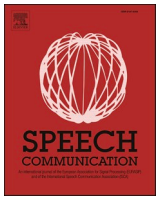
Speech Communication 171 (2025) 103243

Contents lists available at ScienceDirect

Speech Communication

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



Expectation of speech style improves audio-visual perception of English vowels

Joan A. Sereno [a], Allard Jongman [a,*], Yue Wang [b], Paul Tupper [c], Dawn M. Behne [d], Jetic Gu [b], Haoyao Ruan [b]

[a] Department of Linguistics, University of Kansas, Lawrence, KS 66045, USA
[b] Department of Linguistics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada
[c] Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada
[d] Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Keywords:
Expectations
Speech style
Vowel perception
Audio-visual modality

ABSTRACT

Speech perception is influenced by both signal-internal properties and signal-independent knowledge, including communicative expectations. This study investigates how these two factors interact, focusing on the role of speech style expectations. Specifically, we examine how prior knowledge about speech style (clear versus plain speech) affects word identification and speech style judgment. Native English perceivers were presented with English words containing tense versus lax vowels in either clear or plain speech, with trial conditions manipulating whether style prompts (presented immediately prior to the target word) were congruent or incongruent with the actual speech style. The stimuli were also presented in three input modalities: auditory (speaker voice), visual (speaker face), and audio-visual. Results show that prior knowledge of speech style improved accuracy in identifying style after the session when style information in the prompt and target word was consistent, particularly in auditory and audio-visual modalities. Additionally, as expected, clear speech enhanced word intelligibility compared to plain speech, with benefits more evident for tense vowels and in auditory and audio-visual contexts. These results demonstrate that congruent style prompts improve style identification accuracy by aligning with high-level expectations, while clear speech enhances word identification accuracy due to signal-internal modifications. Overall, the current findings suggest an interplay of processing sources of information which are both signal-driven and signal-independent, and that high-level signal-complementary information such as speech style is not separate from, but is embodied in, the signal.

1. Introduction

1.1. Background

Speech perception is a complex process involving managing both signal-driven and signal-independent information (Lindblom, 1990). That is, the processing of incoming speech stimuli is not only driven by the signal itself but is also modulated by the perceiver's knowledge about a language (e.g., lexical sound structure or word frequency, Luce, 1986) and communicative context (e.g., listening environment or talker characteristics, McMurray and Jongman, 2011). The latter contribution implies a high-level signal-complementary process involving accessing both long-term internalized knowledge through prior experience as well as short-term adaptations through exposure to a spontaneous and variable speech context (Kleinschmidt and Jaeger, 2015; Lindblom, 1990). Speech perception theories have recently and consistently acknowledged that high-level knowledge interacts with the incoming speech signal to account for physical variations in the signal, and that perception involves constant comparisons between signal input and knowledge-based expectations (Fowler and Smith, 1986; Kleinschmidt and Jaeger, 2015; Jongman and McMurray, 2017; Lindblom, 1990; McMurray and Jongman, 2011).

Empirical evidence reveals that various types of expectations outside of the speech signal may alter perception (Apfelbaum et al., 2014; McMurray and Jongman, 2015; Niedzielski, 1999). First, studies have demonstrated that knowledge about the talker (e.g., gender, identity) enables experience-based expectations that facilitate the perception of speech segments such as fricatives and vowels (Johnson, 1999;

* Corresponding author.
E-mail address: jongman@ku.edu (A. Jongman).

https://doi.org/10.1016/j.specom.2025.103243
Received 7 September 2024; Received in revised form 7 April 2025; Accepted 11 April 2025
Available online 17 April 2025

McMurray and Jongman, 2015; Strand, 1999). For example, McMurray and Jongman (2015) showed that if perceivers know the upcoming fricative and who produced it, this prior experience enables them to account for variance and use the residual information to make more accurate predictions about the following vowel. Likewise, seeing a talker's face helps perceivers form a more complete foundation for perception, including acoustic cues for vocal tract configurations, visual cues, and talker-idiosyncratic characteristics (Johnson, 1999). As shown in Strand (1999), the presence of a female face shifts the perceived category boundary between /ʃ/ and /s/ to a higher frequency while a male face shifts the boundary down to a lower frequency. Moreover, knowledge about the language such as the talker's dialect also allows the perceiver to extract and use relevant linguistic information. For example, Niedzielski (1999) showed that Detroit listeners matched a vowel target with a raised variant when they thought they were listening to a Canadian speaker but not when they thought they were listening to a Detroiter. Thus, expectations about where a speaker was from affected speech perception.

One type of expectation that has not been explored is speech style, such as clear speech, a type of hyper-articulation intended to improve intelligibility. In addition to the factors discussed above (e.g., gender, talker, dialect), clear speech provides another good vehicle for testing the role of expectations. On the one hand, high-level information about speaking styles may be predicted outside the speech signal or in a signal-universal manner; on the other hand, clear speech cues are also integrated in the signal in a segment-specific manner (Redmon et al., 2020; Smiljanić, 2021).

Indeed, clear speech (compared to plain, conversational speech) arises from two levels of modifications: (1) signal-based, involving signal-universal changes (e.g., higher pitch or intensity overall) to enhance general acoustic saliency, and (2) code-based, involving segment-specific changes to increase sound category distinctions (e.g., altering formants to enlarge acoustic distance between two vowels) (Bradlow and Bent, 2002; Leung et al., 2016; Redmon et al., 2020; Zhao and Jurafsky, 2009). Previous studies have shown that globally enhanced acoustic and articulatory information in clear speech, such as increased duration, decreased articulation rate, and greater intensity, can improve speech intelligibility in both auditory and visual perception (e.g., Ferguson and Kewley-Port, 2002; Kim et al., 2011; Lam et al., 2012; Maniwa et al., 2008; Redmon et al., 2020). However, code-based clear speech cues that are aligned with speech-intrinsic properties appear to be more effective than signal-based cues in aiding auditory and visual intelligibility (Ferguson and Kewley-Port, 2002; Maniwa et al., 2008; Redmon et al., 2020; Smiljanić, 2021; Zeng et al., 2023). Moreover, while code-based category-enhancing cues (e.g., more dynamic formant patterns for lax vowels) are found to increase intelligibility, code-based category-blurring cues (e.g., lengthened lax vowels) may decrease intelligibility (Redmon et al., 2020; Zeng et al., 2023). Aside from signal-driven cues, perceivers are also influenced by additional signal-independent situations, such as effects of talker gender, background noise, and semantic context (Bradlow and Bent, 2002; Smiljanić and Sladen, 2013; Van der Feest et al., 2019; Van Engen, et al., 2014). These findings suggest combined clear-speech effects in signal-dependent processing as well as in higher-level processing abstracted from the input speech (Smiljanić, 2023).

Consistently, research based on game theoretic models (Bens et al., 2006; Jäger, 2008) also predicts that knowledge of speech styles can yield expectations to inform perceivers about sources of signal and contextual variability and enhance accuracy of perception (Tupper et al., 2018). However, such predictions have not been empirically tested. The critical underlying question is how high-level knowledge about a speech signal interacts with the cues that are present in the signal, or whether there exist different pathways for signal-independent and signal-driven clear speech cues.

1.2. The present study

We address these questions by investigating how knowledge of, and exposure to, clear speech affect perceivers' word identification and speech style judgment. For word identification, native English perceivers are presented with English words differing in vowels (tense versus lax) produced in either clear speech or plain speech, in each of the three input modalities: Auditory (AO), Visual (VO), and Audio-visual (AV).

The word and style contrasts employed in the current experiment followed our previous study (Redmon et al., 2020), which revealed a clear-speech benefit for tense vowels across A/V modalities as well as for lax vowels in AO and AV modalities. However, a clear-speech disadvantage was observed for lax vowels in the VO modality, arguably due to a similar extent of visual articulatory clear-speech enhancement and lengthening for both tense and lax vowels (Tang et al., 2015), resulting in clear lax vowels being perceived as tense vowels. Findings from this study have laid the foundation of the current experiment to explore clear-speech effects as a function of expectation and exposure.

Crucially, in the current design, prior to each target word, perceivers are prompted with a hint indicating whether an upcoming token is in clear speech (i.e., screen displaying "clearly spoken" in text). In one session ("congruent-only"), the prompt always correctly indicates the speech style (e.g., "clearly spoken" prompt, clearly produced target word). In another session ("mixed"), the prompt either matches the target in style (e.g., "clearly spoken" prompt, clearly produced target word) or mismatches it (e.g., "clearly spoken" prompt, plainly produced target word, or vice versa) (See Fig. 1). A style identification test is then administered after the word identification session in which perceivers are presented with the same words and are asked to indicate the style (clear or plain) of each word.

This design allows us to examine the extent to which perception is

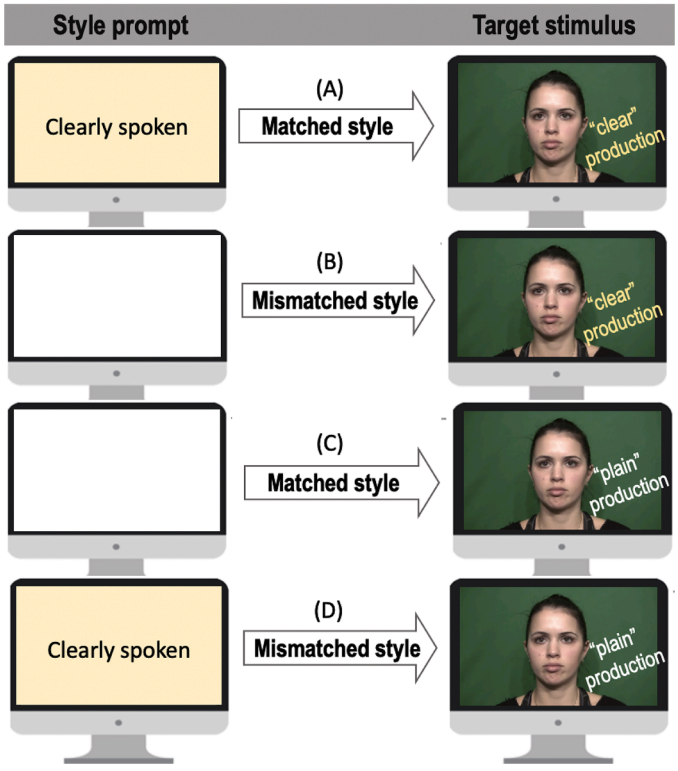


Fig. 1. Four different types of prompt-target matched or mismatched trials in a "mixed" experimental session: (A) matched style (clear prompt, clear target), (B) mismatched style (no clear prompt, clear target), (C) matched style (no clear prompt, plain target), and (D) mismatched style (clear prompt, plain target).

dependent on signal-internal clear-speech cues and the extent to which it is influenced by expectations of an incoming speech style.

As a starting point, we should expect signal-internal clear-speech effects as revealed in previous research (Redmon et al., 2020). That is, clear (relative to plain) speech should present an advantage across vowels and A/V modalities, except for a disadvantage with lax vowels in VO (where clear-speech modifications conflict with visual cues for lax vowels).

Regarding the main goal of comparing signal-internal versus expectation-based clear-speech effects, we predict the following. First, the manipulation of the trial-by-trial "style" prompt enables unraveling these differences: from the "mixed" session, greater word identification accuracy in matched than mismatched prompt-target trials would suggest effects of knowledge-based expectations, while greater accuracy in clear as compared to plain trials, regardless of matching conditions, would favor signal-dependent perception. Additionally, comparisons of prompt-target matched stimuli in congruent-only against mixed sessions would reveal effects of expectations arising from short-term enhancement and exposure to clear speech against those based on inconsistent contextual information which does not match the presented signal. Furthermore, the style identification task allows examining whether participants are sensitive to the longer-term effect of expectations based on high-level, signal-independent knowledge of speech styles. Finally, the above effects may interact with input modality (AO, VO, AV) and vowel characteristics (tense-lax contrast), presumably due to different weighting of signal saliency in the auditory versus visual domains and/or with tense versus lax vowels. For example, prior knowledge may not benefit visual perception of clear lax vowels due to the conflicting articulatory clear-speech modifications of lax vowels mentioned earlier (Redmon et al., 2020).

Specifically, for the mixed session, we predict the following (See Table 1.1): Word (vowel) identification would be more accurate in clear relative to plain speech styles, reflecting a signal-based advantage; and more accurate with a congruent than incongruent prompt, reflecting effects of knowledge-based expectations. Thus: (1) identification would be most accurate in clear speech with a matched prompt (Condition A), benefiting from both signal- and knowledge-based information; (2) identification would be least accurate in plain speech with a mismatched prompt, with no signal-internal or -external benefit (Condition D); and (3) for Conditions B and C, we have two alternative predictions: if the effect of "signal" outweighed that of "expectation", we would expect greater accuracy with a mismatched "clear" target word (Condition B) than a matched "plain" target word (Condition C); however, if the effect of "expectation" outweighed that of "signal", we would expect the reverse.

2. Methods

2.1. Perceivers

A total of 126 native North American English perceivers (aged 21–59 yrs, mean: 35 yrs) recruited via Amazon Mechanical Turk who completed the task were included in the current study. The perceivers reported having English as their native and dominant language, and having normal hearing and vision, and no history of speech, language, or neurological disorders. The participants were randomly assigned to one of the six sessions of the experiment (20–22 per session, See 2.3 below). The participants provided online informed consent and were compensated for their participation.

Table 1.1
Predictions of perceptual accuracy ranking in the four prompt-target conditions.

	Trial condition			Predicted accuracy ranking (1: most accurate; 4: least accurate)	
	prompt (perceiver knowledge)	signal style	prompt-signal congruency	favors signal	favors expectation
A	Clear	clear	match	1	1
B	Plain	clear	mismatch	**2**	**3**
C	Plain	plain	match	**3**	**2**
D	Clear	plain	mismatch	4	4

2.2. Stimuli

The stimuli included six English words ("keyed, kid, cod, cud, cooed" and "could") produced in plain and clear speech styles. These words contain three pairs of American English tense and lax vowels: /i-ɪ/, /ɑ-ʌ/ and /u-ʊ/ (Gopal, 1990; Lam et al., 2012; Leung et al., 2016).

2.2.1. Talkers

The audio-visual stimuli were provided by native talkers of Western Canadian English ($n = 4$; 2 M, 2F; aged 17–30 yrs, mean: 22 yrs). They were recruited from the undergraduate and graduate student population at Simon Fraser University (SFU). They reported no history of speech or language impairment.

These four talkers were selected from a pool of eighteen native Western Canadian talkers as our previous analysis (Tang et a., 2015) revealed that their productions contained the most contrastive visible articulatory features in plain versus clear speech. Additionally, these takers' plain and clear productions differ acoustically, with increased plain-to-clear spectral and/or temporal change (Leung et al., 2016) (Appendix 1). Moreover, based on our previous intelligibility study (Redmon et al., 2020), these speakers' productions of the target stimuli in clear (relative to plain) speech improved intelligibility.

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 (Maniwa et al., 2009; Redmon et al., 2020). On each trial, one of the six target words was displayed on a computer screen. The talker was instructed to produce the word naturally, as if in casual conversation (thus eliciting plain style productions). Then, the program would "guess" and indicate on the screen what word the talker produced. If the guess was correct (20 % of the occurrences), the program would move on to the next stimulus. If it was incorrect (80 % of the occurrences), the program would instruct the talker to repeat the stimulus as clearly as possible for any incorrect guess (thus eliciting clear style productions).

Audio-video recordings were acquired in a sound-attenuated booth in the Language and Brain Lab at SFU. Front-view videos were captured with a Canon Vixia HF30 camera at a recording rate of 30 fps. Audio recordings were acquired 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, 20 cm away from the talker's mouth. The recorded words were judged as correct productions of the target words by two native English speakers.

2.2.3. Editing of stimuli

Three sets of stimuli were created for the three presentation modalities: audio only (AO), visual only (VO), and audio-visual (AV). The AO stimuli were excised from the microphone audio recordings as individual word clips of two seconds each, using Audacity v.2.1. The AV stimuli were created by replacing the on-camera audio track with the high-quality audio recordings from the microphone through synchronization of the two waveforms, 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 induce sufficient errors for comparisons between plain and clear speech in the AO and AV conditions, the audio stimuli were first normalized at 60 dB and then embedded in cafeteria noise (primarily containing natural background conversations by multiple talkers without any discernible

words), recorded at a cafeteria at SFU. The signal-to-noise ratio (SNR) was set at a −15 dB. This SNR level was empirically established by a pilot study, with the target error rate set at 30 %, based on similar previous studies (Gagné et al., 2002; Wang et al., 2008) (See Redmon et al., 2020 for details of the pilot study.)

2.3. Procedures

The experiment was created using a custom version of jsPsych-6.1.0 and put on a JATOS server. Six experimental sessions were developed based on stimulus input modality and the composition of stimuli: AO-mixed, VO-mixed, AV-mixed, AO-congruent, VO-congruent, and AV-congruent.

The core design of this experiment was the inclusion of a "style prompt" prior to the presentation of a stimulus token (target word), indicating that the word to be perceived was "clearly spoken". The motivation for showing a prompt for the clear speech trials only, and not the plain speech, is to highlight their difference.

2.3.1. Mixed sessions

In a "mixed" session, the "clear" style prompt (a 2-second yellow screen with the text "clearly spoken") was presented either before a "clear" speech token, creating a prompt-target "matched" trial; or before a "plain" speech token, creating a prompt-target "mismatched" trial. Additionally, the session also contained plain and clear stimuli without a prompt (a blank screen preceding the target token), indicating that a "plain" token was to be presented. Such trials could also be "matched" (when the actual token was "plain") or "mismatched" (when the actual token was "clear"). Fig. 1 illustrates the composition of the four types of "matched" and "mismatched" trials.

Thus, the prompt provided the perceiver with either correct or incorrect information about the speech style of the target word before the AO, VO, or AV file of the word was presented. The perceiver was then asked to indicate (within 3 s) which word they perceived by clicking on one of the six target words on the response screen.

Each of the plain and clear stimuli in the "mixed" session was paired with each style prompt once. In total, a "mixed" session contained 96 trials, including two prompt-target style matching conditions (matched, mismatched) x 2 styles (clear, plain) x 6 target words x 4 talkers. Presentation of the stimuli was randomized, split into four blocks of 24 trials each with short breaks in-between. There were three "mixed" sessions, one for each input modality (AO, VO, and AV).

2.3.2. Congruent-only sessions

In addition to the "mixed" session, three "congruent-only" sessions were also developed, one for each modality (AO, VO, and AV). The main difference between a "congruent-only" and a "mixed" session was that the former did not contain any mismatched prompt-target trials. All clear tokens were prompted with the yellow "clearly spoken" screen, and all plain tokens were preceded with a blank screen. A "congruent-only" session contained a total of 48 trials (2 styles x 6 words x 4 talkers). Presentation of the stimuli was randomized, split into four blocks of 12 trials each with short breaks in-between.

2.3.3. Style identification

After both the "mixed" and "congruent-only" sessions, a "style test" was administered, in which participants were to identify the "speech style" of the target word. The stimuli involved the same six target words in plain and clear speech by the same four talkers, presented in the AO, VO, or AV modality after the corresponding main session. Thus, there were three versions of the style test: AO-style, VO-style, or AV-style, administered after either the mixed or congruent-only main sessions. The audio stimuli were not embedded in noise, since the task was to identify the speech style rather than the target word. For a style test trial, the text of the target word was first shown for 2 s, followed by presentation of the target stimulus, which was in turn followed by a response screen (up to 3 s) showing the options of "Plain" or "Clear". The style test contained 48 trials (2 styles x 6 words x 4 talkers). The presentation of stimuli was randomized within each version of the test.

2.3.4. Summary of sessions and tasks

To summarize (See Table 2.1), the experiment involved six sessions for six different groups of participants: AO-mixed, VO-mixed, AV-mixed, AO-congruent, VO-congruent, and AV-congruent. The mixed sessions involved a clear-speech prompt which either matched or did not match the speech style of the target word, while only the matched trials were included in the congruent-only sessions. Each session contained a main test involving word identification given the style prompt, followed by a style test for speech style identification.

2.3.5. Online experiment setup

All sessions were conducted online. Participants were asked to take the tests in a quiet room, using a screen size of 13-inches or more. Those who were assigned to the AO- and AV- sessions were required to wear a wired headphone. Each session started with a language background questionnaire, followed by instructions to help participants calibrate their computer browser size and audio-video settings. A short practice session was also included to familiarize participants with the target words, speech styles, and tasks, using stimuli and talkers not included in the testing sessions.

3. Results

Effects of perceiver experience of speech style were analyzed through three sets of comparisons. Analysis 1 examined if participants could better identify a word after being provided with correct (as opposed to incorrect) information about the speech style of the word, involving within-subjects comparisons between matched and mismatched prompt-target trials within the "mixed" session. Analysis 2 evaluated if word identification would benefit from knowledge of the correct speech style information (than signal-intrinsic enhancements) through between-subjects comparisons of the matched prompt-target trials in the "congruent-only" session versus those in the "mixed" session". Analysis 3 involved between-subjects comparisons from the style test, focusing on whether identification of the speech style of a word improved after completing the "congruent-only" session as compared to the "mixed" session. These three analyses were separate for each input modality (AO, VO, AV), a between-subjects factor. Thus, a total of nine datasets were used (3 types of analysis x 3 modalities) .[1]

Each of the nine datasets was submitted to a logistic mixed-effects model using the 'lme4' package in R. For Analysis 1, the fixed effects included Matching (matched vs mismatched prompt-target trials), Style (clear vs plain stimuli), and Tensity (tense- vs lax-vowel words). For Analysis 2, fixed effects included Session (congruent-only vs mixed), Style, and Tensity. For both analyses, the dependent variable was word identification accuracy. The fixed effects for Analysis 3 involved Post-session (post-congruent vs post-mixed) and Style, the dependent variable being speech style identification accuracy. Table 3.1 displays an overview of the analyses.

For each analysis, random effects were added to the intercept term to account for different participants, talkers and words. After the model was finalized, a Type III Wald chi-square test was applied (using the Anova() function in the 'car' package) to assess the fixed effects including all the possible interaction terms. For significant interactions,

[1] For each session and each response metric (word identification accuracy or style identification accuracy), data screening was performed using the 1.5 interquartile range (IQR) rule, where data points above [75th percentile +1.5*interquartile] or below [25th percentile −1.5*interquartile] were considered outliers (Upton and Cook, 1996). Only two participants in the AO-mixed session were identified as outliers and removed from all the analyses.

Table 2.1
Overview of the experiment.

	Mixed sessions style matched & mismatched prompt-target trials			Congruent sessions style matched prompt-target trials		
Participant group	1 (n = 22)	2 (n = 22)	3 (n = 20)	4 (n = 20)	5 (n = 20)	6 (n = 20)
Main test word identification	AO-mixed	VO-mixed	AV-mixed	AO-congruent	VO-congruent	AV-congruent
Style test style identification	AO-style	VO-style	AV-style	AO-style	VO-style	AV-style

Table 3.1
Analysis overview.

		AO, VO, AV	
		Measure	Fixed effects
Analysis 1	matched vs mismatched trials in mixed session	Word ID	Matching x Style x Tensity
Analysis 2	matched trials in mixed session vs congruent-only session	Word ID	Session x Style x Tensity
Analysis 3	post-congruent-only session vs post-mixed session	Style ID	Post-session x Style

Note: Random effects for each analysis include Participant, Talker, and Word.

subsequent post-hoc pairwise comparisons were conducted using the multivariate adjustment method ('mvt') in the 'emmeans' package.

Table 3.2 summarizes the overall descriptive results. For all three sets of analyses, the mean perception accuracy values revealed a trend in favor of congruent conditions across input modalities and styles of the stimuli. Detailed results from mixed-effects modeling are reported in separate sections below.

3.1. Analysis 1: matched vs mismatched prompt-target trials

Analysis 1 examined the effects of prompt-target matching on word identification and its interactions with the factors of speech style and vowel tensity of the target word through three sets of mixed-effects logistic regression analyses, one for each modality (AO, VO and AV). The generic model formula was Word accuracy ~ Matching*Style*Tensity + (1|Participant) + (1|Talker) + (1|Word). Model coefficient estimates are listed in Appendix 2 (for all three sets of analyses). Fig. 2 displays the comparisons of identification accuracy in these conditions.[2]

Modeling results for AO showed no significant main effects for Matching, with comparable mean accuracy for the matched (70 %) and mismatched (69 %) conditions. However, a significant main effect of Style was observed [$\chi^2_{(1)} = 5.89, p = 0.015$], showing higher accuracy for clear (73 %) than plain (66 %) speech.

In VO, no significant main effect was observed between matched (51.1 %) and mismatched (50.5 %) trials. For the main effect of Style, accuracy was significantly higher in clear (54.5 %) than in plain (47.1 %) speech [$\chi^2_{(1)} = 18.29, p < 0.001$]. Additionally, a significant interaction of Style x Tensity was found [$\chi^2_{(1)} = 14.46, p < 0.001$]. Post-hoc pairwise comparisons reveal that, for the tense-vowel words, accuracy was higher in the clear (61.7 %) than plain (44.1 %) style [Clear/Plain = 2.236, CI = (1.613, 3.10), $z = 6.030, p < 0.001$].

In AV, no significant main effect of Matching was observed, despite the higher mean accuracy for matched (81.5 %) than mismatched (78.2 %) trials. For the main effect of Style, accuracy in clear speech (82.9 %) was significantly higher than in plain speech (76.8 %) [$\chi^2_{(1)} = 9.69, p = 0.019$].

Table 3.2
Summary of the descriptive results: mean accuracy (& standard error) in %.

		AO	VO	AV
Analysis 1	Matched trial	70.1 (1.4)	51.1 (1.6)	81.5 (1.3)
	Mismatched trial	69.1 (1.4)	50.5 (1.5)	78.2 (1.3)
Analysis 2	Congruent session	73.8 (1.4)	53.0 (1.6)	88.2 (1.1)
	Mixed session	70.1 (1.4)	51.1 (1.6)	81.5 (1.3)
Analysis 3	Post-congruent session	75.6 (1.4)	68.6 (1.5)	77.7 (1.4)
	Post-mixed session	72.9 (1.4)	66.6 (1.5)	66.1 (1.5)

[2] For brevity, statistical results are reported only for significant and marginally significant (trends) main effects and interactions in this study.

3.2. Analysis 2: congruent-only vs mixed sessions

Analysis 2 examined whether word intelligibility benefited more from the correct style prompt in the session containing only the matched prompt-target information compared to the mixed-congruency session, where both correct and incorrect style information was provided. A mixed-effects logistic regression model was built for each of the three input modalities, using the equation Word accuracy ~ Session*Style*Tensity + (1|Participant) + (1|Talker) + (1|Word). Accuracy comparisons in these conditions are shown in Fig. 3.

The AO modality exhibited a marginally significant main effect of Session [$\chi^2_{(1)} = 3.23, p = 0.072$], showing the expected direction of higher accuracy in the congruent-only session (73.8 %) than in the mixed session (70.1 %). Moreover, the model revealed a positive clear-speech effect [$\chi^2_{(1)} = 3.85, p = 0.050$], with clear style (74.7 %) outperforming plain (69.0 %) style.

For VO, no significant main effect of Session was observed, despite the higher mean accuracy in the congruent-only session (53 %) relative to the mixed session (51.1 %). For Style, accuracy was significantly higher in clear speech (54.5 %) than in plain speech (49.5 %) [$\chi^2_{(1)} = 16.17, p < 0.001$]. Further, a significant interaction of Style x Tensity was found [$\chi^2_{(1)} = 21.99, p < 0.001$]. Post-hoc analyses showed that, for tense-vowel words, accuracy was higher in clear (64.1 %) than plain (46.4 %) speech [Clear/Plain = 2.225, CI = 1.591, 3.113), $z = 5.830, p < 0.001$], whereas for lax-vowel words, accuracy was lower for clear (44.8 %) than plain (52.7 %) speech [Clear/Plain = 0.6703, CI = (0.507, 0.975), $z = -2.640, p = 0.030$].

The AV results showed a marginally significant main effect of Session [$\chi^2_{(1)} = 3.01, p = 0.083$], with higher accuracy in the congruent-only (88.2 %) than the mixed (81.5 %) session. For Style, accuracy was expectedly higher in clear speech (88.2 %) than in plain speech (81.5 %) [$\chi^2_{(1)} = 5.26, p = 0.022$].

3.3. Analysis 3: speech style identification after congruent vs mixed sessions

Analysis 3 tested the ability to identify the speech style of the target words following the "congruent-only" versus "mixed" session (i.e., post-congruent session and post-mixed session, respectively, for the main effect of "Post-session"). For each modality, a mixed-effects logistic regression model was formulated as Style accuracy ~ Post-session*Style + (1|Participant) + (1|Talker) +(1|Word). Style accuracy results are illustrated in Fig. 4.

The AO results revealed a reliably higher style identification accuracy in the post-congruent test (75.6 %) than in the post-mixed test (72.9

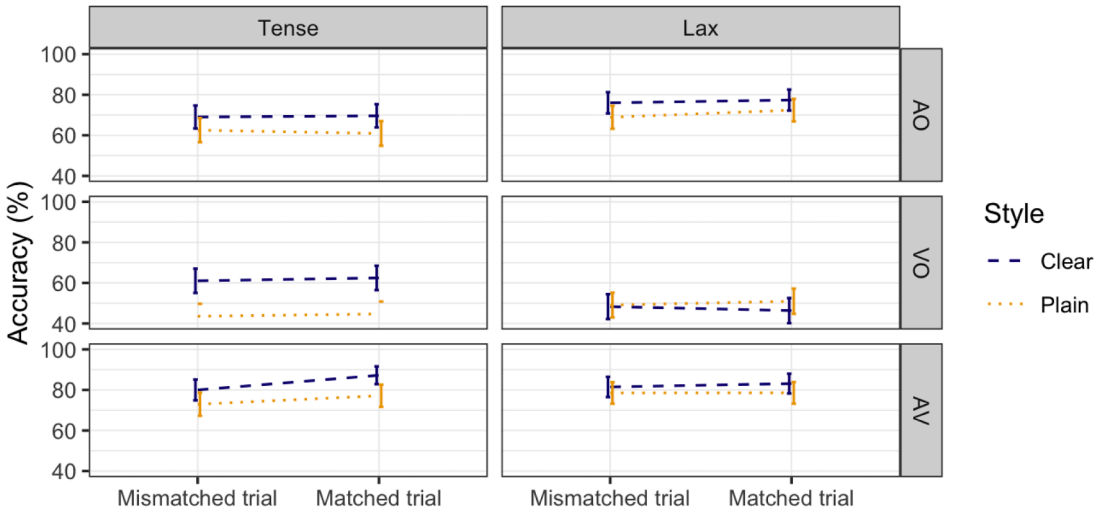


Fig. 2. Analysis 1: Mean word identification accuracy (%) comparisons between (prompt-target) style matched and mismatched trials in the "mixed" session for each speech style (clear, plain), stimulus vowel tensity (tense, lax), and input modality (AO, VO, AV). Error bars indicate 95 % confidence interval.

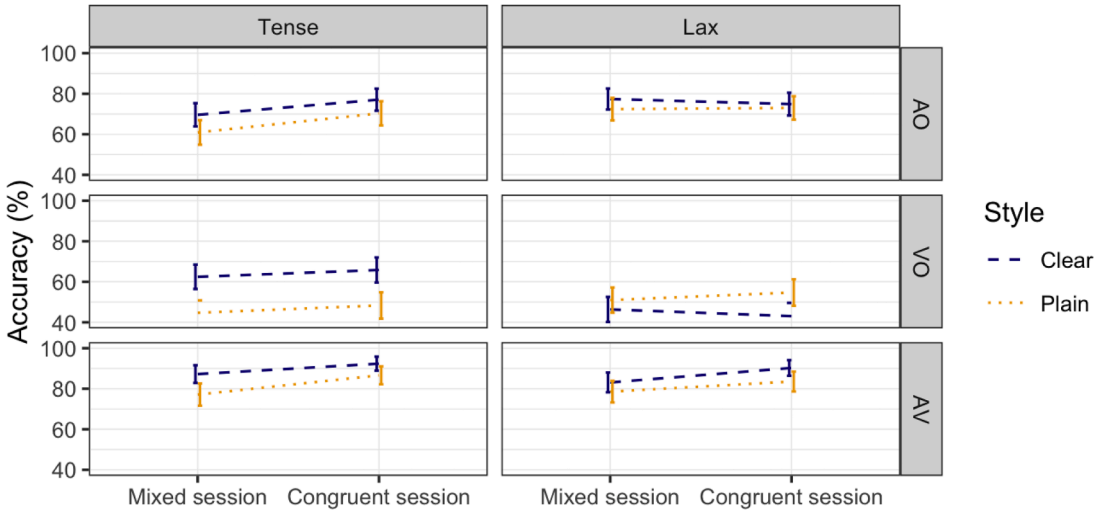


Fig. 3. Analysis 2: Mean word identification accuracy (%) comparisons of the prompt-target matched stimuli in the "congruent-only" session and the "mixed" session for each speech style (clear, plain), stimulus vowel tensity (tense, lax), and input modality (AO, VO, AV). Error bars indicate 95 % confidence interval.

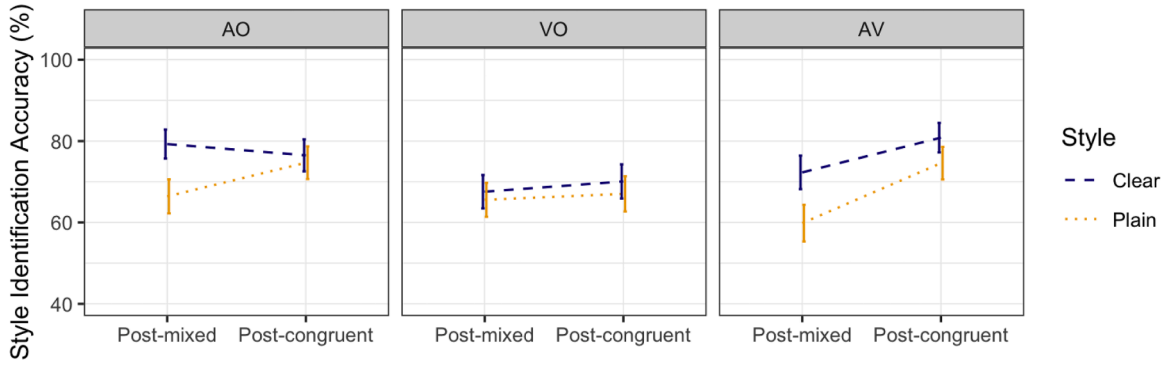


Fig. 4. Analysis 3: Mean style identification accuracy (%) comparisons in the post-congruent versus post-mixed session for each speech style (clear, plain) and input modality (AO, VO, AV). Error bars indicate 95 % confidence interval.

%) [$\chi^2_{(1)} = 5.53$, $p = 0.019$], as predicted. However, there was no significant difference in style identification for clear (78.0 %) and plain (70.4 %) stimuli, despite the mean data showing a higher speech style accuracy for clear stimuli. A significant interaction of Post-session x Style was found [$\chi^2_{(1)} = 4.85$, $p = 0.028$]. Post-hoc analysis showed that, for the post-mixed session, style identification accuracy was higher

in the clear (79.3 %) than plain (66.4 %) speech [clear/plain = 2.014, CI=(1.404, 2.89), $Z = 4.760$, $P < 0.001$]. Additionally, for the plain stimuli, style identification accuracy was marginally higher in the post-congruent test (74.7 %) than in the post-mixed test (66.4 %) [post-mixed/post-congruent = 0.67, CI = (0.436, 1.02), $z = -2.350$, $p = 0.066$].

In VO, no significant difference in style identification accuracy was detected between post-congruent (68.6 %) and post-mixed (66.6 %) tests, or between clear (68.8 %) and plain (66.3 %) stimuli.

For AV, style identification accuracy was reliably higher in the post-congruent test (77.7 %) than in the post-mixed test (66.1 %) [$\chi^2_{(1)} = 9.63$, $p = 0.002$]. For the main effect of Style, accuracy was significantly higher in clear (76.6 %) than plain speech (67.2 %) [$\chi^2_{(1)} = 5.81$, $p = 0.016$].

3.4. Summary of results

Statistically significant main effects and interactions are summarized in Table 3.3.

Overall, prior knowledge of the speech style tended to be helpful in conditions where consistent correct style information is provided. Specifically, in the AO and AV modalities, participants could more accurately identify the speech style of the target words after completing the congruent-only session compared to the mixed session (Analysis 3). There was also a trend of better performance in word identification in the congruent-only session than in the mixed session (Analysis 2). However, congruency did not significantly benefit VO perception. Moreover, across input modalities, perception accuracy did not reliably favor the prompt-target matched stimuli within the mixed session (Analysis 1).

In contrast, a robust effect of speech style on word intelligibility was consistently observed. Across input modalities, word identification results exhibited a significantly higher accuracy in clear speech than in plain speech. Moreover, the interaction between speech style and vowel tensity in VO revealed that clear speech only benefited the perception of tense-vowel words (Analyses 1 & 2). Consistently, the style identification results (Analysis 3) showed more accurate recognition of speech style when the stimuli were in clear (relative to plain) speech, especially in AO and AV modalities.

Table 3.3
Summary of statistically significant main effects and interactions.

		AO	VO	AV
Analysis 1	Style	clear > plain	clear > plain	clear > plain
	Style x Tensity		Tense: clear > plain	
Analysis 2	Session	.congruent > mixed		. congruent > mixed
	Style	clear > plain	clear > plain	clear > plain
	Style x Tensity		Tense: clear > plain Lax: plain > clear	
Analysis 3	Post-session Style	post-congruent >post-mixed		post-congruent > post-mixed clear > plain
	Post-session x Style	plain: . post-congruent >post-mixed		

Note: ">": greater accuracy; ".": marginally significant ($0.05 < p < 0.10$). All other effects in this table are significant ($p < 0.05$).

4. Discussion and conclusion

We examined how expectations of, and exposure to, clear speech affect perceivers' word identification and speech style judgment. This study was motivated by the theoretical claims that speech perception involves processing both signal-driven and signal-independent information (Fowler and Smith, 1986; Kleinschmidt and Jaeger, 2015; Lindblom, 1990; McMurray and Jongman, 2011; Tupper et al., 2018). On the one hand, processing of the incoming speech signal is driven by the signal itself, with the source of information being the signal. At the same time, processing of the incoming speech signal is driven by signal-independent sources of information, such as information modulated by the perceiver's knowledge about a language, including contextual information, exposure to language, listening environment, expectations, and communicative setting. The question in essence is how such high-level signal-independent cues work in tandem with (or separately from) signal-inherent cues to improve speech intelligibility.

The present study addressed this question from two novel perspectives. First, we consider speech style (clear speech), a novel test case. Previous studies have established that knowledge of information, such as talker identity and gender, forms expectations about the incoming signal and aids perception (Johnson, 1999; McMurray and Jongman, 2015). Clear speech provides another unique source of information, which is not only predictable based on prior knowledge but also accessible within the signal (as clear-speech cues are integrated in the speech signal). This allowed us to compare the relative contributions of both expectation and signal. Second, the current design, including both trial-level and session-level comparisons and both word and style identification, enabled us to evaluate expectations formed through short-term exposure to a speech context as well as through long-term prior experience.

Our first hypothesis was tested through the manipulation of trial-by-trial "style" prompt-target matching (Analysis 1), predicting knowledge-based expectations with greater word identification accuracy in matched than mismatched prompt-target trials, or signal-dependent perception with greater accuracy in clear than plain trials regardless of matching condition. Specifically, Analysis 1 involves a word identification task. First, the results show that identification of clear tokens is more accurate than for plain tokens across all three modalities (AO, VO, and AV), consistent with previous findings in similar studies indicating a facilitative effect of clear speech (e.g., Redmon et al., 2020). Interestingly, participants did not differ in identification of a word after being provided with correct, as opposed to incorrect, information about the speech style of the word, with similar identification rates in a within-subjects comparison between matched and mismatched prompt-target trials. These patterns reveal that information in the signal from the productions was critical in facilitating word identification for clear productions over plain productions, with neither matching nor mismatching speech style information contributing to identification accuracy. The results from Analysis 1 thus favor signal over expectation; that is, information in the signal of the clear productions, rather than knowledge about clear speech, contributes to better identification.

The second question addressed effects of expectations arising from short-term enhancement and exposure to clear speech as compared to those effects based on inconsistent contextual information. To this end, Analysis 2 explored both signal contributions and contextual contributions to word identification by examining conditions when matched speech-style prompts were presented. For these congruent-only conditions, only matched prompt-target trials were examined with clear tokens prompted by the "clearly spoken" screen, and plain tokens preceded by a blank screen, so expectations of the productions matched the prompts. These matched trials were presented in either congruent-only sessions (only matched prompts) or in mixed sessions (with both matched and mismatched prompts), as a way to directly examine expectations. As with the previous analysis, identification of clear tokens was more accurate than plain tokens across all three modalities (AO, VO

and AV), even when style information is provided to the participants, showing a robust contribution of signal-dependent information. Comparing the matched prompt-target trials in the congruent-only session with those in the mixed session, participants showed a marginal benefit from knowledge of the correct speech style information when consistent style information was present, particularly in the AO and AV conditions. This suggests a contribution of expectation within the congruent-only sessions (in contrast to the mixed session where exposure to both congruent and incongruent style information cannot help form consistent expectations). Thus, expectations result from the exposure to, and presence of, consistent information in the experimental context. Such repetitive information about speech style and subsequent exposure to the signal congruent with this information reinforced the signal characteristics, providing a statistical learning context to facilitate more accurate perception. Together, these results are aligned with previous discussion that a robust speech perception system may adapt statistical knowledge acquired in spontaneous situations as well as perceivers' prior knowledge (Feldman et al., 2009; Kleinschidt and Jaeger, 2015; McMurray and Jongman, 2015; Norris et al., 2008).

Beyond segment-level identification, we explored the longer-term effect of expectations based on high-level, signal-independent knowledge of speech styles. Analysis 3 focussed on an explicit determination of speech style, examining whether identification of the speech style of a word improved after completing the congruent-only session compared to the mixed session (which included both congruent and incongruent trials with matching and mismatching style information provided, respectively). The results reveal that identification of speech style was more accurate for clear tokens compared to plain tokens in the AV modality, with a trend observed in the AO modality. Clearly produced tokens were more accurately identified as a clear speech style production, showing a contribution of signal-only information even in a speech style identification task, with more accurate recognition of speech style when the stimuli were in clear (compared to plain) speech, in AV and AO modalities. Most interestingly, for this speech style determination, identification of speech style after congruent sessions was significantly more accurate than identification of speech style after mixed sessions, with congruent sessions aiding in speech style determination while mixed information, including both matching congruent and mismatching incongruent speech style information, perturbed expectations and interfered with identification of speech style. These results suggest longer term effects of conflicting speech style information, an additional complementary process in addition to signal-level information, involving accessing exposure information (matching versus mismatching), and comparing this information to internalized speech style knowledge. Longer-term effects of consistent speech style information resulted in enhanced knowledge of style. In sum, Analysis 3 suggested that the identification of high-level information of speech style is affected by both signal- and knowledge-based processes, corroborating the notion that high-level signal-complementary information (such as speech style) is embodied in, rather than separate from, the signal (Lindblom, 1990).

Finally, we observed that patterns of knowledge-based expectations interact with specific features of the speech signal (i.e., tense vs lax vowels, auditory vs visual speech information). Analyses 2 and 3 showed that, in contrast to the auditory domain (AO, AV), visual-only (VO) perception did not benefit from prior experience with different speech styles. A closer inspection also revealed an interaction between speech style and vowel tensity in VO, where clear speech, while beneficial for tense vowel perception, turned out to be harmful for lax vowel perception. This detrimental effect has also been observed previously and had been attributed to incompatible visual clear-speech modifications and the intrinsic properties of lax vowels (Leung et al., 2016; Redmon et al., 2020). Thus, for VO, where speech style may result in ambiguity, it may have been difficult for perceivers to form reliable expectations. Knowledge-based expectations were only observed in AO and AV modalities, where consistent cues to speech style are present in the signal. These patterns again imply the integrated effects of signal and expectation.

Overall, these data clearly show that speech perception involves processing sources of information which are driven by both the signal and contextual signal-independent expectations. Perception involves constant comparisons between signal input and linguistic knowledge in the form of prior exposure, as well as expectation and contextual contributions.

These findings point to possible avenues for future research. As the current study tested perception without a spontaneous interlocutor or communicative goal, the dynamicity of signal input and contextual expectations cannot be evaluated. Future studies could involve both talkers and perceivers in a conversational setting, where interlocutors may constantly shift their cue-weighting patterns for intelligibility gain based on their awareness of the talker's strategy and the perceiver's condition. Such research will unravel how conversational interlocutors strike a balance between making signal-based adaptations to strengthen overall speech salience and optimizing mutual intelligibility through experience- and knowledge-based processes.

CRediT authorship contribution statement

Joan A. Sereno: Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Allard Jongman:** Writing – review & editing, Methodology, Conceptualization. **Yue Wang:** Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization. **Paul Tupper:** Formal analysis, Conceptualization. **Dawn M. Behne:** Conceptualization. **Jetic Gu:** Software, Methodology. **Haoyao Ruan:** Visualization, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Yue Wang reports financial support was provided by Government of Canada Social Sciences and Humanities Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was funded by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC Insight Grant 435–2019–1065). We thank Simrin Bains, Junette Gonzales, Alysha Milne, and Samantha Sundby from the Language and Brain Lab at Simon Fraser University (SFU) for their assistance in data collection; and the SFU Linguistics Tech Team for technical assistance.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.specom.2025.103243.

Data availability

Data will be made available on request.

References

Apfelbaum, K.S., Bullock-Rest, N., Rhone, A., Jongman, A., McMurray, B., 2014. Contingent categorization in speech perception. Lang. Cogn. Neurosci. 29, 1070–1082.

Benz, A., Jäger, G., Van Rooij, R., 2006. An introduction to game theory for linguists. Game Theory and Pragmatics. Springer, pp. 1–82.

Bradlow, A.R., Bent, T., 2002. The clear speech effect for non-native listeners. J. Acoust. Soc. Am. 112, 272–284.

Feldman, N.H., Griffiths, T.L., Morgan, J.L., 2009. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychol. Rev. 116, 752–782.
Ferguson, S.H., Kewley-Port, D., 2002. Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 112, 259–271.
Fowler, C.A., Smith, M., 1986. Speech perception as "vector analysis": an approach to the problems of segmentation and invariance. In: Perkell, J.S., Klatt, D. (Eds.), Invariance and Variability in Speech Processes. Erlbaum, Hillsdale, NJ, pp. 123–136.
Gagné, J.P., Rochette, A.J., Charest, M., 2002. Auditory, visual and audiovisual clear speech. Speech. Commun. 37, 213–230.
Gopal, H.S., 1990. Effects of speaking rate on the behavior of tense and lax vowel durations. J. Phon. 18, 497–518.
Johnson, K., Strand, E.A., D'Imperio, M., 1999. Auditory–visual integration of talker gender in vowel perception. J. Phon. 27, 359–384.
Jongman, A., McMurray, B., 2017. On invariance: acoustic input meets listener expectations. In: Lahiri, A., Kotzor, S. (Eds.), The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches. Mouton De Gruyter, Berlin.
Jäger, G., 2008. Applications of game theory in linguistics. Lang. Linguist. Compass. 2, 406–421.
Kim, J., Sironic, A., Davis, C., 2011. Hearing speech in noise: seeing a loud talker is better. Perception. 40, 853–862.
Kleinschmidt, D.F., Jaeger, T.F., 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol. Rev. 122, 148–203.
Lam, J., Tjaden, K., Wilding, G., 2012. Acoustics of clear speech: effect of instruction. J. Speech, Language, Hear. Res. 55, 1807–1821.
Leung, K.K.W., Jongman, A., Wang, Y., Sereno, J.A., 2016. Acoustic characteristics of clearly spoken English tense and lax vowels. J. Acoust. Soc. Am. 140, 45–58.
Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W., Marchal, A. (Eds.), Speech Production and Speech Modelling. Springer, Dordrecht, pp. 403–439.
Luce, P.A., 1986. Neighborhoods of Words Lexicon. Psychology, Indiana University. Doctoral dissertation.
Maniwa, K., Jongman, A., Wade, T., 2009. Acoustic characteristics of clearly spoken English fricatives. J. Acoust. Soc. Am. 125, 3962–3973.
Maniwa, K., Jongman, A., Wade, T., 2008. Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. J. Acoust. Soc. Am. 123, 1114–1125.
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. Psychol. Rev. 118, 219–246.
McMurray, B., Jongman, A., 2015. What comes after [f]? Prediction in speech is a product of expectation and signal. Psychol. Sci. 27, 43–52.
Niedzielski, N., 1999. The effect of social information on the perception of sociolinguistic variables. J. Lang. Soc. Psychol. 18, 62–85.
Norris, D., McQueen, J.M., Shortlist, B., 2008. A Bayesian model of continuous speech recognition. Psychol. Rev. 115, 357–395.
Redmon, C., Leung, K., Wang, Y., McMurray, B., Jongman, A., Sereno, J.A., 2020. Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information. J. Phon. 81, 1–25.
Smiljanić, R., 2023. Clear speech processing benefits beyond intelligibility. J. Acoust. Soc. Am. 153 (3_supplement). A121-A121.
Smiljanić, R., 2021. Clear speech perception. In: Nygaard, L.C., Pardo, J., Pisoni, D., Remez, R. (Eds.), The Handbook of Speech Perception, 2nd ed., pp. 177–205.
Smiljanić, R., Sladen, D., 2013. Acoustic and semantic enhancements for children with cochlear implants. J. Speech, Language, Hear. Res. 56, 1085–1096.
Strand, E.A., 1999. Uncovering the role of gender stereotypes in speech perception. J. Lang. Soc. Psychol. 18, 86–100.
Tang, L.Y.W., Hannah, B., Jongman, A., Sereno, J., Wang, Y., Hamarneh, G., 2015. Examining visible articulatory features in clear and plain speech. Speech. Commun. 75, 1–13.
Tupper, P., Jian, J., Leung, K., Wang, Y., 2018. Game theoretic models of clear versus plain speech. In: Proceedings of the 40th Annual Meeting of the Cognitive Science Society. CogSci 2018, pp. 1133–1138.
Upton, G., Cook, I., 1996. Understanding Statistics. Oxford University Press, p. 55. ISBN 0-19-914391-9.
van der Feest, S.V.H., Blanco, C.P., Smiljanic, R., 2019. Influence of speaking style adaptations and semantic context on the time course of word recognition in quiet and in noise. J. Phon. 73, 158–177.
Van Engen, K.J., Phelps, J.E., Smiljanić, R., Chandrasekaran, B., 2014. Enhancing speech intelligibility: interactions among context, modality, speech style, and masker. J. Speech, Language, Hear. Res. 57, 1908–1918.
Wang, Y., Behne, D.M., Jiang, H., 2008. Linguistic experience and audiovisual perception of non-native fricatives. J. Acoust. Soc. Am. 124, 1716–1726.
Zeng, Y., Leung, K.K.W., Jongman, A., Sereno, J.A., Wang, Y., 2023. Multi-modal cross-linguistic perception of Mandarin tones in clear speech. Front. Hum. Neurosci. 17, 1–13.
Zhao, Y., Jurafsky, D., 2009. The effect of lexical frequency and Lombard reflex on tone hyperarticulation. J. Phon. 37, 231–247.