WHAT YOU SEE IS (NOT) WHAT YOU HEAR: THE ROLE OF SOCIAL EXPECTATIONS ON SPEECH INTELLIGIBILITY IN A BRITISH CONTEXT

Ruohan Guo, Fengjie Cheng, Bronwen G. Evans

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
ucjurgu@ucl.ac.uk, fengjie.cheng.20@ucl.ac.uk, bronwen.evans@ucl.ac.uk

ABSTRACT

Several studies have found that giving listeners information about the talker affects speech comprehension. This study investigated whether exposing listeners to regionally-meaningful cues has similar effects on the intelligibility of native accents in a British context. Two groups of participants, who differed in their familiarity with Glaswegian (GE) and Standard Southern British English (SSBE), transcribed GE-accented and SSBE-accented sentences at three SNRs (+3 dB, 0 dB, -3 dB) while they were exposed to visual cues indicating (1) a congruent accent region, (2) an incongruent accent region or (3) silhouetted cues representing no specific region. As expected, noise level and accent familiarity affected performance. However, contrary to previous findings, congruence between regional cues and spoken accent did not affect speech intelligibility in either listener group and did not interact with noise level or familiarity. This null result raises questions about the generalizability of the effects of social expectations on speech perception.

Keywords: sociophonetics, speech perception, exemplar-models, speech intelligibility in noise.

1. INTRODUCTION

A major area of interest in sociophonetics is how non-linguistic information is integrated into speech processing. There is evidence that perceived talker characteristics can influence listeners behaviour in perceptual tasks. For example, Niedzielski [1] found that listeners from Detroit categorised /aʊ/ differently depending on the description they were given of the speaker's regional origin. Niedzielski suggested that listeners' phoneme categorization decisions reflected awareness of phonetic features that they stereotypically attributed to certain regional accents. Subsequent studies have also found that listeners use social information to guide speech processing and that even implicit cues such as stuffed toys [2] or newspapers [3] can affect perceptual performance.

The observed effects have led to an interest in usage-based and exemplar approaches to speech perception [4, 5]. These argue that, based on observations or stereotypes, listeners form and store in memory associations between linguistic instances and socioindexical information (about speakers or contexts) and exploit these when processing newly-encountered speech.

In recent years, a growing body of research has found that exposure to socially-meaningful cues can also affect the processing of utterances beyond the phoneme level. In a transcription task, Babel and Russell [6] reported that listeners who were shown a photo of a Chinese speaker transcribed L1-English sentences less accurately than listeners who were shown a photo of a white Caucasian. McGowan [7] complemented this finding, demonstrating that Chinese-accented sentences in noise were more intelligible to American listeners when paired with an Asian face than with a Caucasian face. In contrast, misleading social expectations (i.e., incongruent pairing of face and voice) led to activation of inappropriate associations and hindered word recognition [6]. These findings are consistent with an exemplar-based account in which congruent speaker information activates appropriate associations in listeners' memory and facilitates recognition of utterances (see also [8]).

The reported effects of the role of social expectations in speech intelligibility have not only contributed to theories of speech perception but have also helped us better understand how stereotypes can disadvantage a speaker in communication (e.g., [9, 10]). However, most existing studies have been conducted in either New Zealand (e.g., [2]) or North America, with those in North America focussing primarily on the effect of perceived talker ethnicity on the intelligibility of standard native and/or foreign accents (e.g., [6, 7, 8]). It is unclear if this effect generalizes to different geographical contexts and to the perception of regional identity.

This study contributes to existing knowledge by testing how exposure to regionally-meaningful cues affects speech intelligibility in a British context. In two online experiments, we measured listeners' sentence transcription accuracy in two British English accents, Glaswegian English (GE) and Standard Southern British English (SSBE), whilst they were exposed to visual cues indicating; (1) a congruent accent region, (2) an incongruent accent region or (3) silhouetted cues representing no specific region (control condition). The first experiment involved

listeners from Glasgow who were familiar with both accents, while the second recruited listeners from the south of England, highly familiar with SSBE but not GE. Based on previous findings [7], we hypothesised that listeners would transcribe sentences in noise more accurately in the congruent condition than in the control condition and least accurately in the incongruent condition, but that this would be modulated by familiarity such that greater familiarity would lead to a larger congruence effect. In addition, stimuli were presented at three signal-to-noise ratios (SNR) (+3 dB, 0 dB, -3 dB). We expected that transcription accuracy would decrease with noise [11, 12] but of interest here, was whether effects of regional cues would be modulated by listening conditions.

2. EXPERIMENT ONE

2.1. Methods

2.1.1 Participants

Twenty-nine monolingual native English speakers aged 18-50 yrs were recruited via Prolific (www.prolific.co) [13]. All participants had lived in Glasgow for at least 3 months and reported being very familiar with both GE and SSBE. Twenty-six self-identified as having a GE accent. The participants reported no history of any speech, hearing, or language disorders.

2.1.2 Auditory stimuli

The stimuli were 108 sentences from the Institute of Electrical and Electronics Engineers (IEEE) sentence list [14], taken from a previous study [15]. Sentences were recorded at 44.1 kHz, 16-bit resolution, then segmented in Praat [16] and downsampled to 22,050 kHz. Amplitude was normalized to 70 dB. The sentences were produced by two GE and two SSBE speakers, all male, monolingual English speakers, aged 26-46 yrs, who identified as middle class.

Half of the sentences (N=54) were randomly assigned to the GE accent condition, and the other half (N=54) to the SSBE accent condition. For each accent, a third of the sentences (N=18) were paired with images representing a congruent accent region; the second set (N=18) was paired with images representing an accent region incongruent with the sentence accent; the final set (N=18) was paired with neutral images representing no particular place.

2.1.3 Visual stimuli

Visual stimuli were chosen to represent Glasgow (GE) and London (SSBE). The images included iconic cartoon-style images of the city skylines, football club mascots and a culturally salient figure representing each city (London – Beefeater; Glasgow – piper in traditional Scottish dress). For the neutral background, a silhouetted cityscape and two stick figures were selected. The icons were arranged in the centre of the screen and around the transcription input box to ensure that participants would notice them.

A naïve assessor confirmed that the images represented each city or, for the neutral background, no specific place. Post-experiment questionnaires showed that participants also linked the visual cues with each city or a neutral place.

2.1.4 Procedure

The experiment was designed and conducted online in Gorilla Experiment Builder (www.gorilla.sc) [17].

Before beginning the experiment, participants were told that their performance would be used to assess whether the experimental setup was suitable for children and that, as such, they would see images as well as hear a sentence. This was to encourage them to pay attention to the images. However, consistent with some previous studies [2], these regional cues were not attributed to the speaker.

For the transcription task, participants were instructed that they would hear each sentence only once and were asked to type in anything they heard. Participants completed 108 trials in a randomised order. Sentences were played at a self-adjusted comfortable listening level. The experiment was self-paced. Participants clicked a button to start each trial: the screen then simultaneously displayed the images and a play button which participants clicked to hear the sentence.

Finally, participants completed a post-experiment questionnaire. They rated the images and gave basic information about their language background and experience with GE and SSBE.

2.1.5 Data analysis

Data was pre-processed to remove punctuation. A custom-built Excel script was used to automatically calculate the proportion of words correctly transcribed in each sentence. Manual spot checks were completed afterwards, e.g., to correct typos ("squireel" for "squirrel").

Data was then analysed using R [18]. A binomial generalized linear mixed model was fitted. The proportion of correctly transcribed words (bounded between 0 and 1) was used as the dependent variable. Based on the predictions, the model included *Congruence* (Congruent, Control, Incongruent), *Noise Level* (-3 dB, 0 dB, 3 dB) and *Sentence Accent*

(GE, SSBE) as fixed effects, as well as the interactions between them. Random intercepts for participant and sentence were also included. The *mixed*() function in the *afex* package was used to compute the p-values of all fixed effects using likelihood ratio tests [19]. Follow-up tests were conducted with the *emmeans* package [20] and corrected pairwise comparisons with Holm correction.

2.2 Results

As expected, the likelihood ratio tests for the full model demonstrated a significant main effect of *Noise Level*, $\chi^2(2) = 50.70$, $p<0.001$, but a non-significant effect of *Sentence Accent*, χ^2 (1) = 1.13, p =0.28. Overall, GE listeners performed more poorly at higher noise levels but performed equally well with GE and SSBE (see Fig. 1). Follow-up tests showed that, on average, listeners transcribed sentences more accurately at +3 dB than at 0 dB (z=2.16, p =0.03) and -3 dB (z=7.757, p <0.001). Additionally, transcription accuracy at 0 dB was significantly higher than at -3 dB (z=5.624, p <0.001).

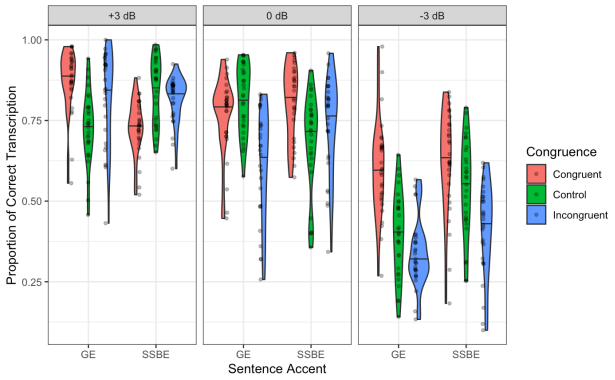

Figure 1: Proportion of correctly transcribed words for GE listeners in each accent (GE, SSBE) split by noise level and congruence condition.

However, there was a non-significant effect of *Congruence*, χ^2 (2) = 5.96, p = 0.051, and no significant interactions. Although the data, as displayed in Fig. 1, showed a pattern matching our prediction (i.e., *Congruent* > *Control* > *Incongruent*) at the highest SNR, -3 dB, this was not statistically significant.

2.3 Summary

Contrary to our expectations, we failed to find clear evidence for the effect of regional cues on speech intelligibility under adverse listening conditions.

Consistent with previous findings, GE listeners transcribed sentences less accurately at higher noise levels [11] and performed equally well with both GE and SSBE [12]. This likely reflects GE listeners' high level of familiarity with both these accents. However, the fact that GE listeners had high levels of experience with and, therefore, presumably detailed phonetic representations of both accents did not lead to any reliable congruence effects. Although there was some indication that sentences were identified more accurately when presented with congruent visual cues, particularly at higher noise levels, this was not reliable. This contrasts with previous studies [7, 8] which have found that listeners are able to benefit from matching social information.

One possibility is that these GE listeners, highly familiar with both GE and SSBE, have detailed enough phonetic representations and so either do not recruit non-linguistic information to support word recognition or are better able to ignore incongruent visual cues. In Experiment 2, we investigate whether listeners from the South of England who are highly familiar with SSBE but not GE perform similarly.

3. EXPERIMENT TWO

3.1 Methods

3.1.1 Participants

Twenty-four monolingual native English speakers aged 18-50 yrs were recruited via Prolific [13]. They had been born and raised in London and South East England and described their accent as SSBE. In the post-experiment questionnaire, all reported that they were highly familiar with SSBE but not GE, that they had not spent any significant amount of time in Scotland and did not have regular contact with GE speakers (e.g., family members, partner). Participants reported no history of any speech, hearing, or language disorders.

3.1.2 Stimuli and procedure

The design and procedure were the same as in Expt 1.

3.2 Results

We performed a binomial generalized linear mixed model on the SSBE group data, with *Congruence*, *Noise Level* and *Sentence Accent* as fixed effects and interaction terms between them. The model included random intercepts for both participant and sentence.

The results of likelihood ratio tests for the mixed effects model showed that, as predicted, there was a significant main effect of *Noise Level*, χ^2 (2) = 52.56, p <0.001. As in Expt 1, intelligibility decreased as the noise level increased (see Fig 2). Follow-up tests showed that listeners recognized utterances more accurately at +3 dB than at 0 dB (z=2.168, p =0.03) and -3 dB (z=7.93, p <0.001). Transcription accuracy

was also significantly higher at 0 dB than at -3 dB (z=5.78, p <0.001) where participants performed particularly poorly for GE in all conditions. There was also a significant main effect of *Sentence Accent*, χ^2 (1) = 28.66, p <0.001: SSBE listeners performed better with SSBE than GE (Fig 2). Follow-up tests showed that GE-accent sentences were less accurately transcribed than SSBE-accent sentences in noise (z=-5.72, p <0.001).

However, there was no significant main effect of *Congruence*, χ^2 (2) = 4.27, p =0.118, and no interaction terms were statistically significant. Although there was a trend for listeners to perform better with GE sentences in the congruent than incongruent condition, particularly at higher noise levels (Fig 2), SSBE listeners' transcription accuracy in noise was not reliably affected by the match or mismatch between the visual regional cues and spoken accent.

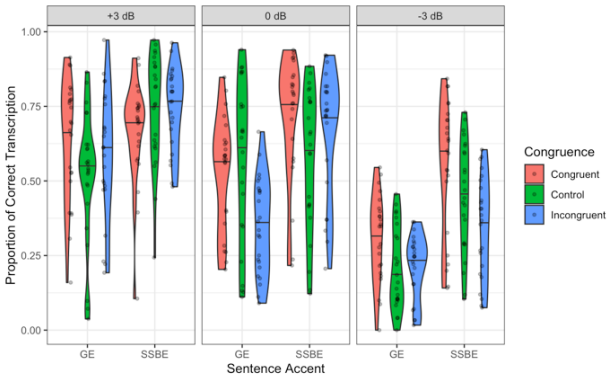

Figure 2: Proportion of correctly transcribed words for SSBE listeners in each accent (GE, SSBE) split by noise level and congruence condition.

3.3 Summary

The results of this experiment are similar to those reported in Expt 1. Listeners performed more poorly as the noise level increased. As predicted, SSBE listeners, highly familiar with SSBE but not GE, performed more poorly with GE than SSBE (cf. [12]), but there was no evidence for a congruence effect: there was no reliable difference in performance in the congruent, control or incongruent conditions for either accent at any noise level. This indicates that the effect of regional cues on sentence transcription was not modulated by familiarity with a given accent.

4. GENERAL DISCUSSION

This study explored how exposure to regionally-meaningful visual cues affects speech intelligibility in a British context. Although both experiments showed expected effects of noise level and familiarity, there was no evidence that transcription accuracy was affected by visual cues to regional accents. This null result is inconsistent with earlier studies showing that alignment between social cues and spoken accents increases sentence intelligibility in noise (e.g., [7]).

The null effect may be attributed to the subtle visual prompts used in this study. Earlier studies have primarily examined how perceived ethnicity affects speech intelligibility of standard L1- or foreign (L2-) accents in North America (e.g., [6, 7, 8]). In these experiments, listeners are typically led to believe that the speech they hear comes from the talker, depicted in a photo [6, 7] or explicitly described [8]. This may lead to a closer association of social cues and auditory information than is possible in an experiment like the present one, where visual stimuli were cartoon-like icons, and may in turn lead to a congruence effect.

Another explanation is that although regional cues may have guided social expectations, at least some of the listeners did not have sufficient experience with the accents and instead used stereotypical or simplified structures that did not match the speech they actually heard (cf. [7]). This could explain why SSBE listeners showed no effect of congruence for GE, with which they were unfamiliar. Nevertheless, this cannot explain why listeners did not show the expected priming effect for accents with which they were highly familiar. Although it could be the case that listeners discard mismatching visual information when they can reliably access the auditory signal (i.e., at easier SNRs), it is still unclear why visual cues appeared to have no effect on speech intelligibility at the higher noise levels (0dB, -3dB), similar to those used in previous studies (e.g., [7]).

Some studies have found a significant effect of regional cues on vowel categorization, mainly in New Zealand (e.g., [2] and more recently, [21]). Researchers in Britain have attempted to replicate these effects but have failed to find convincing evidence that exposure to regional cues affects phoneme matching [22, 23]. The current study is not directly comparable to these previous experiments due to methodological differences (sentence transcription *vs* phoneme matching). However, the lack of a significant effect in these different studies may indicate either that the effect of social cues on speech processing is more limited than expected or that listeners do use this information but these paradigms are not always able to detect this [cf. 22, 23]. Further studies using a variety of approaches are needed to investigate the role of these and other factors, in particular sociocultural context, on the role of social expectations in speech processing.

5. REFERENCES

[1] Niedzielski, N. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1), 62–85.

[2] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48(4), 865–892.

[3] Portes, C., German, J. S. 2019. Implicit effects of regional cues on the interpretation of intonation by Corsican French listeners. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10, 1-26.

[4] Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34, 485–499.

[5] Docherty, G. J., Foulkes, P. 2014. An evaluation of usage-based approaches to the modelling of sociophonetic variability. *Lingua* 142, 42-56.

[6] Babel, M., Russell, J. 2015. Expectations and speech intelligibility. *The Journal of the Acoustical Society of America 137*(5), 2823-2833.

[7] McGowan, K. B. 2015. Social expectation improves speech perception in noise. *Language and Speech 58*(4), 502-521.

[8] Vaughn, C. R. 2019. Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America* 145(5), 3218-3232.

[9] Kang, O., Rubin, D. L. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28, 441–456.

[10] Kutlu, E., Tiv, M., Wulff, S., Titone, D. 2022. Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications* 7(1), 1-16.

[11] Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., Abrams, H. B. 2006. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics* 27(3), 465-485.

[12] Adank, P., Evans, B.G., Stuart-Smith, J., Scott, S.K. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance* 35 (2), 520-529.

[13] Palan, S., Schitter, C. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17, 22-27.

[14] Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust* 17, 227–246.

[15] Evans, B.G., Adank, P. 2008. Differences in the time-course of accent adaptation: a comparison of adaptation to foreign-accented and unfamiliar regionally-accented speech. *Journal of the Acoustical Society of America* 123(5), 3703-3703.

[16] Boersma, P. & Weenink, D. 2021. Praat: Doing phonetics by computer (Version 6.2.14)

[17] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52(1), 388-407.

[18] R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[19] Singmann, H., Bolker, B., Westfall, J. 2015. afex: Analysis of factorial experiments. http://CRAN.R-project.org/package=afex.

[20] Lenth. R., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H. & Singmann, H. 2022. *emmeans: Estimated Marginal Means, aka Least-Squares Means.* https://github.com/rvlenth/emmeans.

[21] Hurring, G., Hay, J., Drager, K., Podlubny, R., Manhire, L., Ellis, A. 2022. Social priming in speech perception: revisiting Kangaroo/Kiwi priming in New Zealand English. *Brain Sciences* 12(6), 684.

[22] Lawrence, D. 2015. Limited evidence for social priming in the perception of the BATH and STRUT vowels. *Proc. 18th ICPhS* Glasgow.

[23] Juskan, M. 2018. *Sound Change, Priming, Salience: Producing and Perceiving Variation in Liverpool English*. Language Science Press.

[24] Hanulíková, A. 2021. Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PLoS ONE* 16(10), e0259230.