Difficult to hear but easy to see: Audio-visual perception of the /r/-/w/ contrast in Anglo-English

Hannah King, Ioana Chitoran

Difficult to hear but easy to see: Audio-visual perception of the /r/-/w/ contrast in Anglo-English

Hannah King[a)] and Ioana Chitoran

Université Paris Cité, UFR Linguistique, CLILLAC-ARP, F-75013 Paris, France

This paper investigates the influence of visual cues in the perception of the /r/-/w/ contrast in Anglo-English. Audio-visual perception of Anglo-English /r/ warrants attention because productions are increasingly non-lingual, labiodental (e.g., [ʋ]), possibly involving visual prominence of the lips for the post-alveolar approximant [ɹ]. 40 native speakers identified [ɹ] and [w] stimuli in 4 presentation modalities: auditory-only, visual-only, congruous audio-visual and incongruous audio-visual. Auditory stimuli were presented in noise. The results indicate that native Anglo-English speakers can identify [ɹ] and [w] from visual information alone with almost perfect accuracy. Furthermore, visual cues dominate the perception of the /r/-/w/ contrast when auditory and visual cues are mismatched. However, auditory perception is ambiguous because participants tend to perceive both [ɹ] and [w] as /r/. Auditory ambiguity is related to Anglo-English listeners' exposure to acoustic variation for /r/, especially to [ʋ], which is often confused with [w]. It is suggested that a specific labial configuration for Anglo-English /r/ encodes the contrast with /w/ visually, compensating for the ambiguous auditory contrast. An Audio-Visual Enhancement Hypothesis is proposed, and the findings are discussed with regard to sound change.

To cite:

King, H., and Chitoran, I. (2022). "Difficult to hear but easy to see: Audio-visual perception of the /r/-/w/ contrast in Anglo-English", *The Journal of the Acoustical Society of America* 152, 368-379, https://doi.org/10.1121/10.0012660.

[a)]Also at: UFR Études Anglophones; hannahhmking@gmail.com

I. INTRODUCTION

A. The influence of visual labial cues on speech perception

The lips play a fundamental role in spoken language. As well as contributing to the size and shape of the vocal tract, and thus to the acoustics of speech, the lips are also a visible articulator, providing a complementary source of information to the auditory stream in face-to-face communication. Although audition is the primary mode of perception in spoken language, perception is influenced by what we see as well as by what we hear. By presenting information about the position of a speaker's articulators, the lips may provide visual phonetic cues to the place of articulation of speech sounds. A large body of research has shown that visual information aids speech perception (Massaro, 1987, 1998), and that perception is more accurate when listeners are able to see the speaker as well as hear them. For example, speech comprehension is dramatically improved by visual cues from the speaker's lips when the auditory conditions are degraded due to hearing loss or environmental noise (e.g., Grant *et al.*, 1998; Lalonde and Werner, 2019; Ross *et al.*, 2007; Sumby and Pollack, 1954).

The most famous demonstration of the impact of visual speech cues on auditory speech perception occurs in the McGurk Effect, in which the phonetic properties of conflicting auditory and visual cues combine to form a new, fused auditory percept (McGurk and Macdonald, 1976). A similar but lesser known illusion, *visual capture*, is arguably even more dramatic. It occurs when listeners who are perceiving incongruous audio-visual speech report hearing the visually presented sound instead of the auditory one (Mattheyses and Verhelst, 2015). Visual capture may be anticipated when the visible articulation unambiguously specifies the phoneme under presentation (Werker *et al.*, 1992). This illusion suggests that in some cases, phonetic cues provided by vision may be salient enough to override auditory ones, indicating that visual speech cues may hold as much perceptual weight as auditory ones, and more under certain conditions.

B. The influence of labial cues in the production and perception of Anglo-English /r/

In this paper, we assess the impact of visual cues on the perception of word-initial /r/ in non-rhotic varieties of English spoken in England, henceforth Anglo-English. The term Anglo-English, rather than British English, is employed to avoid confusion with the varieties

of English spoken in Scotland, Wales and Northern Ireland (as in Lawson *et al.*, 2018, among others).

The role of the lips in the production and perception of /r/ in Anglo-English warrants attention because non-lingual, labiodental productions (e.g., [ʋ]) commonly occur. This labiodentalization of /r/ seems part of an accent levelling process which typically affects consonants, and has origins in the south east of England (Foulkes and Docherty, 2000). Although [ʋ] is established as a widespread feature of non-standard south-eastern accents (Foulkes and Docherty, 2000; Wells, 1982), instances have been reported throughout the country including Norwich (Trudgill, 1974), Milton Keynes, Reading, Hull (Williams and Kerswill, 1999), Derby (Foulkes and Docherty, 2000), Leeds (Marsden, 2006), Middlesborough (Llamas, 1998) and Newcastle (Foulkes and Docherty, 2000). This variant is considered one of the most rapidly advancing changes in Anglo-English at present, and some even suggest that it is becoming the norm among younger speakers in urban areas (Hornsby, 2014). Where [ʋ] was once considered 'defective' or an affectation of upper class speech (e.g., Gimson, 1980, p. 207), dialectological evidence suggests that [ʋ] has since become less stigmatized (see Foulkes and Docherty, 2000).

Labiodental variants may have emerged due to speakers dropping the lingual articulation of the post-alveolar approximant [ɹ], leaving the labial one to form the primary constriction (Docherty and Foulkes, 2001; Jones, 1972). The lingual articulation of [ɹ] is well known for its substantial variability. Tongue shapes range from tip-down bunched to curled-back retroflex in rhotic Englishes, e.g., North America (Delattre and Freeman, 1968; Mielke *et al.*, 2016; Tiede *et al.*, 2004; Zhou *et al.*, 2008) and Scotland (Lawson *et al.*, 2011, 2014), and in non-rhotic Englishes, e.g., New Zealand (Heyne *et al.*, 2018) and Anglo-English (King and Ferragne, 2020b). The different tongue shapes result in equivalent acoustic signals up to the first three formants (Zhou *et al.*, 2008), characterized by a low F3, generally below 2 000 Hz (e.g., Boyce and Espy-Wilson, 1997; Delattre and Freeman, 1968), in proximity to F2 (e.g., O'Connor *et al.*, 1957; Stevens, 1998).

Although acoustic data is scarce, [ʋ] is characterized by a higher F3 than its lingual counterpart [ɹ], at around 2 200 Hz (Foulkes and Docherty, 2000). Therefore, [ʋ] may actually share more acoustic properties with [w] than with [ɹ] (Dalcher *et al.*, 2008), and perceptual confusion between [ʋ] and [w] is widely reported anecdotally (Foulkes and Docherty, 2000).

[ɹ] in English is often described as labialized, which can be considered an articulatory enhancement strategy, as lip protrusion contributes to F3 lowering (King and Ferragne, 2020b). If the rise in labiodental variants of /r/ is due to speakers retaining the labial gesture

for [ɹ] at the expense of the lingual one, [ɹ] should be produced with a secondary labiodental gesture in Anglo-English. This hypothesis was examined in a study comparing the labial postures of [ɹ] and [w] in Anglo-English (King and Ferragne, 2020a). It was predicted that if [ɹ] is labiodental, the labial gesture for [w], which is unequivocally considered rounded, should differ substantially. Techniques from deep learning were used to automatically classify and measure the lip postures for [ɹ] and [w] from static images of the lips in 23 native speakers. The results indicate that there is a recognizable difference between the lip postures for [ɹ] and [w], which a convolutional neural network can detect with a very high degree of accuracy. Measurements of the lip area acquired using an artificial neural network indicated that [ɹ] indeed has a more labiodental-like lip posture than [w].

It has been suggested that the change towards exclusively labiodental variants of /r/ in Anglo-English may be due to the heavy visual prominence of the lips for [ɹ] (Docherty and Foulkes, 2001). We investigate this proposal in the present study by manipulating the audio-visual experience of native Anglo-English perceivers in normal, noisy listening conditions. Visual cues provide the greatest contribution to speech perception in noise, which is the normal context in which spoken language is communicated (Sumby and Pollack, 1954). If the labial posture for [ɹ] is visually salient, we expect visual cues from the lips to enhance auditory perception of the /r/-/w/ contrast when presented in noise. Audio-visual should therefore be better than auditory-only perception. The addition of noise should also prevent participants from reaching ceiling in auditory-only perception and allow them room for improvement with the addition of visual cues (as described in Van Engen *et al.*, 2017). Previous research generally shows that subjects perform less well in visual-only than auditory-only speech perception (e.g., Summerfield *et al.*, 1992). However, a study of English fricative perception found that /f, v/ are better distinguished from /θ, ð/ based on visual information alone than on auditory information, the former being no less informative than the combined audio-visual condition (Jongman *et al.*, 2003). Thus, if visual information from the lips is particularly perceptually salient for [ɹ], we may expect similar or even better performance in visual-only than auditory-only perception. Finally, if the visual cues for [ɹ] and [w] are phonetically unambiguous, we may anticipate visual capture when subjects are presented with incongruous audio-visual stimuli.

II. MATERIALS AND METHODS

A. Participants

40 native Anglo-English speakers (21F) aged between 18 and 73 (mean = 41.32 ± 17.92) took part in the study, which was conducted in North Yorkshire. 8 participants were recruited at the University of York, where ethical approval had been granted. Subjects at the university were undergraduates and were either financially compensated (£5) for their participation or gained class credit for linguistics courses. The remaining 32 participants were recruited among the first author's connections in the area. They were offered monetary compensation but chose to participate voluntarily. Participants self-identified as speaking with a native Anglo-English accent, which the first author, who is a native speaker, verified by conversing with them. This study did not examine which variant of /r/ the participants themselves used, but a future study matching subjects' production and perception would be valuable to examine links between the two. Subjects signed an informed consent form and completed a background questionnaire. None of them reported having any known speech or language disorders. Participants provided the region in which they spent the most time growing up, until the age of 18. 3 subjects spent most of their childhood outside the UK. The remaining subjects grew up in the following regions of England: North East ($n = 23$), North West ($n = 5$), Midlands ($n = 2$), South East ($n = 5$), South West ($n = 1$), various ($n = 1$). All subjects but one had normal or normal-corrected vision. A hearing performance score out of 30 was attributed to each subject based on their responses to 6 questions, which asked them to judge to what extent their hearing suffered in typical listening scenarios[1]. A list of these questions is included as supplementary materials[2]. Subjects scored a maximum of 30 points if they never experienced hearing problems. Mean hearing was 25.25 ± 4.35. Based on the questionnaire, no participants were eliminated from the study.

B. Stimuli

Stimuli were a list of monosyllabic CV(C)C minimal pairs contrasting /r, w, l/ word initially. [l] was included as a control because it is not produced with labialization, contrary to [w] and [ɹ]. To avoid labial coarticulation with the following vowel, the onsets occurred in the context of the non-rounded vowels [iː, ɪ, e, æ, eɪ, aɪ]. The coda consonant(s) was also never a labial. Each onset and vowel combination was assigned two items, resulting in

36 test words. For fillers and controls, we produced a list of the same number of minimal pairs contrasting word-initial /h/, /s/ and /θ/ (or /ð/ when minimal pairs with /θ/ are not attested). As our main concern was to find perfect minimal pairs, the materials could not be controlled for word frequency. The appendix presents a list of the test words and their respective frequency scores (in Zipf-scale) according to the SUBTLEX-UK database, which includes word frequencies from a corpus of 201.3 million words (van Heuven *et al.*, 2014).

A female 22-year-old native Anglo-English speaker was video-recorded reading the word list in a sound-attenuated booth at Université Paris Cité, France. Five tokens of each word were recorded in a semi-randomized order, avoiding sequences of words with /r/ and /w/ onsets. Audio and video recordings were made using a Zoom Q2HD Handy Video Recorder. Video was set to a resolution of 1 280 × 720 pixels recording 59.94 frames per second. The video camera's built-in condenser microphone was used to record the audio signal, which was digitized as a PCM stereo file with a 44 100 Hz sampling rate and 16-bit quantization. The resulting audio file was converted from stereo to mono during the post-processing stage in Praat (Boersma and Weenink, 2019) by extracting the left channel. The video camera was stabilized relative to the head using a bicycle helmet to which the camera was attached using a flexible arm (recycled from a pop filter) and a handheld tripod. We positioned the camera to capture front-facing images of the bottom half of the speaker's face, from her nose to her chin. The speaker audibly produced [ɹ], which was later confirmed with acoustic analysis conducted in Praat by extracting formant values using the Burg algorithm. For each token, formant parameters were manually adjusted to reach an optimal match between formant estimation and the underlying spectrogram by adjusting the ceiling of the formant search range. The most salient acoustic features of [ɹ] and [w] were used for formant extraction. The point at which F3 was minimally low for [ɹ] and F2 was minimally low for [w] was labelled by hand for each token and the first three formants (F1-F3) were extracted at these points. On average, F3 was 628 Hz lower and F2 was 380 Hz higher for [ɹ] than for [w] (see Table S1 in the supplementary material for summary statistics[2]).

We measured the speaker's lip dimensions in [ɹ] and [w] tokens, as well as in a neutral setting prior to speech. A physical ruler was placed below the speaker's lips touching her chin, which was video recorded. One video frame presenting an image of the ruler was extracted and opened in ImageJ (Schneider *et al.*, 2012). A straight line was positioned from 0 to 10 cm along the ruler, which yielded a global measurement scale for all subsequent measures. Video files for each token were opened in ImageJ and the image presenting maximum labial constriction for [ɹ] and [w] was selected by holistically examining sequential video frames.

Lip width was measured by placing a quasi-horizontal line from lip corner to corner. We measured lip aperture by positioning a straight vertical line from the vermilion border of the top lip, just below the philtrum dimple, down to the vermilion border of the bottom lip. The position of the mid-point of the lip aperture line along the y-axis was used to measure the vertical position of the lips. Figure 1 presents example images of the placement of lip dimension lines. We observed the same pattern as was previously reported (King and Ferragne, 2020a): the speaker's lips were wider and higher for [ɹ] than they were for [w], possibly indicative of a non-rounded, labiodental lip posture (see Table S2 for summary statistics of lip dimensions[2]). Figure 2 presents an example image of the lips in a neutral setting and of maximum labial constriction for [ɹ] and [w] from the minimal pairs *red* and *wed*. While [w] has a visibly rounded posture, the bottom lip is in proximity to the front surface of the upper incisors for [ɹ], again suggestive of a labiodental posture.

FIG. 1. Lip width (left) and lip aperture (right) lines positioned to generate lip dimension measures.

Auditory stimuli for the perception experiment were embedded in pink noise. Pink rather than white noise was selected because it is the most effective masker (Adachi *et al.*, 2006). Pink noise was mixed with the audio files at a signal-to-noise (SNR) ratio of -12 dB and mean amplitude was scaled to 70 db using a Praat script adapted from McCloy (2013). An SNR of -12 dB was used because it has been identified as a 'special zone' where audio-visual benefit is maximal (Ross *et al.*, 2007).

Perception trials were created using VirtualDub (Lee, 2000) in the following modalities: visual-only (VO), auditory-only (AO), congruous audio-visual (AVc) and incongruous audio-visual (AVi). Example videos of perception trials in each modality are included as supplementary material[2], as well as an overview of the auditory and visual cues presented

FIG. 2. Example images depicting a neutral lip setting (left); maximum labial constriction for [ɹ] (middle) and [w] (right).

within (Table S3). We generated VO trials by replacing the audio track from audio-visual stimuli with pink noise. AO trials were produced by embedding pink noise in the audio track, combined with a still image of the speaker's face prior to speech to enforce ongoing attention to visual speech cues throughout the experiment. The same video still was used for all VO trials, which was the one used to calculate neutral lip dimensions, as presented in Fig. 2. We created AVc trials by embedding pink noise in the audio track from the audio-visual stimuli[3]. To create AVi trials, we dubbed the audio track of one sound (mixed with pink noise) over the video track of another.

Three tokens of each item were used to generate perception trials. The same token was used to create AO and VO trials. AVc trials were produced with different tokens from the ones used in AO and VO to avoid familiarization with the same token in the course of the experiment. 216 trials were generated for these three modalities, including test words, fillers and controls. To reduce experimental time, half of the AO, VO and AVc trials were presented to a single group of participants (108 trials per group), counterbalanced across modalities and words. Different tokens of word-initial /r/ and /w/ were used to generate 24 AVi trials, in which the audio-visual word pairings were matched as closely as possible in word length (mean difference = 8.65 ± 6.41 ms). A further 24 AVi trials, which combined word-initial audio-visual /s/-/θ/ (or /ð/), were generated for controls because they allow for straightforward predictions. The interdental articulation of [θ] and [ð] should be relatively visible contrary to [s] whose primary articulation occurs inside the mouth. We therefore predicted that incongruous audio-visual /s/-/θ/ pairs would induce visual capture, contrary to /θ/-/s/. Participants were shown all 48 AVi items. There were thus 156 perception trials per group.

We created 10 catch trials to enforce ongoing attention to the video (Irwin *et al.*, 2011). Catch trials consisted of a random auditory token from the dataset (one which was not used in the experimental conditions), mixed with pink noise and combined with a still image of the speaker in which the region corresponding to the lips within the image was colored in. An example catch trial is included in the supplementary materials (Fig. S1)[2]. Five colors were used in total (blue, green, pink, purple and black). In these cases, participants were instructed to respond with the color of the speaker's lips and not with the word she said. All subjects but one correctly responded to at least 7/10 catch trials. The remaining subject, who correctly responded to 2 catch trials only, reported to have an uncorrected sight problem, and was therefore excluded from subsequent analyses.

C. Procedure

The perception experiment, which was carried out in PsychoPy (Peirce, 2007), took place in a quiet room either at the University of York, or in the participant's home or workplace. Subjects were seated in front of a portable laptop computer with a 13-inch screen. Audio was presented through a pair of AKG K271 headphones with the volume set to a comfortable level. Stimulus presentation of trials in all four modalities was randomly intermixed (as in Ross *et al.*, 2007), and trial order was unique to each participant. Each trial was preceded by a fixation cross of duration 2 000 ms, which participants were instructed to look at. Directly after stimulus presentation, participants identified the word-initial consonant they perceived by clicking on a word from two options using a wireless optical mouse. A 2 000 ms time limit was imposed on responses, after which the program automatically advanced to the next stimulus (as in Havenhill and Do, 2018). Subjects were instructed that their first mouse click would be recorded and were asked to respond as quickly and accurately as possible. They were provided with four equally-distributed, self-timed breaks. The experiment took around 20 minutes to complete.

D. Pre-processing

Three measures from the unimodal and congruous modalities (AO, VO, AVc) were analyzed in the experimental trials: accuracy, sensitivity and response bias. Accuracy scores were recorded for each trial for each participant by coding responses as correct or incorrect. Perceptual sensitivity to each contrast (/l/-/w/, /l/-/r/, and /r/-/w/) in each modality was

measured per participant using d'. Hit and false alarm rates were calculated by arbitrarily assigning correct responses for one of the phonemes in each pair as hits (as in McGuire and Babel, 2012). Hits were assigned to correct /l/ responses in the /l/-/w/ and /l/-/r/ contrasts, and to correct /r/ responses in the /r/-/w/ contrast. False alarms were assigned to incorrect responses of the same phonemes in each of the respective contrasts. Hit and false alarm rates of 0 and 1 were converted to $1/(2N)$ and $1 - 1/(2N)$ respectively, where N is the number of trials on which the proportion is based (Macmillan and Creelman, 2005). A $d' = 0$ indicates that a subject shows no sensitivity to a contrast. The maximal d' was just over 2.9, which we consider near-perfect perception. In an unplanned, exploratory analysis, the same hits and false alarms were used to measure participants' bias to respond with one of the phonemes in each contrast by calculating Criterion Location c (Macmillan and Creelman, 2005). Each subject's bias to respond with /l/ in the /l/-/w/ and /l/-/r/ contrasts, and with /r/ in the /r/-/w/ contrast was measured. A $c = 0$ indicates no response bias.

One measure from the incongruous audio-visual modality was analyzed: visual capture. Visual capture was recorded for each incongruous /r/-/w/ (test) and /s/-/θ/ (control) trial for each participant by coding visual and auditory responses as 1 and 0, respectively.

E. Statistical analysis

Accuracy and visual capture were assessed using binomial generalized linear mixed-effects models (GLMM) in R (R Core Team, 2018), using the glmer() function of the *lme4* (Bates *et al.*, 2015) package. Sensitivity and response bias were examined with linear mixed-effects models (LMM) using the lmer() function. Wherever possible, models included an interaction between modality and stimulus (or contrast), as well as the following main effects: age (continuous variable), subject origin (abroad or England), hearing score (continuous variable), sex (female, male), and word frequency (continuous Zipf-scale, as in Table III). We note that the dataset was far from balanced for subject origin, as only 3 out of the 39 participants grew up abroad. Continuous effects were converted to z-scores. We included the maximal set of successfully converging random slopes and intercepts for subjects, and, wherever possible, items. The significance of interactions and main effects was tested using likelihood ratio tests with the mixed() function of the *afex* package (Singmann *et al.*, 2015). Model output tables of the best-fitting models are provided as supplementary material[2], in which the lmerTest library (Kuznetsova *et al.*, 2017) was used to calculate indications of significance, which uses values from Satterthwaite's approximations for the degrees of

freedom. Post-hoc pairwise comparisons of significant interactions, which are also presented in tables as supplementary material[2], were conducted in *emmeans* (Lenth, 2021) and were adjusted for the multiple comparisons via Bonferroni correction. Model fit was assessed with a comparison of Akaike Information Criterion. For LMMs, model residuals were plotted to test for deviations from homoscedasticity or normality.

III. RESULTS

A. Perception of unimodal and congruous audio-visual trials

39 participants responded to 54 unimodal (AO, VO) and congruous audio-visual (AVc) trials presenting monosyllabic words beginning with /r/, /w/ and /l/ in noise, resulting in a total of 2 106 observations. Table I presents raw stimulus-response confusion matrices for these modalities. 126 trials were left unanswered, over a third of which occurred in the context of visual-only /l/. All unanswered trials were excluded, resulting in 1 980 analyzable observations. Descriptive statistics for sensitivity scores, response bias and the proportion of correct responses are provided in Table II. We next analyze these results by running a model of perceptual accuracy and a model of perceptual sensitivity.

1. Analysis of accuracy

We ran a GLMM with accuracy as the binary outcome variable, regressed against stimulus and modality with an interaction term, as well as age, subject origin, hearing score, sex and word frequency as main effects. Random intercepts for subjects and items were included.

The interaction between stimulus and modality was significant ($\chi^2(4) = 89.55$, $p < .001$). The main effects of both stimulus and modality were significant (Stimulus: $\chi^2(2) = 8.25$, $p = 0.02$; Modality: $\chi^2(2) = 171.04$, $p < .001$), as well as subject sex and origin (Sex: $\chi^2(1) = 5.47$, $p = 0.02$; Origin: $\chi^2(1) = 5.22$, $p = 0.02$). However, age, hearing score and word frequency failed to reach significance (Age: $\chi^2(1) = 2.37$, $p = 0.13$; Hearing: $\chi^2(1) = 0.64$, $p = 0.43$; Zipf: $\chi^2(1) = 2.44$, $p = 0.12$).

The results reveal that women and subjects who grew up in England had significantly higher accuracy rates overall (see Table S4 in the supplementary material for model output[2]). With regards to sex, this trend follows previous studies where women have been shown to be more sensitive to visual speech cues than men (e.g., Aloufy *et al.*, 1996; Watson *et al.*,

TABLE I. Raw stimulus-response confusion matrices in unimodal and congruous audio-visual modalities. ∅ corresponds to no response.

	Presented					
Responded	AO		VO		AVc	
	/l/	/w/	/l/	/w/	/l/	/w/
<l>	91	44	79	4	112	1
<w>	15	68	16	107	3	114
∅	11	5	22	6	2	2
	/l/	/r/	/l/	/r/	/l/	/r/
<l>	100	37	75	4	115	1
<r>	14	77	22	102	1	116
∅	3	3	20	11	1	0
	/r/	/w/	/r/	/w/	/r/	/w/
<r>	100	55	103	6	112	11
<w>	12	51	3	107	1	101
∅	5	11	11	4	4	5

1996). While the results could indicate that linguistic experience plays a role in perceptual accuracy, given the small number of participants born outside of England in the dataset ($n = 3$), the significance of subject origin remains inconclusive.

With regards to the significant interaction between stimulus and modality, Fig. 3 presents a plot of the predicted probability of accurately identifying each stimulus across the three modalities according to the best-fitting model. Post-hoc pairwise comparisons are supplied in Table S5 in the supplementary material[2]. The results indicate that auditory perception is enhanced by visual cues: accuracy is significantly better in AVc than in AO perception across the board ($p < .001$).

The model predicts perceptual accuracy to be near-perfect in the VO perception of /w/ (0.96 ± 0.02) and /r/ (0.97 ± 0.01). Accuracy is significantly lower in AO than VO perception for both /w/ and /r/ ($p < .001$), and accuracy is predicted to be lower than chance for

TABLE II. Mean d', c (0 = no bias, negative value = bias to respond with the first phoneme of the contrast), and proportion of correct responses (Pc). Standard deviations appear within parentheses.

Contrast	Mod.	d'	c	Pc
/l/-/w/	AO	1.26 (0.62)	-0.33 (0.43)	0.73 (0.12)
	VO	2.02 (0.87)	0.19 (0.32)	0.89 (0.15)
	AVc	2.71 (0.39)	0.04 (0.17)	0.98 (0.06)
/l/-/r/	AO	1.51 (0.86)	-0.27 (0.42)	0.78 (0.16)
	VO	1.90 (1.02)	0.23 (0.43)	0.87 (0.19)
	AVc	2.60 (0.27)	0.00 (0.13)	0.99 (0.04)
/r/-/w/	AO	1.05 (1.01)	-0.53 (0.47)	0.69 (0.19)
	VO	2.44 (0.54)	-0.04 (0.27)	0.96 (0.09)
	AVc	2.43 (0.03)	-0.12 (0.32)	0.95 (0.12)

/w/ in AO (0.46 ± 0.08). However, the model predicts accuracy to be well above chance for both AO /l/ (0.86 ± 0.04) and /r/ stimuli (0.77 ± 0.06), suggesting that despite the addition of pink noise, participants were still sensitive to acoustic cues in the experiment. Indeed, while the presence of a visual cue aided the identification of /r/ and /w/, /l/ tokens were still best identified when an auditory cue was also present. The probability of accurately identifying /l/ stimuli was lower in VO (0.77 ± 0.06) than in AO (0.86 ± 0.04), although this difference did not reach significance. Contrary to /r/ and /w/, the probability of accurately identifying /l/ tokens significantly improved from the VO modality with the presence of auditory cues in AVc ($p < .001$). /l/ was thus the only stimulus to benefit from the combination of auditory and visual cues. The perception of /r/ patterns with that of /w/ across all three modalities, suggesting that /r/, like /w/, has a visual labial cue which speakers may use as reliable phonetic information in perception.

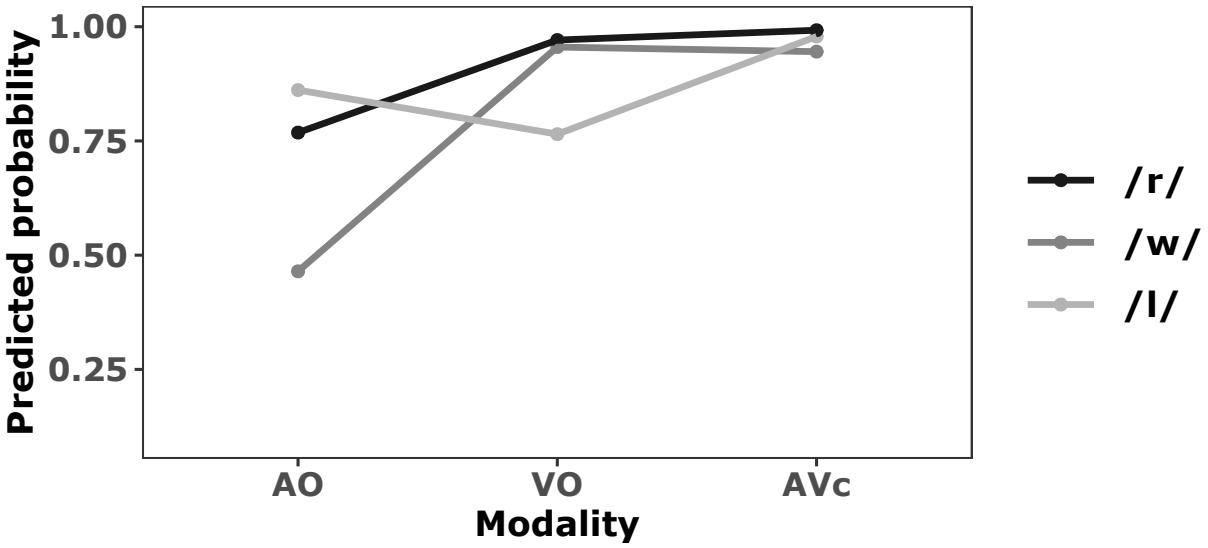


FIG. 3. Predicted probability of accurately identifying stimuli in each modality from a GLMM.

2. Analysis of sensitivity

Having determined that accuracy was very high for both /r/ and /w/ stimuli in VO, we now ask to what extent subjects were able to distinguish between the labial configurations for /r/ and /w/. To this end, an LMM analysis was implemented predicting subjects' sensitivity to each of the three contrasts in the three modalities. d' was the outcome variable which was regressed against contrast (/l/-/r/, /l/-/w/, /r/-/w/) and modality (AO, VO, AVc) with an interaction term. Subject age, origin, hearing score and sex were also included as main effects. The model contained random intercepts for subjects.

The interaction between contrast and modality was significant ($\chi^2(4) = 23.20$, $p < .001$). The main effect of modality was significant ($\chi^2(2) = 161.2$, $p < .001$), while contrast failed to reach significance ($\chi^2(2) = 0.15$, $p = 0.93$). As in the GLMM analysis of accuracy, the main effects of subject sex and origin were significant (Sex: $\chi^2(1) = 3.87$, $p < .05$; Origin: $\chi^2(1) = 7.11$, $p = 0.008$) with women and subjects who grew up in England predicted to have the highest sensitivity overall. Similarly, neither hearing score nor age reached significance (Hearing: $\chi^2(1) = 0.38$, $p = 0.54$; Age: $\chi^2(1) = 2.86$, $p = 0.09$). Model output is supplied in Table S6 in the supplementary materials[2].

Regarding the significant interaction between modality and contrast, Fig. 4 presents effects plots of sensitivity to the contrasts in each modality according to the best-fitting

model. Table S7 in the supplementary material provides post-hoc pairwise comparisons[2]. As in the GLMM analysis of accuracy, the results indicate that participants were sensitive to acoustic cues despite the addition of noise because the model predicts d' values to be much higher than zero in AO perception for all three contrasts. However, a discrepancy is observed between the sensitivity and accuracy models (cf. Fig. 3 & 4) in that in the sensitivity model, participants were more sensitive to visual cues than to auditory ones across the board. At first glance, the results from the sensitivity model may seem to contradict the general finding from previous research that participants are more successful at identifying speech in audio-only than in visual-only conditions. However, notice that the probability of accurately identifying each consonant (Fig. 3) is predicted to be higher in the visual-only condition only for /r/ and /w/. This is not the case for /l/. The visual-only condition does not present an advantage for the identification of /l/ as it does for /r/ and /w/. The results are consistent with our prediction that visual cues would be perceptually salient for /r/ and /w/, but not available for /l/, which was included in our study as a control. The inclusion of /l/ confirms that /r/ includes labial information that can be used in perception. If /r/ had no visual labial cue, we would expect VO perception of /r/-/l/ to be significantly worse than that of /w/-/l/, which was not the case.

As for the difference in lip postures between /r/ and /w/, while the /r/-/w/ contrast in AO has the lowest predicted d' of all the contexts (0.85), sensitivity to the contrast is significantly higher in VO ($p < .001$). The cumulative benefit of combining auditory and visual cues in AVc is only observed for the contrasts with /l/. For the /r/-/w/ contrast, no benefit was obtained from presenting an auditory stimulus alongside a visual one, i.e., there was no significant difference in sensitivity to the /r/-/w/ contrast between VO and AVc ($p = 0.99$). It seems then that visual cues are no less informative than the combined audio-visual condition, suggesting that visual cues are particularly salient for the /r/-/w/ contrast.

3. Analysis of response bias

The stimulus-response matrices presented in Table I show that in the AO trials, while subjects were generally able to accurately identify /r/ and /l/ tokens, the proportion of correctly identified /w/ tokens was comparatively lower. In fact, when presented with /w/ AO stimuli in the context of /r/ and /w/ responses, participants selected /r/ more often than /w/, suggesting there may be a preference for /r/. This apparent /r/ bias does not

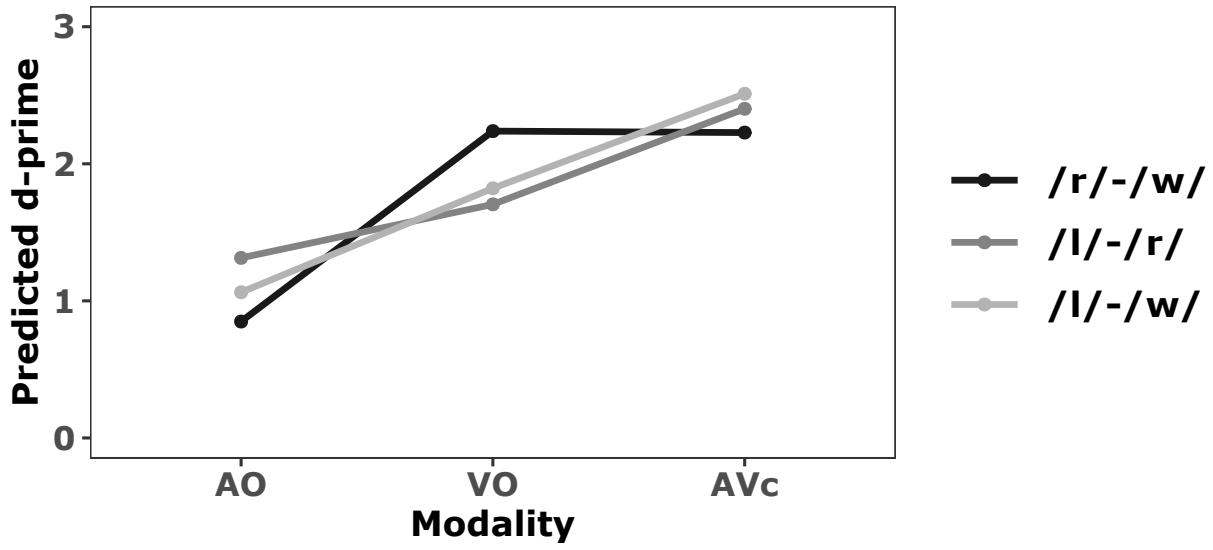


FIG. 4. Predicted sensitivity to contrasts in each modality from a LMM.

seem to extend to the other modalities. An unplanned, exploratory analysis of response bias was thus carried out using measures of Criterion Location (c). The mean (c) values presented in Table II paint a similar picture, as the mean c value that is furthest from zero occurs in the context of the /r/-/w/ contrast in the AO modality (-0.53). The mean c values in the other contexts are much closer to zero, suggestive of little response bias. Given these patterns, we decided to subset the data and only analyze response bias in the /r/-/w/ contrast. Another motivation for this choice is that the estimates from a model of criterion location on all the data would be extremely difficult to interpret because criterion location was calculated with respect to different phonemes in each contrast.

We implemented a LMM analysis predicting response bias in the /r/-/w/ contrast. c was the outcome variable, which was regressed against modality (AO, VO, AVc), subject age, origin, hearing score and sex. We also included random intercepts for subjects.

The only significant main effect was modality ($\chi^2(2) = 41.21$, $p < .001$). None of the other effects reached significance (Age: $\chi^2(1) = 0.04$, $p = 0.84$); Origin $\chi^2(1) = 1.46$, $p = 0.23$; Hearing $\chi^2(1) = 0.29$, $p = 0.59$; Sex: $\chi^2(1) = 0.03$, $p = 0.87$). Model output is supplied in Table S8 as supplementary material[2]. The modality effect indicated a significant difference in response bias between VO and AO ($p < 0.001$). There was no significant difference between VO and AVc ($p = 0.95$). A negative c corresponds to a bias to respond

with /r/, and the model estimates significantly lower c values in AO than VO, which can be interpreted as a significant bias for /r/ in the AO modality.

B. Perception of incongruous audio-visual trials

Subjects each responded to 48 trials in which auditory /s/ was dubbed over visual /θ/ (or /ð/) and vice versa, and auditory /r/ was dubbed over visual /w/ and vice versa. 74 trials were left unanswered and were excluded, leaving 1 798 analyzable observations. Fig. 5 presents plots of the overall proportions of auditory and visual responses induced by the four contexts. As predicted, higher numbers of visual responses arose from visual /θ/, /w/ and /r/ than /s/. We observe a very large proportion of visual responses in the case of visual /r/ paired with auditory /w/ (83.7%). The opposite context (visual /w/ with auditory /r/) resulted in a smaller proportion of visual responses, although visual responses were still more frequent than auditory ones (63.1%).

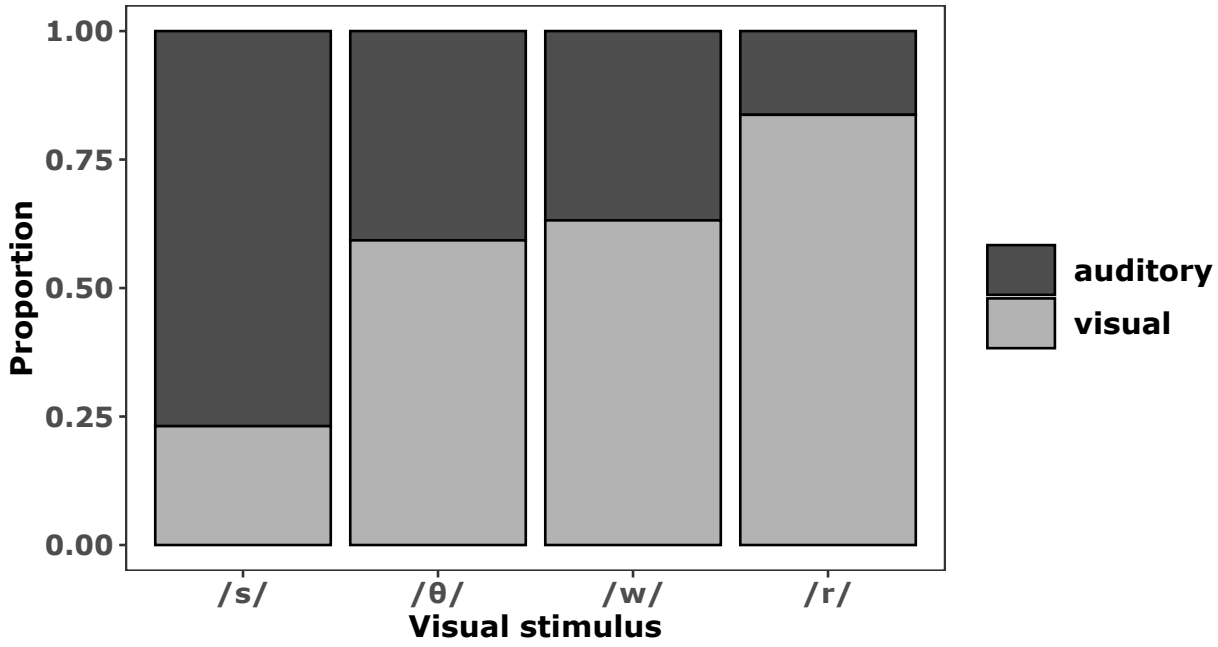


FIG. 5. Proportion of auditory and visual responses in incongruous audio-visual trials.

1. Analysis of visual capture

We ran a GLMM with visual capture as the binary outcome variable. Main effects included visual stimulus (/s/, /θ/, /w/, /r/), sex, age, origin and hearing score. Random intercepts were included for subjects and items.

The results reveal that visual stimulus was a significant main effect ($\chi^2(3) = 43.20$, $p < .001$). Subject sex was also significant ($\chi^2(1) = 4.51$, $p = 0.03$) with female subjects being more likely to select a visual response than males. None of the other main effects reached significance (Age: $\chi^2(1) = 0.20$, $p = 0.65$; Origin: $\chi^2(1) = 0.70$, $p = 0.40$; Hearing: $\chi^2(1) = 0.07$, $p = 0.79$).

Model output is supplied in Table S9 in the supplementary material[2]. Visual /s/ paired with auditory /θ/ resulted in significantly less visual capture than the three other contexts. By changing the reference level of the visual stimulus effect to /w/ and rerunning the model, we found no significant difference between /w/ and /θ/ visual stimuli ($p = 0.59$). In contrast, the model predicts that /r/ induces significantly more visual capture than /w/ ($p < .01$). /r/ has the highest predicted probability of visual capture among all the visual stimuli (0.88 ± 0.05), followed by /w/ (0.62 ± 0.11), /θ/ (0.56 ± 0.12) and /s/ (0.12 ± 0.05). The results support our prediction that visual capture arises for /r/ due to its unambiguous visual labial cue with respect to /w/.

IV. DISCUSSION

This study's main finding is that the visually distinct labial postures of Anglo-English [ɹ] and [w] are used by native observers as perceptual information. Participants were able to distinguish /r/ from /w/ from visual cues alone with near-perfect accuracy. Contrary to the contrasts involving /l/, the perceptual advantage from visual cues did not require auditory input whatsoever for the /r/-/w/ contrast, indicating that visual cues from the lips provide sufficient phonetic evidence to accurately perceive the contrast. High rates of visual capture in incongruous audio-visual pairings, especially in trials containing visual [ɹ] paired with auditory [w], suggest that the labial posture for [ɹ] is unambiguous with respect to [w]. The lip posture for [ɹ] induces significantly higher visual capture than those for [w] and [θ], despite all three having visible articulations. The results are consistent with the proposal by Docherty and Foulkes (2001) that the labial cue for Anglo-English [ɹ] is particularly visually

prominent. This could account for the change towards exclusively labial variants of /r/ in Anglo-English.

A. Reliance on visual cues for /r/

The question remains why the lips should be used as such a reliable visual cue for Anglo-English /r/ in the first place, given that audition is consistently defined as the primary mode of communication in spoken language. We will propose tentative answers to this question, drawing on Ohala's perception-oriented account of sound change (e.g., Ohala, 1981, 1996) as a framework from which to illustrate our argument. Ohala's account proposes that the main source of variation in speech, and hence the driving force behind sound change, is the misperception of the acoustic signal by the listener. He argues that much of the variation which underpins the acoustic speech signal is phonetically predictable. When confronted with variation, a key factor at play is the listener's phonetic experience, without which the listener is forced to take the acoustic signal at face value. Although Ohala's approach focuses on auditory perception, we extend his perceptual account of sound change to include visual cues, in accordance with the multimodal nature of perception. Further below, in Fig. 7, we present and discuss the relevant scenarios. Given that the current study presents synchronic data, we stress that these are potential implications of our findings, which will require further study with new data to confirm or deny any predictions made about sound change.

Anglo-English listeners are regularly confronted with phonetic variation for /r/. Tongue shapes vary widely for the post-alveolar approximant, yet the acoustic output of these articulations remains comparatively stable. However, productions without a specified lingual component, i.e., [ʊ], do not produce the same acoustic output. Due to its high F3, [ʊ] may share more acoustic properties with [w] than with [ɹ]. As labiodental variants are rapidly spreading throughout England (Foulkes and Docherty, 2000; Llamas, 1998; Marsden, 2006; Trudgill, 1974; Williams and Kerswill, 1999), experience with these variants must also be on the rise. Incidentally, increased exposure may also explain why [ʊ] is becoming less stigmatized. The sound change from [ɹ] to [ʊ] is likely to be phonetically gradient. The lingual articulation is gradually lost over time, resulting in the steady raising of F3. Exposure to this acoustic variation may shape speakers' tolerance of what constitutes an acceptable /r/ in perception. Indeed, it has been suggested that high exposure to [ʊ] may have resulted in a shift in the perceptual weighting of auditory cues for the /r/-/w/ contrast. Dalcher *et al.* (2008) compared the auditory perception of copy-synthesized approximant sounds in

American- and Anglo-English native listeners. A stimulus which combined a low [ɹ]-like F3 with a low [w]-like F2 was almost systematically perceived as /r/ by American listeners, but was more often identified as /w/ by Anglo-English subjects. Fig 6 presents stylized formant contrasts for [ɹ], [w] and [ʊ] based on those provided by Dalcher *et al.* (2008). As Fig 6 shows, [ɹ] contrasts with [w] both with respect to F2 and F3. However, F3 does not hugely differ between [ʊ] and [w]. Dalcher *et al.* (2008) concluded that /r/ is increasingly defined by F2 in Anglo-English, due to high exposure to the [ʊ] variant. In a similar vein, we propose that exposure to phonetic variation allows Anglo-English listeners to reconstruct [ʊ] as /r/, despite the acoustic proximity of [ʊ] to [w]. This scenario is consistent with Ohala's depiction of *perceptual compensation*, presented in Fig. 7(a).

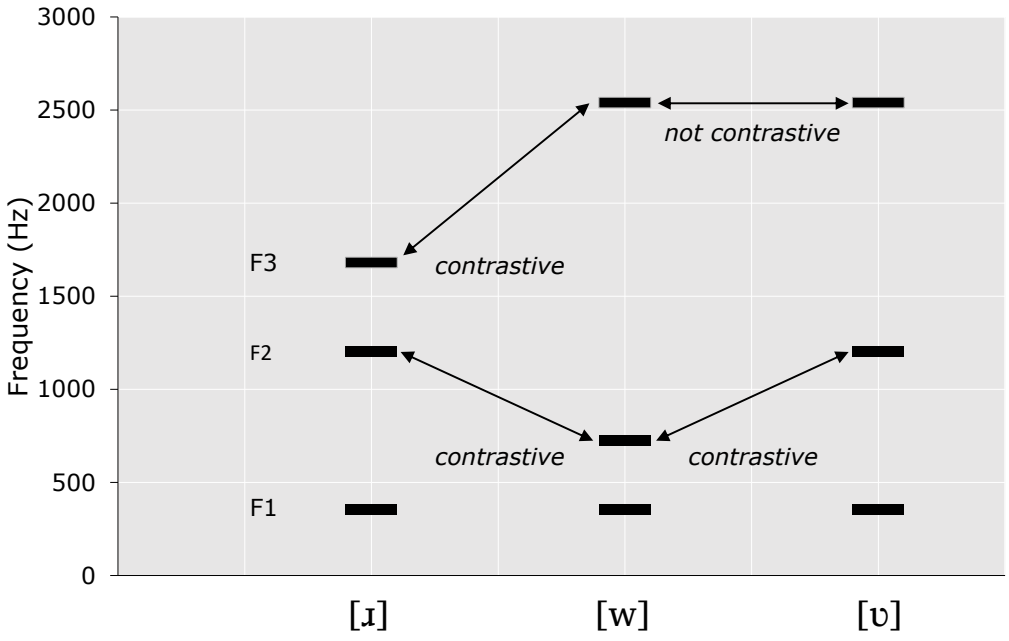
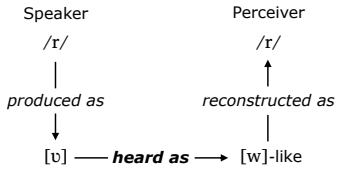
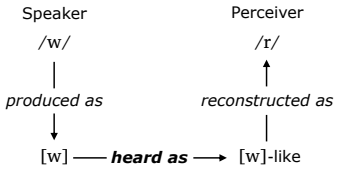


FIG. 6. Illustration of formant contrasts between pronunciation variants of /r/ and /w/ (based on those in Dalcher *et al.*, 2008)

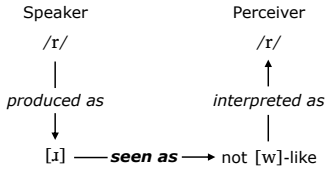
Tolerance for high F3 variants of /r/ may impact the perception of /w/. An unexpected result emerged in the auditory perception of the /r/-/w/ contrast in the present study. When presented with AO [w] stimuli, subjects reported perceiving /r/ more often than /w/, resulting in a bias for /r/. Given the acoustic similarity between [ʊ] and [w], we propose that a speaker's [w] productions may be erroneously reconstructed as /r/ by the listener, which would account for the observed /r/ bias in the identification of /w/-/r/ target-distractor



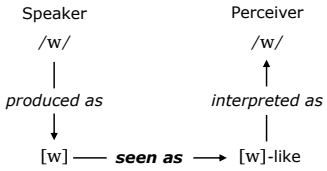
(a) Perceptual compensation



(b) Hypercorrection



(c) Accurate perception of [ɪ]



(d) Accurate perception of [w]

FIG. 7. Auditory (a, b) and visual (c, d) perception scenarios involving the Anglo-English /r/-/w/ contrast based on the perception-oriented account of sound change proposed by Ohala (1981). Slashes represent lexical forms, and square brackets denote surface phonetic forms.

pairs in AO perception. This is an example of what Ohala defines as *hypercorrection* and is schematized in Fig. 7(b).

We suggest that Anglo-English listeners tolerate such a high degree of acoustic variation for /r/ that even canonical productions of [w] may be reconstructed as /r/ in perception. However, when presented with the accompanying visual cues, the bias for /r/ responses disappears and sensitivity to the contrast is significantly enhanced. The two scenarios involving perceptual compensation (Fig. 7(a)) and hypercorrection (Fig. 7(b)) therefore no longer apply when listeners are able to see the speaker's lips. We contend that a specific labial cue for Anglo-English /r/ encodes and disambiguates the contrast with /w/. As such, when the listener sees a lip posture that is not [w]-like, they will likely interpret that production as /r/ and not as /w/, given their phonetic knowledge of the distinct lip configurations for [ɹ] and [w]. This scenario is presented in Fig. 7(c).

Similarly, a [w]-like visual cue will allow listeners to accurately identify the speaker's intended form as /w/ (cf. Fig. 7(d)). If we compare this scenario with the one of hypercorrection in Fig. 7(b), we notice that the presence of visual cues prevents the hypercorrection of [w] to /r/. By preventing hypercorrection, visual cues may avert potential misperception-based sound change. Ohala proposes that sound change may arise due to misperception in

the listener when the listener turns speaker. We may imagine an extension of Fig. 7(b), in which hypercorrection of [w] catalyzes a sound change toward more [w]-like realizations of /r/ when the listener turns speaker, as they increasingly associate [w]-like productions with /r/. However, visual cues may render such a sound change less likely by allowing the listener to accurately interpret productions of [ɹ] and [w] as /r/ and /w/, respectively.

An alternative account for the observed /r/ bias in AO perception of the /r/-/w/ contrast could involve word frequency. Although the frequency of the test words was not a significant predictor of perceptual accuracy in the experiment, a higher type frequency of word-initial /r/ than word-initial /w/ could explain why listeners tend to select /r/ rather than /w/ responses when presented with auditory [w]. To test this possibility, we again used the SUBTLEX-UK database (van Heuven *et al.*, 2014). A list of /r/- and /w/-initial words in the corpus was generated, based on the words' written forms[4]. /w/-initial words account for 6.67% of the dataset, while /r/-initial ones make up just 2.16%. We can therefore conclude that neither lexical nor type frequency account for the observed /r/-bias. We argue that /r/-bias is more likely to stem from native Anglo-English speakers' high tolerance for acoustic variability for /r/.

Finally, using this framework, we may make predictions about the perception of /r/ in Englishes in which [ʊ] does not occur, e.g., American English. Lack of experience with /r/ variation would force listeners to take the acoustic signal at face-value. It is therefore likely they would interpret [ʊ] as /w/, given its acoustic proximity to [w]. We would thus not expect the hypercorrection of auditory [w] to /r/ in American English listeners, nor would we expect them to be as influenced by visual cues as Anglo-English listeners. This remains to be tested.

B. Towards an Audio-Visual Enhancement Hypothesis

The results from this study indicate that visual information is reliably used in the perception of Anglo-English /r/. We propose that exposure to multiple phonetic variants of /r/ leads speakers to tolerate a high degree of acoustic variation for /r/, resulting in perceptual ambiguity between /r/ and /w/. However, visual cues from the speaker's lips allow perceivers to accurately differentiate /r/ from /w/, thus enhancing the contrast. In the event of auditory ambiguity, if visual speech cues are available, speakers will rely on them in perception. This idea of optimizing the distinctiveness of phonological contrasts is central to the *Auditory Enhancement Hypothesis* (Diehl and Kluender, 1989), which proposes that

the phonetic structure of languages is driven by properties of speech sounds that reinforce phonological contrasts. A typical example involves lip rounding in back vowels. Back vowels are generally produced with lip rounding, which enhances the auditory effect of tongue backing by contributing to F2 lowering. In contrast, fewer instances of lip rounding occur in front vowels where lip rounding counteracts the acoustic effect of tongue fronting.

In terms of sound change, the Auditory Enhancement Hypothesis would predict that new combinations of articulations would develop when an existing phonological contrast is insufficiently perceptually salient. Parallels may thus be drawn between this framework and the conclusions from this study. However, by definition, the Auditory Enhancement Hypothesis is concerned exclusively with auditory speech perception. A logical extension, which we put forward here, would be an *Audio-Visual Enhancement Hypothesis*, thereby incorporating both auditory and visual speech cues. With this hypothesis, we may then predict new auditory and/or visual features to combine in a given language, compensating for a phonetically ambiguous contrast.

Other evidence for the optimization of phonological systems through visually salient phonetic features may be drawn from commonly occurring contrasts in phonemic inventories. For example, the visual distinction between bilabial and coronal articulations, such as [m] and [n], may maximize the perceptual distinctiveness of these sounds, which may explain why they occur so frequently cross-linguistically, despite their relatively similar acoustic cues (Dohen, 2009). In a production and audio-visual perception study of the American English /ɔ/-/ɑ/ contrast, which is currently undergoing a merger in some dialects, it was found that a visual labial contrast was retained, despite the merging acoustics (Havenhill and Do, 2018). The authors themselves argued that visual cues may play a role in the shaping of phonological systems by inhibiting misperception of the speech signal in cases where two sounds are acoustically similar. Another noteworthy sound change in English involves /u/-fronting. /u/-fronting manifests itself acoustically as the raising of F2. As the term implies, it is generally assumed that /u/-fronting is the result of the fronting of the palatal constriction from an originally back position. However, a similar acoustic effect of F2 raising may also be a consequence of lip unrounding. Harrington *et al.* (2011) assessed the lingual and labial articulation of /u/ in Standard British English speakers and found that fronting indeed affects the position of the tongue and not the rounding of the lips. What all these sound change examples have in common is the retention of the labial articulation as a visual encoding of a phonological contrast. In the present study, we consider the retention

of the lip gesture for /r/ to be enhancement because it results in the differentiation of /r/ from /w/ in the visual domain.

V. CONCLUSION

By considering the audio-visual perception of the /r/-/w/ contrast in Anglo-English, we have shown that auditory perception in noise is not only enhanced by seeing the speaker's lips, but that visual speech cues provide reliable phonetic information which native speakers use in perception. Exposure to acoustic variation for /r/ may have resulted in perceptual ambiguity between /r/ and /w/ in England. Listeners must tolerate such a high degree of acoustic variability for /r/ that even canonical productions of [w] may be reconstructed as /r/ in perception. While auditory perception of the /r/-/w/ contrast may pose a challenge to Anglo-English listeners, visual cues from the speaker's lips allow them to disambiguate the contrast with an exceptionally high degree of accuracy. The Anglo-English /r/-/w/ contrast is therefore difficult to hear but easy to see. In proposing an Audio-Visual Enhancement Hypothesis, we suggest that languages select audio-visual properties of speech sounds which reinforce phonological contrasts. Finally, given the perceptual reliability of the visual cues relative to the auditory ones for Anglo-English /r/, we might predict a continued increase in the change from lingual to labial articulations. Predicting sound change should be undertaken with caution, and so we conclude that for now, the articulation of Anglo-English /r/ remains to be seen!

ACKNOWLEDGMENTS

We gratefully acknowledge the University of York, particularly Paul Foulkes, for data collection. The study received funding from the French National Research Agency (ANR-10-LABX-0083-LabEx EFL). We thank Emmanuel Ferragne for his help in this project. Finally, we thank the associate editor, Jianjing Kuang, and two anonymous reviewers for their valuable comments.

TABLE III. Test words and their lexical frequencies (in parentheses) in Zipf-scale from the SUBTLEX-UK database (van Heuven *et al.*, 2014).

Vowel	/r/	/w/	/l/
/iː/	*reeds* (3.36)	*weeds* (3.66)	*leads* (4.41)
	reek (2.61)	*week* (5.66)	*leak* (3.96)
/ɪ/	*rit* (2.20)	*wit* (3.67)	*lit* (4.02)
	rick (4.07)	*wick* (3.16)	*lick* (3.86)
/e/	*red* (5.41)	*wed* (3.16)	*led* (4.83)
	rent (4.53)	*went* (5.68)	*lent* (3.74)
/æ/	*rack* (3.81)	*whack* (3.74)	*lack* (4.62)
	rag (3.59)	*wag* (3.30)	*lag* (3.10)
/eɪ/	*rate* (4.85)	*wait* (5.37)	*late* (5.21)
	rake (3.40)	*wake* (5.12)	*lake* (4.49)
/aɪ/	*right* (6.36)	*white* (5.25)	*light* (5.28)
	rise (4.78)	*wise* (4.53)	*lies* (4.56)

APPENDIX: TEST WORDS

[1]An audiologist was consulted who recommended the set of questions we used to judge hearing performance, although she stressed that this technique could not replace clinical evaluations.

[2]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0012660 for additional methods details, additional results, an example catch trial, and for MP4 video files presenting example perception trials.

[3]We could not use different tokens of the same word to produce congruous audio-visual trials because no remaining tokens matched closely enough in word duration to create naturalistic materials. As congruous and incongruous audio-visual perception are not directly compared, this should not affect our results.

[4]<wh> words pronounced with initial /h/ were not included, and <wr> words were grouped with /r/-initial words.

Adachi, T., Akahane-Yamada, R., and Ueda, K. (**2006**). "Intelligibility of English phonemes in noise for native and non-native listeners," *Acoustical Science and Technology* **27**(5), 285–289, https://doi.org/10.1250/ast.27.285.

Aloufy, S., Lapidot, M., and Myslobodsky, M. (**1996**). "Differences in susceptibility to the "blending illusion" among native Hebrew and English speakers," *Brain and Language* **53**(1), 51–57, https://doi.org/10.1006/brln.1996.0036.

Bates, D., Mächler, M., Bolker, B., and 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.

Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (version 6.0.50)," http://www.praat.org (Last viewed December 10, 2019).

Boyce, S. E., and Espy-Wilson, C. Y. (**1997**). "Coarticulatory stability in American English /r/," *The Journal of the Acoustical Society of America* **101**(6), 3741–3753, https://doi.org/10.1121/1.418333.

Dalcher, C. V., Knight, R.-A., and Jones, M. J. (**2008**). "Cue switching in the perception of approximants: Evidence from two English dialects," University of Pennsylvania Working Papers in Linguistics **14**(2), 9.

Delattre, Pierre, C., and Freeman, D. C. (**1968**). "A dialect study of American r's by X-ray motion picture," *Linguistics* **6**(44), 29–68, https://doi.org/10.1515/ling.1968.6.44.29.

Diehl, R. L., and Kluender, K. R. (**1989**). "On the objects of speech perception," *Ecological Psychology* **1**(2), 121–144, https://doi.org/10.1207/s15326969eco0102_2.

Docherty, G., and Foulkes, P. (**2001**). "Variability in (r) production – Instrumental perspectives," in 'R-Atics: Sociolinguistic, phonetic and phonological characteristics of /r/, edited by H. Van de Velde and R. van Hout (Université Libre de Bruxelles, Brussels, Belgium), pp. 173–184.

Dohen, M. (**2009**). "Speech through the ear, the eye, the mouth and the hand," in Multimodal Signals: Cognitive and Algorithmic Issues, edited by A. Esposito, A. Hussain, M. Marinaro, and R. Martone, Lecture Notes in Computer Science (Springer, Berlin/Heidelberg, Germany), pp. 24–39.

Foulkes, P., and Docherty, G. J. (**2000**). "Another chapter in the story of /r/: 'Labiodental' variants in British English," *Journal of Sociolinguistics* **4**(1), 30–59, https://doi.org/10.1111/1467-9481.00102.

Gimson, A. C. (**1980**). An Introduction to the Pronunciation of English, 3rd ed. (Arnold, London).

Grant, K. W., Walden, B. E., and Seitz, P.-F. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," *The Journal of the Acoustical Society of America* **103**(5), 2677–2690, https://doi.org/10.1121/1.422788.

Harrington, J., Kleber, F., and Reubold, U. (**2011**). "The contributions of the lips and the tongue to the diachronic fronting of high back vowels in Standard Southern British English," *Journal of the International Phonetic Association* **41**(2), 137–156, https://doi.org/10.1017/S0025100310000265.

Havenhill, J., and Do, Y. (**2018**). "Visual speech perception cues constrain patterns of articulatory variation and sound change," *Frontiers in Psychology* **9**(728), https://doi.org/10.3389/fpsyg.2018.00728.

Heyne, M., Wang, X., Derrick, D., Dorreen, K., and Watson, K. (**2018**). "The articulation of /ɹ/ in New Zealand English," *Journal of the International Phonetic Association* 1–23, https://doi.org/10.1017/S0025100318000324.

Hornsby, D. (**2014**). Linguistics: A Complete Introduction (Teach Yourself, London).

Irwin, J. R., Frost, S. J., Mencl, W. E., Chen, H., and Fowler, C. A. (**2011**). "Functional activation for imitation of seen and heard speech," *Journal of Neurolinguistics* **24**(6), 611–618, https://doi.org/10.1016/j.jneuroling.2011.05.001.

Jones, D. (**1972**). An Outline of English Phonetics, 9 ed. (Cambridge University Press).

Jongman, A., Wang, Y., and Kim, H., B. (**2003**). "Contributions of Semantic and Facial Information to Perception of Nonsibilant Fricatives," *Journal of Speech, Language, and Hearing Research* **46**(6), 1367–1377, https://doi.org/10.1044/1092-4388(2003/106).

King, H., and Ferragne, E. (**2020**a). "Labiodentals /r/ here to stay: Deep learning shows us why," *Anglophonia. French Journal of English Linguistics* (30), https://doi.org/10.4000/anglophonia.3424.

King, H., and Ferragne, E. (**2020**b). "Loose lips and tongue tips: The central role of the /r/-typical labial gesture in Anglo-English," *Journal of Phonetics* **80**, 100978, https://doi.org/10.1016/j.wocn.2020.100978.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest Package: Tests in Linear Mixed Effects Models," *Journal of Statistical Software* **82**(13), https://doi.org/10.18637/jss.v082.i13.

Lalonde, K., and Werner, L. A. (**2019**). "Infants and adults use visual cues to improve detection and discrimination of speech in noise," *Journal of Speech, Language, and Hearing Research* **62**(10), 3860–3875, https://doi.org/10.1044/2019_JSLHR-H-19-0106.

Lawson, E., Scobbie, J. M., and Stuart-Smith, J. (**2011**). "The social stratification of tongue shape for postvocalic /r/ in Scottish English," *Journal of Sociolinguistics* **15**(2), 256–268, https://doi.org/10.1111/j.1467-9841.2011.00464.x.

Lawson, E., Scobbie, J. M., and Stuart-Smith, J. (**2014**). "A Socio-Articulatory Study of Scottish Rhoticity," in Sociolinguistics in Scotland, edited by R. Lawson (Palgrave Macmillan UK, London), pp. 53–78, https://doi.org/10.1057/9781137034717_4.

Lawson, E., Stuart-Smith, J., and Scobbie, J. M. (**2018**). "The role of gesture delay in coda /r/ weakening: An articulatory, auditory and acoustic study," *The Journal of the Acoustical Society of America* **143**(3), 1646–1657, https://doi.org/10.1121/1.5027833.

Lee, A. (**2000**). "Virtual Dub (version 1.10.4)," http://www.virtualdub.org (Last viewed December 10.2019.

Lenth, R. V. (**2021**). "emmeans: Estimated marginal means, aka least-squares means," R package version 1.5.5-1, https://CRAN.R-project.org/package=emmeans (Last viewed July 27, 2021).

Llamas, C. (**1998**). "Language variation and innovation in Middlesborough: A pilot study," Leeds Working Papers in Linguistics and Phonetics **6**, 97–114.

Macmillan, N. A., and Creelman, C. D. (**2005**). Detection Theory: A User's Guide, second ed. (Lawrence Erlbaum Associates, Mahwah, NJ, USA).

Marsden, S. (**2006**). "A sociophonetic study of labiodental /r/ in Leeds," Leeds Working Papers in Linguistics and Phonetics (11), 153–172.

Massaro, D. W. (**1987**). Speech Perception By Ear and Eye: A Paradigm for Psychological Inquiry (Lawrence Erlbaum Associates, Hillsdale, NK).

Massaro, D. W. (**1998**). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle (MIT press, Cambridge, MA).

Mattheyses, W., and Verhelst, W. (**2015**). "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication* **66**, 182–217, https://doi.org/10.1016/j.specom.2014.11.001.

McCloy, D. (**2013**). "Mix speech with noise," Praat script licensed under the GNU General Public Licence v3.0, https://github.com/drammock/praat-semiauto/blob/master/MixSpeechNoise.praat (Last viewed December 10, 2019).

McGuire, G., and Babel, M. (**2012**). "A cross-modal account for synchronic and diachronic patterns of /f/ and /θ/ in English," *Laboratory Phonology* **3**(2), 251–272, https://doi.org/10.1515/lp-2012-0014.

McGurk, H., and Macdonald, J. (**1976**). "Hearing lips and seeing voices," *Nature* **264**(5588), 746–748, https://doi.org/10.1038/264746a0.

Mielke, J., Baker, A., and Archangeli, D. (**2016**). "Individual-level contact limits phonological complexity: Evidence from bunched and retroflex /ɹ/," *Language* **92**(1), 101–140, https://doi.org/10.1353/lan.2016.0019.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (**1957**). "Acoustic cues for the perception of initial /w, j, r, l/ in English," *Word* **13**(1), 24–43, https://doi.org/10.1080/00437956.1957.11659626.

Ohala, J. J. (**1981**). "The listener as a source of sound change," in Papers from the Parasession on Language and Behavior, edited by C. Masek and R. Hendrick, Chicago Linguistic Society, Chicago, IL, USA, pp. 178–203.

Ohala, J. J. (**1996**). "Speech perception is hearing sounds, not tongues," *Journal of the Acoustical Society of America* **99**(3), 1718–1725, https://doi.org/10.1121/1.414696.

Peirce, J. W. (**2007**). "PsychoPy – Psychophysics software in Python," *Journal of Neuroscience Methods* **162**(1), 8–13, https://doi.org/10.1016/j.jneumeth.2006.11.017.

R Core Team (**2018**). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (**2007**). "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex* **17**(5), 1147–1153, https://doi.org/10.1093/cercor/bhl024.

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (**2012**). "NIH Image to ImageJ: 25 years of image analysis," *Nature Methods* **9**(7), 671–675, https://doi.org/10.1038/nmeth.2089.

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (**2015**). "afex: Analysis of factorial experiments," R package version 0.13–145, http://CRAN.R-project.org/package=afex (Last viewed July 27, 2020).

Stevens, K. N. (**1998**). Acoustic Phonetics, **30** (MIT press, Cambridge, MA, USA).

Sumby, W., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," *The Journal of the Acoustical Society of America* **26**(2), 212–215, https://doi.org/10.1121/1.1907309.

Summerfield, Q., Bruce, V., Cowey, A., Ellis, A. W., and Perrett, D. (**1992**). "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **335**(1273), 71–78, https://doi.org/10.1098/rstb.1992.0009.

Tiede, M. K., Boyce, S. E., Holland, C. K., and Choe, K. A. (**2004**). "A new taxonomy of American English /r/ using MRI and ultrasound," *The Journal of the Acoustical Society of America* **115**(5), 2633–2634, https://doi.org/10.1121/1.4784878.

Trudgill, P. (**1974**). The Social Differentiation of English in Norwich (Cambridge University Press).

Van Engen, K. J., Xie, Z., and Chandrasekaran, B. (**2017**). "Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect," *Attention, Perception, & Psychophysics* **79**(2), 396–403, https://doi.org/10.3758/s13414-016-1238-9.

van Heuven, W. J. B., Mandera, P., Keuleers, E., and Brysbaert, M. (**2014**). "Subtlex-UK: A New and Improved Word Frequency Database for British English," *Quarterly Journal of Experimental Psychology* **67**(6), 1176–1190, https://doi.org/10.1080/17470218.2013.850521.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., and Li, X. (**1996**). "Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition," *The Journal of the Acoustical Society of America* **100**(2), 1153–1162, https://doi.org/10.1121/1.416300.

Wells, J. C. (**1982**). Accents of English (Cambridge University Press, Cambridge).

Werker, J. F., Frost, P. E., and McGurk, H. (**1992**). "La langue et les lèvres: Cross-language influences on bimodal speech perception.," *Canadian Journal of Psychology* **46**(4), 551–568, https://doi.org/10.1037/h0084331.

Williams, A., and Kerswill, P. (**1999**). "Dialect levelling: Change and continuity in Milton Keynes, Reading and Hull," in Urban voices: Accent studies in the British Isles, edited by P. Foulkes and G. J. Docherty (Arnold, London, UK), pp. 141–162.

Zhou, X., Espy-Wilson, C. Y., Boyce, S. E., Tiede, M. K., Holland, C., and Choe, A. (**2008**). "A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/," *The Journal of the Acoustical Society of America* **123**(6), 4466–4481, https://doi.org/10.1121/1.2902168.