	1.46	1.46	1.35	1.33	2.77	2.35	2.27	1.97
		(0.03)	(0.03)	(0.10)	(0.09)	(0.08)	(0.09)	(0.05)	(0.05)	(0.12)	(0.10)	(0.13)	(0.09)
Spectral	Female	3.06	3.11	3.13	3.14	3.12	3.13	3.12	3.14	3.14	3.15	3.15	3.15
Angle		(0.02)	(0.02)	(0.00)	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.00)	(0.00)	(0.01)	(0.01)
	Male	3.07	3.07	3.14	3.14	3.14	3.13	3.14	3.13	3.15	3.14	3.14	3.16
		(0.02)	(0.03)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.00)	(0.00)	(0.02)	(0.02)
Word	Female	632	455	609	444	639	473	532	358	511	381	539	373
Duration		(20.5)	(10.0)	(20.8)	(9.1)	(22.7)	(12.4)	(24.0)	(8.6)	(18.4)	(9.8)	(22.1)	(9.6)
(ms)	Male	608	511	564	483	634	535	443	373	444	387	478	398
		(14.2)	(10.0)	(12.3)	(9.8)	(18.2)	(15.1)	(12.3)	(8.3)	(12.1)	(9.6)	(12.7)	(9.3)

(*continued on next page*)

Table B1 (*continued*)

		Tense						Lax					
		/i/		/ɑ/		/u/		/ɪ/		/ʌ/		/ʊ/	
		Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain
Vowel Duration (ms)	Female	364 (12.4)	251 (7.1)	349 (13.3)	261 (6.3)	390 (16.5)	274 (8.9)	230 (12.8)	167 (4.4)	217 (7.1)	177 (4.5)	250 (11.1)	182 (5.0)
	Male	340 (9.0)	279 (5.7)	307 (7.3)	269 (5.9)	369 (11.5)	307 (8.8)	163 (4.1)	146 (3.1)	162 (3.5)	150 (3.2)	185 (4.9)	163 (3.5)
Intensity (dB)	Female	62.3 (0.4)	62.2 (0.4)	65.6 (0.5)	64.5 (0.4)	63.8 (0.4)	63.2 (0.4)	66.4 (0.4)	64.9 (0.4)	66.7 (0.5)	65.8 (0.4)	66.4 (0.4)	65.8 (0.5)
	Male	61.1 (0.5)	59.9 (0.6)	64.0 (0.5)	62.8 (0.5)	62.1 (0.6)	60.3 (0.6)	66.0 (0.6)	65.1 (0.7)	65.1 (0.5)	64.0 (0.6)	65.7 (0.6)	64.0 (0.6)
f0 (Hz)	Female	233 (2.9)	242 (4.2)	213 (2.5)	221 (3.6)	236 (4.0)	243 (4.9)	235 (3.6)	241 (5.1)	229 (3.6)	236 (4.7)	233 (3.6)	244 (5.7)
	Male	117 (1.6)	114 (1.6)	110 (2.5)	109 (2.5)	117 (1.7)	116 (1.8)	124 (1.8)	120 (2.1)	117 (1.7)	114 (1.8)	124 (2.0)	121 (2.1)

Table B2
Means and (standard errors) of visual parameters in Tang et al. (2015).

		Tense						Lax					
		/i/		/ɑ/		/u/		/ɪ/		/ʌ/		/ʊ/	
		Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain	Clear	Plain
Vertical Lip Stretch (norm.)	Female	1.05 (0.02)	0.98 (0.01)	1.13 (0.02)	1.08 (0.02)	0.96 (0.02)	0.92 (0.02)	1.05 (0.02)	0.98 (0.02)	1.06 (0.02)	1.01 (0.02)	1.03 (0.02)	0.97 (0.02)
	Male	1.36 (0.03)	1.19 (0.02)	1.40 (0.04)	1.28 (0.03)	1.15 (0.03)	1.07 (0.02)	1.33 (0.04)	1.16 (0.02)	1.31 (0.03)	1.17 (0.02)	1.26 (0.03)	1.13 (0.03)
Horizontal Lip Stretch (norm.)	Female	0.85 (0.01)	0.84 (0.01)	0.79 (0.01)	0.80 (0.01)	0.81 (0.01)	0.81 (0.01)	0.82 (0.01)	0.81 (0.01)	0.81 (0.01)	0.81 (0.01)	0.80 (0.01)	0.78 (0.01)
	Male	0.99 (0.01)	0.93 (0.01)	0.88 (0.01)	0.89 (0.01)	0.88 (0.01)	0.89 (0.01)	0.97 (0.01)	0.94 (0.01)	0.91 (0.01)	0.90 (0.01)	0.87 (0.01)	0.88 (0.01)
Lip Rounding (norm.)	Female	0.77 (0.01)	0.75 (0.01)	0.76 (0.01)	0.74 (0.01)	0.77 (0.01)	0.78 (0.01)	0.75 (0.01)	0.74 (0.01)	0.75 (0.01)	0.75 (0.01)	0.75 (0.01)	0.74 (0.01)
	Male	0.75 (0.01)	0.75 (0.01)	0.77 (0.01)	0.77 (0.01)	0.72 (0.01)	0.75 (0.01)	0.76 (0.01)	0.77 (0.01)	0.77 (0.01)	0.77 (0.01)	0.73 (0.01)	0.76 (0.01)
Jaw Disp. (norm.)	Female	0.11 (0.01)	0.10 (0.01)	0.13 (0.01)	0.12 (0.01)	0.09 (0.01)	0.09 (0.01)	0.17 (0.06)	0.10 (0.01)	0.13 (0.01)	0.13 (0.01)	0.10 (0.01)	0.09 (0.01)
	Male	0.13 (0.01)	0.13 (0.01)	0.19 (0.01)	0.16 (0.01)	0.14 (0.01)	0.11 (0.01)	0.18 (0.01)	0.13 (0.01)	0.21 (0.01)	0.16 (0.01)	0.14 (0.01)	0.13 (0.01)
Word Duration (ms)	Female	1052 (21.4)	917 (15.6)	1036 (25.8)	904 (14.2)	1089 (27.7)	908 (18.0)	966 (25.7)	816 (14.3)	970 (20.7)	865 (16.6)	957 (21.6)	836 (13.7)
	Male	1077 (14.4)	947 (11.8)	1024 (13.9)	915 (10.1)	1095 (24.0)	959 (16.8)	916 (13.7)	818 (11.5)	912 (12.0)	823 (11.5)	932 (15.7)	832 (10.8)

References

Assmann, P. F., & Katz, W. F. (2005). Synthesis fidelity and time-varying spectral change in vowels. *Journal of the Acoustical Society of America, 117*(2), 886–895.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bradlow, A., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America, 112*(1), 272–284.

Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities. *Journal of Speech, Language, and Hearing Research, 46*(1), 80–97.

Clopper, C. G., Pisoni, D. B., & De Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118*(3), 1661–1676.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics, 38*(2), 167–184.

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *Journal of the Acoustical Society of America, 128*(4), 2059–2069.

Fenwick, S., Davis, C., Best, C. T., & Tyler, M. D. (2015). The effect of modality and speaking style on the discrimination of non-native phonological and phonetic contrasts in noise. In Proceedings of the 1st joint conference on facial analysis, animation, and auditory-visual speech processing (pp. 67–72), Vienna, Austria.

Ferguson, S. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America, 116*(4), 2365–2373.

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

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.

Ferguson, S. H., & Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *Journal of the Acoustical Society of America, 135*(6), 3570–3584.

Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology, 27*, 135–158.

Gagné, J. P., Querengesser, C., Folkeard, P., Munhall, K. G., & Masterson, V. M. (1995). Auditory, visual, and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech. *The Volta Review, 97*, 33–51.

Gagné, J. P., Rochette, A. J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Communication, 37*(3–4), 213–230.

Gopal, H. S. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations. *Journal of Phonetics, 18*(4), 497–518.

Granlund, S., Hazan, V., & Baker, R. (2012). An acoustic–phonetic comparison of the clear speaking styles of Finnish-English late bilinguals. *Journal of Phonetics, 40*(3), 509–520.

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *Journal of the Acoustical Society of America, 130*(4), 2139–2152.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52*(11–12), 996–1009.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *Journal of the Acoustical Society of America, 119*(3), 1740–1751.

Helfer, K. (1997). Auditory and auditory-visual perception of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 40*(2), 432–443.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*(5), 3099–3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105*(6), 3509–3523.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America, 106*(3), 1532–1542.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *Journal of the Acoustical Society of America, 119*(2), 1118–1130.

Jongman, A., & McMurray, B. (2017). On invariance: Acoustic input meets listener expectations. In A. Lahiri & S. Kotzor (Eds.), *The speech processing lexicon: Neurocognitive and behavioural approaches* (pp. 21–50).

Kim, J., & Davis, C. (2014). Comparing the consistency and distinctiveness of speech produced in quiet and in noise. *Computer Speech & Language, 28*(2), 598–606.

Kim, J., Sironic, A., & Davis, C. (2011). Hearing speech in noise: Seeing a loud talker is better. *Perception, 40*(7), 853–862.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *Journal of the Acoustical Society of America, 117*(4), 2238–2246.

Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America, 112*(5), 2165–2172.

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research, 55*(6), 1807–1821.

Lander, K., & Capek, C. (2013). Investigating the impact of lip visibility and talking style on speechreading performance. *Speech Communication, 55*(5), 600–605.

Leung, K. K., Jongman, A., Wang, Y., & Sereno, J. A. (2016). Acoustic characteristics of clearly spoken English tense and lax vowels. *Journal of the Acoustical Society of America, 140*(1), 45–58.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Springer.

Liu, S., Del Rio, E., Bradlow, A. R., & Zeng, F. G. (2004). Clear speech perception in acoustic and electric hearing. *Journal of the Acoustical Society of America, 116*(4), 2374–2383.

Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *Journal of the Acoustical Society of America, 124*(5), 3261–3275.

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America, 125*(6), 3962–3973.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246.

McMurray, B., & Jongman, A. (2015). What comes after [f]? Prediction in speech is a product of expectation and signal. *Psychological Science, 27*(1), 43–52.

Moon, S. J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America, 96*(1), 40–55.

Ohala, J. (1995). Clear speech does not exaggerate phonemic contrast. In *Proceedings of the 4th European conference on speech communication and technology* (pp. 1323–1325).

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. *Journal of the Acoustical Society of America, 95*(3), 1581–1592.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 28*(1), 96–103.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America, 103*(6), 3677–3689.

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *Journal of the Acoustical Society of America, 118*(3), 1677–1688.

Smiljanić, R., & Bradlow, A. R. (2008). Stability of temporal contrasts across speaking styles in English and Croatian. *Journal of Phonetics, 36*(1), 91–113.

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.

Smiljanić, R., & Bradlow, A. R. (2011). Bidirectional clear speech perception benefit for native and high-proficiency non-native talkers and listeners: Intelligibility and accentedness. *Journal of the Acoustical Society of America, 130*(6), 4020–4031.

Tang, L. Y., Hannah, B., Jongman, A., Sereno, J., Wang, Y., & Hamarneh, G. (2015). Examining visible articulatory features in clear and plain speech. *Speech Communication, 75*, 1–13.

Tasko, S. M., & Greilick, K. (2010). Acoustic and articulatory features of diphthong production: A speech clarity study. *Journal of Speech, Language, and Hearing Research, 53*(1), 84–99.

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics, 35*(2), 244–258.

Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech, Language, and Hearing Research, 39*(3), 494–509.

Van Engen, K. J., Phelps, J. E., 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.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *Journal of the Acoustical Society of America, 124*(3), 1716–1726.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*(4), 539–552.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555–568.

Tagliaferri, B. (2005). *Paradigm: Perception Research Systems, Inc.* [Software]. Available online at: http://www.perceptionresearchsystems.com.