This is an Accepted Manuscript of an article published by Taylor & Francis in Language, Cognition and Neuroscience, available online:
http://www.tandfonline.com/doi/full/10.1080/23273798.2019.1582787

Individual differences in the link between perception and production and the mechanisms of phonetic imitation

Donghyun Kim[a,c]*, Meghan Clayards[a,b]

[a] *Department of Linguistics, McGill University, Montreal, Canada*

[b] *School of Communication Sciences and Disorders, McGill University, Montreal, Canada*

[c] *Department of Psychology, University of Exeter, Exeter, UK*

*Corresponding author: d.kim2@exeter.ac.uk

Individual differences in the link between perception and production and the mechanisms of phonetic imitation

This study investigates the relationship between speech perception and production using explicit phonetic imitation. We used manipulated natural vowel (*head-had*) stimuli varying in spectral quality and duration in both perception and production tasks to explore the perception-production link in a direct and controlled way. We examined (1) whether individual listeners' perceptual cue weights are related to their patterns of phonetic imitation and (2) phonological and perceptual constraints underlying phonetic imitation. Results showed that better perceptual abilities (i.e., larger cue weights) were related to better imitation of vowel duration. Furthermore, imitation of vowel spectral quality was mediated by contrast maintenance while vowel duration was not. Overall, vowel duration was better imitated despite being the less important cue perceptually. These results suggest that speech perception and production are indeed linked at the individual level, and both linguistic and perceptual-cognitive factors play a role in this process.

Keywords: phonetic imitation; perception-production link; cue weighting; individual differences; vowel contrasts; speech perception; speech production

1 Introduction

Speech communication is two-way. Language users perceive and produce speech sounds, and their communication succeeds when both members understand each other's intended message. The achievement of sufficient equivalence between the forms of speaking and listening is called *parity* and it is required for mutual understanding and successful communication (Gambi & Pickering, 2013; Garrod & Pickering, 2004; Liberman & Whalen, 2000). In speech communication, language users constantly change roles and thus a language user who was once a speaker is subsequently a listener and vice versa, mutually influencing each other's utterances. In terms of cognitive mechanisms of these processes, interlocutors interactively align their representations used in production and comprehension (Garrod & Pickering, 2004). This shows that the

link between speech perception and production is inherent in our use of spoken language. Furthermore, understanding the nature of the relationship as to how and to what extent representations are shared or separate between perception and production may sheds light on speech communication more generally.

Narrowing down the nature of the link is a harder task, however. In the present study, we focus on the link at the level of an individual language user. As reviewed below, previous studies have found mixed evidence for a link in the performance of individuals on perception and production tasks, and these differences may be due to divergent methodologies across the two modalities. The current study addresses these issues by using a phonetic imitation paradigm of manipulated stimuli (Babel, 2012; Goldinger, 1998; Mitterer & Ernestus, 2008; Nielsen, 2011, 2014; Nye & Fowler, 2003; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013; Shockley, Sabadini, & Fowler, 2004; Yu, Abrego-Collier, & Sonderegger, 2013; Zellou, Scarborough, & Nielsen, 2013, 2016, among many others) to examine the acoustic-phonetic information individuals use in both perception and production in a more constrained way.

1.1 Perception-production links

While the link between speech perception and production has largely been investigated at a group level, for example, by examining second language learners or making cross-linguistic comparisons (Rochet, 1995), a growing body of research has addressed the issue of the relationship between perception and production for individuals (e.g., Beddor, Coetzee, Styler, McGowan, & Boland, 2018; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Brunner, Ghosh, Hoole, Matthies, Tiede, & Perkell, 2011; Frieda, Walley, Flege, & Sloane, 2000; Newman, 2003; Perkell, Guenther, Lane, Matthies, Stockmann, Tiede, & Zandipour, 2004; Shultz, Francis, & Llanos, 2012) but evidence of the link between the two processes has been mixed depending on

experimental paradigms and measures.

One set of studies has examined whether individual participants' performance in perception is related to production accuracy. For instance, the results from Perkell et al. (2004) and Brunner et al. (2011) showed that participants who better discriminated vowel and sibilant contrasts in perception also produced the same contrasts with less overlap in production. Bradlow et al. (1997) also found that Japanese speakers who received perceptual training on the English /ɹ/-/l/ distinction also improved production of the contrast. Together, these studies have suggested that more accurate perception of speech sounds is related to more accurate production.

Another set of studies have approached the perception-production link in terms of best exemplars for categories across the modalities. For example, Newman (2003) tested production and perception prototypes for /pæ/ and /ʃæ/ and found that listeners' perceptual prototypes were significantly (though weakly) correlated with their productions. That is, participants whose perceptual prototype had longer VOT or higher peak spectral frequency also tended to produce these consonants with longer VOT and higher peak spectral frequency. Fox (1982) found that estimates of perceptual distances among vowel stimuli correlated with acoustic measures of the corner vowels at an individual level. Frieda et al. (2000) also demonstrated that for an individual participant the best exemplars of the vowel /i/ in a perceptual task were associated with their own hyperarticulated productions although this pattern was observed only for a subset of participants. These studies overall point towards a link in which individuals who prefer more hyperarticulated tokens also produce more hyperarticulated tokens. However, other studies have failed to find such a link between perception and production categories. For example, in Bailey and Haggard (1980), average VOTs for voiced and voiceless consonants in production were not correlated with perceptual category boundaries for a /g/-/k/ continuum in perception.

Another question is whether variation in use of acoustic-phonetic properties is linked across perception and production. For example, Shultz et al. (2012) examined individual participants' weighting of two cues (VOT and onset f0) to stop consonant voicing in English. They found that individual listeners' perceptual weights for these cues were not significantly correlated with the degree of their cue use in production. However, recent work on sound change in Afrikaans stop voicing reported that participants who showed heavy reliance on a voicing cue in perception (relative to f0) also tended to show more pre-voicing in production (Coetzee, Beddor, Shedden, Styler, & Wissing, 2018). Similarly, Beddor et al. (2018) presented findings from a study of vowel nasalisation due to coarticulation in English and found that participants who were more sensitive to nasalisation cues on the vowel in perception (using an eye-tracking paradigm), produced more coarticulatorily nasalised vowels. Zellou (2017) used a paired discrimination task to assess sensitivity to coarticulatory nasalisation in English and found that listeners who had more veridical perception of vowel nasality in nasal contexts (i.e., less attribution of nasality to the coarticulatory source) were those that produced less coarticulatorily nasalised vowels. These studies therefore provide some support for the idea that listeners interpret the acoustic dimensions of speech with reference to their own productions. It should be noted however that the relationships are somewhat weak.

One reason for the comparatively weak results in paradigms examining cue use, and in particular cue weights just reviewed may be disparate perception and production tasks. In particular, it should be noted that perception tasks generally rely on ambiguous or conflicting stimuli to understand how much an individual listener uses relevant cues in phonetic categorisation. In contrast, production tasks are not generally constrained and speakers tend to produce highly unambiguous and sometimes hyperarticulated tokens, leading to minimal differences among those tokens. Thus, the tasks that have

been used put different demands on the participant (e.g., Shultz et al., 2012), which in part makes an understanding of the true nature of the perception-production relationship difficult. In fact, Zellou (2017) observed that the nature of the link at the individual level may vary depending on different task demands, with no link found between production and a more meta-linguistic nasality rating task.

In sum, studies exploring the relationship between speech perception and production at an individual level have yielded mixed results. Some evidence of the link was found when examining measures of performance accuracy and prototypes, and production and perception measures of coarticulation, while evidence of the link was less conclusive when examining individuals' weighting of acoustic-phonetic cues. This might indicate that the nature of the link is restricted to performance accuracy or preferred degree of hyperarticulation (i.e., prototypes) and does not extend to individual choices about cues. The conflicting findings in prior research might also be in part due to disparate experimental methodologies between perception and production. Crucially, these studies explored the relationship by comparing different perception and production tasks (e.g., read a word list vs. chose which of two categories a stimulus belongs to) which might have different task demands (to be maximally unambiguous vs. to resolve ambiguity). Thus, paradigms in which the demands of the speech perception and production tasks are more parallel and also more tightly linked may tell us more about how the two processes are related. In the present study, we first examine whether listeners' cue weighting strategies in a two-alternative forced choice identification task are related to their production patterns at baseline, which is similar to production and perception tasks in previous studies. Then we further explore whether the two modalities are more tightly linked in an imitation task, in which participants are explicitly asked to imitate ambiguous stimuli as well as unambiguous stimuli.

1.2 Imitation of speech

Phonetic imitation, also known as speech imitation, phonetic accommodation, speech alignment, or phonetic convergence, is the process where talkers adjust acoustic-phonetic characteristics of their speech to those of interlocutors (Babel, 2012; Babel, McGuire, Walters, & Nicholls, 2014; Goldinger, 1998; Nielsen, 2011, 2014; Nye & Fowler, 2003; Pardo et al., 2013; Shockley et al., 2004; Walker & Campbell-Kibler, 2015; Yu et al., 2013; Zellou et al., 2016, among many others). In phonetic imitation, it has been suggested that three major processes are generally involved, namely perception of phonetic properties of stimuli, encoding and storage in memory, and reproduction of the input stimuli through motor control (Flege & Eefting, 1988). Because of the involvement of these three components phonetic imitation can inform us about the link between perception and production processes in speech communication, and also the use of phonetic details during speech perception and production (Babel, 2012; Nielsen, 2011, 2014; Zellou et al., 2016).

Studies that examined imitation of speech sounds have observed that phonetic details in the speech signal can be preserved and reflected in phonetic imitation (Goldinger, 1998; Nielsen, 2011). These studies have hypothesized that detailed information in the speech signal persists as an exemplar in memory and can be used in subsequent productions. Some reported examples of the phonetic details that are manifested in imitation include VOT (Nielsen, 2011, 2014; Shockley et al., 2004; Yu et al., 2013), f0 (Babel & Bulatov, 2012; Postma-Nilsenová, & Postma, 2013), vowel quality (Babel, 2012; Dufour, & Nguyen, 2013; Tilsen, 2009; Walker & Campbell-Kibler, 2015), vowel duration (Delvaux & Soquet, 2007), and vowel nasalisation (Zellou, Dahan, & Embick, 2017; Zellou, et al., 2013, 2016).

Although phonetic imitation has been observed to be influenced by a range of social factors such as gender, attractiveness, and social attitudes (e.g., Babel et al., 2014;

Yu et al., 2013), the present study is primarily concerned with the effects of linguistic factors such as maintenance of the phonological contrast on phonetic imitation (Flege & Eefting, 1988; Kwon, 2015; Mitterer & Ernestus, 2008; Mitterer & Müsseler, 2013; Nielsen, 2011; Podlipský & Simáčková, 2015; Zellou et al. 2013, 2016). Previous research has suggested that there is a role for phonological categorisation in imitation. For example, Nielsen (2011) examined imitability of shortened and extended VOT in the English voiceless stop /p/. She found that only extended VOT was imitated and suggested that this is because imitating reduced VOT undermines a phonological contrast (i.e., /b/-/p/) whereas imitating extended VOT has no such consequences. Flege and Eefting (1988) also examined imitation of a VOT continuum ranging from /da/ to /ta/ and found that VOT was not imitated in a gradient fashion. Rather, participants made categorical responses to the VOT continuum, revealing discontinuous patterns of imitation at the phoneme boundary between /d/ and /t/. Based on the results, they suggested that imitation was influenced by phonological categorisation. Similarly, Mitterer and Ernestus (2008) examined imitation using a shadowing task and found that longer pre-voicing in Dutch was not imitated. They again suggested that this was because the degree of pre-voicing is not relevant to Dutch phonological contrasts. More recently, Kwon (2015) examined imitation of manipulated VOT and f0 in aspirated stops of Seoul Korean speakers. These stops are distinguished from tense and lax stops primarily by higher f0 on the following vowel but also by longer VOT (a secondary cue). Kwon found that enhancing the secondary cue (extended VOT) promoted increases in both VOT and f0 whereas enhancing the primary cue (higher f0) only induced imitation of high f0. This result also suggests a role for phonological contrast in phonetic imitation as the primary cue was most sensitive to enhancement.

In addition to the phonological influence on phonetic imitation, recent studies have suggested that cognitive factors such as perceptual salience affect imitation

separately from phonological categorisation. Salience has been referred to as the extent to which something is more noticeable than related items (Honeybone & Watson, 2013; Rácz, 2013). A salient item thus draws more attention than less salient items and an individual is more likely to pay attention to it and become aware of it. In speech perception, when a certain stimulus is particularly salient, it is easily attended to and encoded in memory (Nosofsky, 1991). In exemplar-based models (e.g., Goldinger, 1998), salient items, which stand out relative to others, might be stored and processed differently for subsequent speech production. In their study of spontaneous imitation, Zellou et al. (2013) examined the imitability of reduced and extended nasality in English vowels and found that both were imitated. However, only those exposed to reduced nasality showed carry-over to a post-shadowing word-reading task. They attributed the persistence of this property to the fact that reduced nasality is less typical in English and thus perceptually more salient. Similarly, Babel, McGuire, Walters, and Nicholls (2014) investigated the effects of cognitive novelty and social preference in an auditory naming task and showed that more imitation was elicited by atypical voices, which may be considered more salient, as well as by socially preferred attractive voices. Podlipský and Simáčková (2015) also suggested that perceptual salience plays an important role in phonetic imitation. They tested imitation of reduced and extended pre-voicing for a Czech voiced stop and reduced and extended vowel duration for a Czech long vowel. They observed imitation of extended pre-voicing but no imitation of reduced pre-voicing consistent with contrast preservation, however they also argued that it is consistent with reduced pre-voicing being less salient than extended pre-voicing to Czech listeners. Further, they found that unlike pre-voicing, reduced as well as extended vowel duration was imitated even though shortening vowel duration undermined the vowel length contrast. Consequently, they concluded that perceptual salience may be more important than contrast preservation in phonetic imitation.

In short, imitation of speech sounds is not an automatic consequence of the unmediated link between perception and production. Rather, a variety of social, linguistic, and cognitive factors come into play in the process of phonetic imitation. In particular, studies have shown some inconsistent findings in the phonological influence on imitation. Some studies have suggested that phonetic imitation is not likely to be induced when it threatens a phonological contrast (e.g., Nielsen, 2011) while others have suggested that preservation of the phonological contrast does not necessarily preclude imitation of perceptually salient properties (e.g., Podlipský & Simáčková, 2015). In the present study, we also investigate whether and how imitability of model speech is influenced by preservation of the phonological contrast and perceptual salience to better understand previous inconsistent findings. We further explore how the effects of phonological categorisation and perceptual salience on imitation interact with the link between speech perception and production described above.

An important difference between the present study and the aforementioned studies in phonetic imitation is that the previous studies examined spontaneous phonetic imitation (e.g., spontaneous word shadowing). In spontaneous imitation, participants are not aware of whether they modified their speech, and social and attitudinal factors likely come into play in this implicit imitation process. However, the current study uses a forced imitation task in which participants are explicitly told to imitate target speech as closely as possible (see also Zetterholm, 2006 and D'Imperio & German, 2015 for examples of explicit imitation tasks). Although it is not inconceivable that social and attitudinal factors might still be at work in imitation processes to some extent, these factors (their willingness to imitate) would be less likely to be involved in explicit (forced) imitation than in implicit (spontaneous) imitation. Also, participants would be more likely to show their imitative ability in explicit imitation than in implicit imitation. One study that compared spontaneous and explicit phonetic imitation (Delaney, Savji,

and Babel, 2010) found that participants overall showed greater imitation of target vowels in explicit imitation, consistent with the prediction that explicit imitation focuses on what participants 'can' do. We expect that using explicit imitation allows us to more directly investigate how perceptual and production representations are linked, minimizing confounding factors such as social influences that may modulate what listeners are willing to imitate.

1.3 The present study

The present study examines the link between speech perception and production using explicit phonetic imitation and a perceptual identification task. The first goal of this study is to test the hypothesis that patterns of phonetic imitation are related to perceptual cue weighting at an individual level. Previous work suggested that phonetic imitation reflects listeners' ability to convert the phonetic forms they hear into their subsequent production (Nielsen, 2011, 2014). We predict that an individual who is more sensitive to acoustic-phonetic information in identifying speech sound categories, will imitate variation in that information more. To this end, we manipulate the spectral and duration cues in an English vowel contrast (i.e., /ɛ/-/æ/), and examine the relative contribution of the two cues to vowel categorisation in perception and to production in imitation.

We chose the English /ɛ/-/æ/ contrast based on previous findings that these English low front vowels, especially /æ/, are some of the most regionally variable vowels in North America due to sound change such as the Canadian Shift (Boberg, 2008). Thus, we expect considerable individual differences in the production and perception of these vowels. Secondly, this is a vowel contrast where we expect both spectral and duration cues to contribute to vowel categorisation. For vowel contrasts in English, it has been observed that native English listeners will rely primarily on spectral

quality (Escudero, 2000; Hillenbrand, Clark, & Houde, 2000; Kondaurova & Francis, 2008, 2010; Liu & Holt, 2015). Previous work showed that /ɛ/-/æ/ are spectrally distinct, but also that /æ/ tends to be longer than /ɛ/ in productions from native English speakers (Hillenbrand, Getty, Clark, & Wheeler, 1995; Hillenbrand et al., 2000). Hillenbrand et al. (2000) also reported that the /ɛ/-/æ/ contrast is one of the vowel contrasts in English that shows the most robust duration effects in perception. In other words, although native listeners primarily use spectral cues to distinguish /ɛ/-/æ/, they are affected by vowel duration more than other vowel contrasts in English.

The second goal of the present work is to examine the underlying mechanisms of phonetic imitation as to whether and to what extent imitability of model speech is modulated by phonological categorisation and perceptual salience. Our manipulated vowel contrast provides a way to distinguish between these two hypotheses. If phonetic imitation is constrained by preservation of the phonological contrast, unambiguous vowels (i.e., two naturally-produced endpoints) should be better imitated compared to ambiguous vowels as these threaten the contrast. This may be more true of spectral cues than duration cues, as spectral cues are most important to the contrast, but we might still expect that duration values which threaten the contrast are imitated less. As in previous studies, we included extended and shortened vowel duration for both vowels. In the case of /æ/-like vowels, extended duration enhances the contrast while shortened duration threatens it. In the case of /ɛ/-like vowels, shortened duration enhances the contrast while extended duration threatens it.

Although a formal definition and quantification of perceptual salience that is used consistently by researchers remains elusive (MacLeod, 2015), we use the term perceptual salience on two grounds in the present study: (1) perceptual or cognitive noticeability (Honeybone & Watson, 2013; Rácz, 2013) and (2) listeners' expectations based on their language experience (Hay, Drager, & Gibson, 2018; Jaeger &

Weatherholtz, 2016). According to these grounds, we assume that extended and reduced vowel durations are perceptually salient because they are more extreme than intermediate vowel durations, although we do not have an independent measure of salience in the present study. We also assume that extended /ɛ/-like vowel tokens and shortened /æ/-like vowel tokens are perceptually salient because they differ from listeners' experience-driven expectations (i.e., /ɛ/ is generally short and /æ/ is generally long). Based on these assumptions, vowel duration may be especially salient as it was imitated in Czech vowels even when it undermined the vowel length contrast (Podlipský & Simáčková, 2015). Thus, if perceptual salience drives imitation, we may see imitation of both shortened and extended durations for both spectrally /ɛ/-like and spectrally /æ/-like vowels. This includes extended /ɛ/-like vowel tokens and shortened /æ/-like vowel tokens which should be especially threatening to contrast preservation but may still be perceptually salient.

In sum, this study uses phonetic imitation as a direct method to explore the link between perception and production in speech based on the assumption that phonetic imitation reflects listeners' ability to convert the phonetic forms they hear into their subsequent production. In addition, the imitation processes provide an opportunity to investigate how listeners map acoustic-phonetic properties in the speech signal onto their production outputs by reducing or enhancing relevant properties. We expect that close investigation into the imitation of carefully manipulated target speech stimuli will shed light on the mechanisms of phonetic imitation and will further contribute to our understanding of the perception-production link in terms of individuals' use of acoustic-phonetic information.

2 Methods

2.1 Participants

Twenty-three native speakers of North American English (mean age = 20.7, range = 18–31) were paid for their participation. In order to remove potential complications due to imitations across gender, we recruited only female participants and used a female talker for imitation targets. All participants were monolingual speakers of North American English either from Canada or the US. Based on the primary dialect regions in North America in Labov, Ash, and Boberg (2006), the participants were from Canada (n = 12), New England (n = 4), New York City (n = 2), the West (n = 2), and the South (n = 1).[1] None of the participants reported speech or hearing impairments.

2.2 Stimulus creation

Figure 1 illustrates both perception stimuli for perceptual cue weighting and production stimuli used for phonetic imitation. Stimuli were created based on natural recordings of *head* and *had* tokens produced by a female Canadian English speaker in her 20s. The natural productions were recorded with a high-quality recorder (Zoom H4n, 44.1 kHz sampling rate, 16-bit) and then resynthesized to create a continuum from *head* to *had* using TANDEM-STRAIGHT in MATLAB, which is a high-quality vocoder that allows for the creation of natural-sounding continua (Kawahara, Takahashi, Morise, & Banno, 2009).

A twenty-step continuum (from /ɛ/ to /æ/) was first created. Then, 5 native speakers of English who were naïve to the purposes of the task were asked to categorise

[1] According to Labov et al. (2006)'s classifications of short-a systems, the pronunciation of the target vowel /æ/ in 'had' is similar for all dialectal regions of the participants in the present study. Although NYC has a split system with /æ/ tensing under some conditions the word 'had' is not thought to be subject to tensing (Labov et al., 2006, p. 172). Because of its final consonant, this word is also not subject to raising before nasals or velars that affects some dialects. While there may also be phonetic variability between regions, Clopper and Pisoni (2006) reported little differences in /hæd/ for the regions affecting our participants.

the continuum and the most ambiguous step (50% *had* responses) was selected. Based on the most ambiguous token and the two end-point tokens, seven well-separated stimuli along the continuum were chosen based on first (F1) and second (F2) formant values (Table 1). From each of the seven steps along the vowel spectral continuum, seven-step vowel duration continua ranging from 100 ms to 340 ms (40 ms/step) were created using the PSOLA method implemented in Praat (ver. 6.0.03, Boersma & Weenick, 2015). These 49 stimuli, varying in two dimensions—vowel spectrum and duration—from /ɛ/ to /æ/ formed the stimuli for the perception task.

Table 1. F1 & F2 values of the vowels in the *head-had* continuum.

Step	F1 (Hz)	F2 (Hz)
1	685	1988
2	729	1956
3	766	1899
4	791	1840
5	830	1787
6	876	1733
7	918	1640

Production stimuli consisted of a total of 15 stimuli, including 9 stimuli from the perception experiment (blue circles), and also additional stimuli of shortened (60 ms) and extended (380 ms) vowel durations (6 blue diamonds). Shortened and extended vowel durations (salient tokens) will be compared with the other moderate vowel durations (non-salient tokens) in imitability and will test the role of perceptual salience and phonological structure. Spectrally ambiguous tokens were included to test whether participants show imitation of phonetic details or maintenance of phonological contrast.

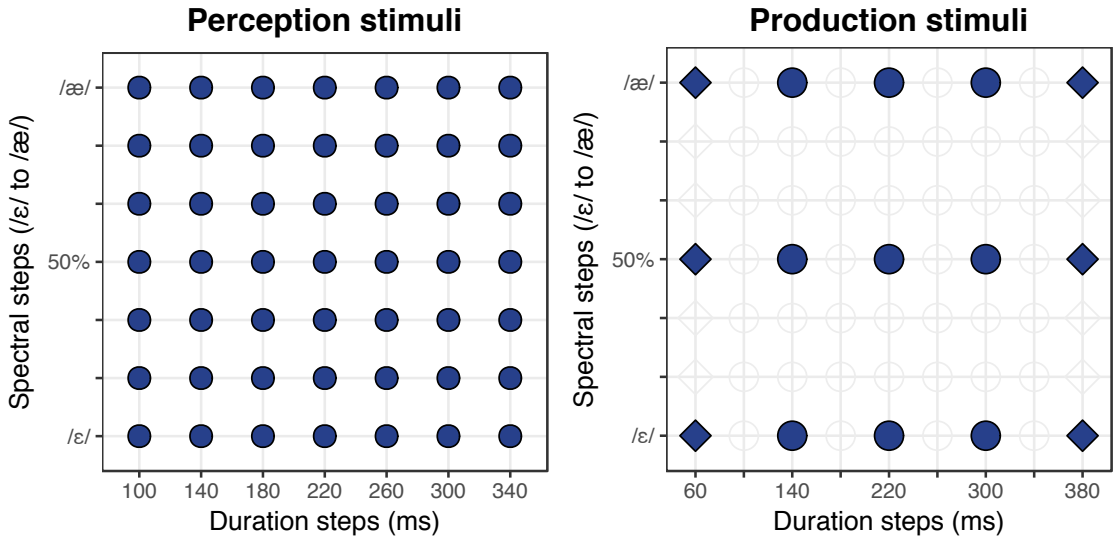

Figure 1. Illustration of stimuli used in perception and production tasks in terms of their spectrum and duration values. Blue circles indicate stimulus items included in both perception and production tasks. Blue diamonds (salient tokens) represent stimuli that were included in the production task only.

2.3 Procedure

The perception and production experiments were conducted at McGill University, Canada. Participants sat in front of a computer and were tested individually in a sound-attenuated booth and received both oral and written instructions about the experiments. They were tested on perception first, and then recorded their baseline productions and carried out the imitation task.

2.3.1 Perception

The perception stimuli were presented with a two-alternative forced choice identification task in MATLAB. Participants heard the words *head* and *had* and responded with a key press to indicate which word they heard. There were 49 stimuli repeated 5 times in each block (245 trials in total), and all the trials within a block were randomly presented through headphones at a comfortable listening level.

2.3.2 Production

Participants first completed a baseline task. The English words *head* and *had*, which were also used in the perception and imitation tasks, were each presented on a computer screen 8 times in a randomized order and participants were instructed to produce each as naturally and clearly as possible. For analysis, only the tokens from the second to sixth productions were analysed because the first and the last productions were expected to be somewhat unnatural. Participants then completed a forced imitation task. In the imitation task, participants were explicitly instructed to imitate the target stimuli as closely as possible after hearing the imitation target. Each target stimulus was played twice so that participants could be well aware of what they would imitate and that they did not miss the imitation target. The inter-stimulus interval was 1.2 seconds. Fifteen stimuli were repeated 6 times in each block (90 trials in total), and all the trials within a block were randomized and presented though headphones. The imitation stimuli were presented in MATLAB and the participants' baseline and imitation productions were recorded using a Logitech H390 headset microphone (22.05 kHz sampling rate, 16-bit).

2.3.3 Imitation measures

For both baseline and imitation productions, vowel spectral quality, namely F1 and F2 values (Hz) at the temporal midpoint of each vowel, and vowel duration (ms) were manually measured in Praat, using both waveforms and spectrograms to determine the vocalic portion. For baseline productions, Linear Discriminant Analysis (LDA) was performed using the *MASS* package in R (R Core Team, 2016) to measure how much each acoustic dimension contributes to categorisation of the vowel contrast within each speaker (Schultz et al., 2012).

For all of the acoustic measures of imitation, degree of imitation on each trial was defined as the distance from the imitated value to the target value, subtracted from

the distance from the baseline value to the target value, namely Degree of imitation: $|X_{target} - X_{baseline}| - |X_{target} - X_{imitation}|$ (Babel, 2012; Phillips & Clopper, 2011; Walker & Campbell-Kibler, 2015). This metric indicates the number of units (Hz or ms) that the participant has moved their production towards the target and away from their own baseline. A negative value of imitation indicates divergence from the target speech whereas a positive value shows convergence to the target speech. Degree of imitation was calculated for vowel spectral quality and duration separately. Degree of imitation for vowel spectral quality was calculated using the F1-F2 Euclidean distance. Degree of imitation for vowel duration was calculated using vowel duration differences. Baseline measures for vowel duration and vowel quality for *head* and *had* target stimuli were the average of the 6 baseline productions of *head* and *had* respectively for each participant. The baseline measures for duration and spectral quality for the spectrally ambiguous targets were the mean of each participant's /ɛ/ and /æ/ productions. Thus, three baseline vowel durations and three baseline vowel qualities (*head, had,* and ambiguous) were calculated for each participant. For example, for the target stimulus with /ɛ/ spectral quality and a duration of 380 ms, each participant's baseline would be the duration and F1 and F2 measures from their 6 productions of *head*.

3 Results

3.1 Speech perception

To examine the extent to which vowel spectral quality and duration are used in vowel categorisation in perception, the participants' responses were analysed using mixed-effects logistic regression using the *glmer* function in the *lme4* (1.1-14) package in R. In addition to perceptual patterns at a group level, individual participants' perceptual weights for vowel spectral quality and duration were also calculated based on spectrum and duration coefficients, using a series of separate logistic regression analyses fitted to

each listener's vowel categorisation responses (Morrison, 2005, 2007; Morrison & Kondaurova, 2009; Schertz, Cho, Lotto, & Warner, 2015; Shultz et al., 2012). The coefficients from the individual models served as measures of the perceptual weights of their respective cues and were used to make relevant comparison with participants' production measures such as imitation performance and baseline productions.

3.1.1 Perceptual cue weighting: Group-level

Figure 2 shows group-level responses for vowel spectrum and duration weighting for categorisation of /ɛ/-/æ/, as well as overall contribution of spectral cues and duration in a heatmap representation. The overall pattern of categorisation responses indicates a strong and categorical effect of spectral cues on vowel categorisation, and it also shows an influence of duration cue on categorisation but to a much weaker degree as expected (Escudero, 2000; Hillenbrand et al., 2000; Kondaurova & Francis, 2008, 2010; Liu & Holt, 2015).

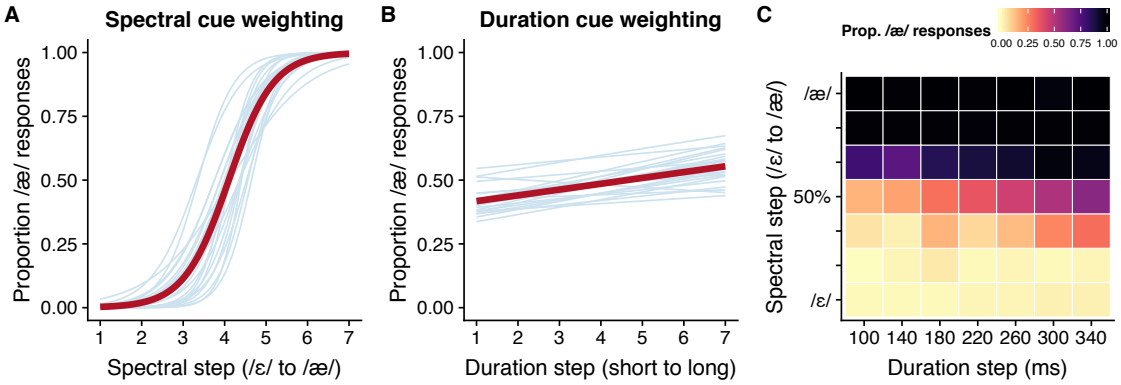

Figure 2. Proportion of /æ/ responses along vowel spectral quality continuum (A) and vowel duration continuum (B) averaging across group, as well as a heatmap plot of overall perceptual weights for spectral quality and duration (C). Logistic curves fit to each individual listener are also shown in light blue in A and B.

A mixed-effects logistic regression analysis with random intercepts and random slopes for spectral and duration steps for participants confirms that listeners rely more

heavily on vowel spectral quality ($\beta = 9.23$, $z = 17.64$, $p < 0.001$) than duration ($\beta = 1.45$, $z = 9.01$, $p < 0.001$) although both dimensions significantly contribute to vowel categorisation.

3.1.2 Perceptual cue weighting: Individual-level

Figure 3 illustrates a scatter plot of individual cue weights based on coefficients from individual regression models. Visual inspection of the plot shows that although listeners rely more heavily on spectral cues, there is considerable individual variability in the magnitude of spectral and duration cue weights. That is, some listeners are more sensitive to spectral dimensions (i.e., larger spectral cue weights) than others while other listeners are more sensitive to duration dimensions (i.e., larger duration cue weights) than others. Also, cue weights in the two dimensions seem to be related to each other within individuals.

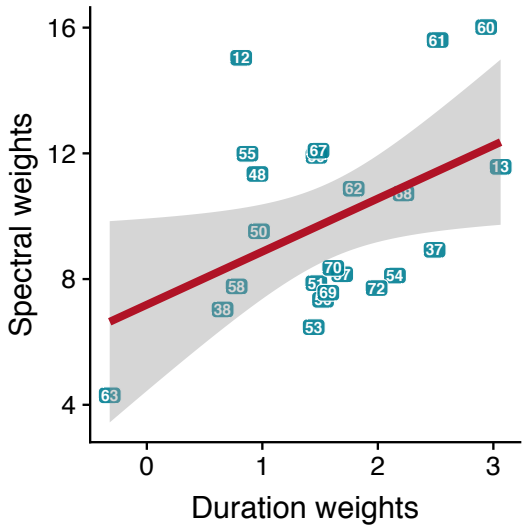

Figure 3. Scatter plots of individual listeners' perceptual cue weights showing the correlation between vowel spectral cue weights and duration cue weights.

A result of the Pearson correlation showed a significant positive correlation between spectral cue weights and duration cue weights ($r = 0.44$, $p = 0.03$). That is, listeners who showed greater use of spectral cues also showed greater use of duration cues. This suggests that some listeners are generally better able to use acoustic-phonetic information in the input signal than others in vowel categorisation. This is consistent with previous trends observed in Shultz et al. (2012) for VOT/f0 as cues to consonant voicing in English.

3.2 Speech production

3.2.1 Baseline production

Baseline production values across all speakers for vowel spectral quality (F1 and F2 values) and vowel durations are provided in Table 2. Mean spectral quality values in baseline productions are also plotted in Figure 5, along with imitated values.

Table 2. Mean vowel spectral and duration values of the baseline productions. Standard deviations are given in parentheses.

	/ɛ/	/æ/
F1 (Hz)	762 (71)	980 (82)
F2 (Hz)	2085 (91)	1786 (79)
Duration (ms)	186 (37)	266 (46)

To explore the relationship between perception and production at baseline in terms of the use of acoustic cues by individuals, Linear Discriminant Analysis (LDA) was conducted for each individual participant (Schertz et al., 2015; Shultz et al., 2012). Each LDA model included F1, F2, and vowel duration and calculated the relative

weighting of predictors in the categorisation of /ɛ/ and /æ/ in production. Each LDA model provided coefficients for predictors, and a larger coefficient indicates a stronger weighting of the predictor. Individuals' cue weighting strategies across perception and production were investigated using Pearson product-moment correlation on individual LDA coefficients in baseline production and cue weights in perception. Although correlations between spectral dimensions were in a positive direction, we did not find significant correlations between the perception and production weights for any of the dimensions as in Table 3 and in Figure 4. That is, we did not find evidence that production patterns by individual speakers at baseline were correlated with their perceptual cue weighting strategies, consistent with previous findings in which there is no or at best weak correlation between these two tasks (Schertz et al., 2015; Shultz et al., 2012). In the next section, we test the hypothesis that the link between perception and production is more robust in phonetic imitation as perceptual patterns are translated into subsequent productions. Thus, we expect participants cue weighting strategies in perception will be predictive of their patterns of phonetic imitation.

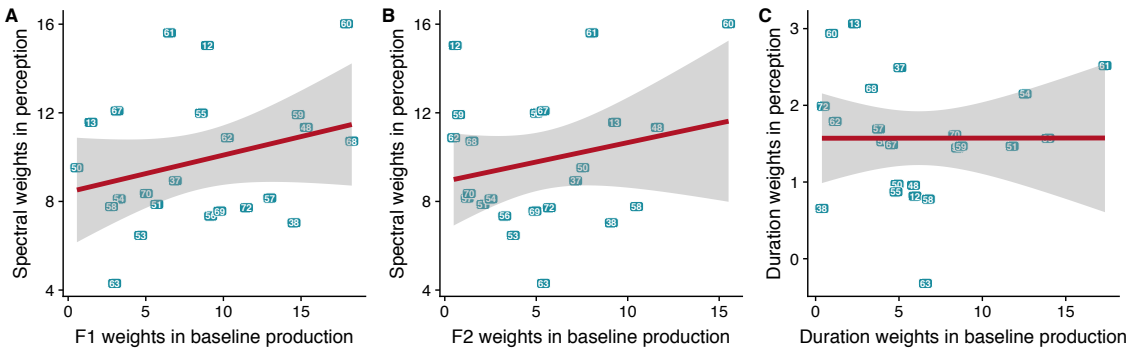

Figure 4. Scatter plots of individual cue weights in perception and individual LDA weights from baseline productions. Spectral weights in perception by F1 weights in production at baseline (A), spectral weights in perception by F2 weights in production at baseline (B), and duration weights in perception by duration weights in production at baseline (C).

Table 3. Results of correlation analyses between perception and production weights.

Dimension	*r*	*p*
Spectral weights/Baseline F1	0.29	0.17
Spectral weights/Baseline F2	0.23	0.28
Duration weights/Baseline duration	0.001	0.99

3.2.2 Imitation

Separate linear mixed-effects models were built for the analysis of spectral and duration imitation using the *lmer* function in the *lme4* (1.1-14) package in R. Dependent variables in each model were degree of imitation (described above). Between-participant predictor variables included spectral cue weights in perception (SPECTRUM), duration cue weights in perception (DURATION), DIALECT, the Euclidean distance between the target and baseline (TARGETBASEED) for the model of vowel quality imitation, and the absolute duration difference between target and baseline (TARGETBASEDD) for the model of vowel duration imitation. To reduce multicollinearity between predictors and also to interpret the predicted mean imitation across all data points, continuous variables namely SPECTRUM, DURATION, TARGETBASEED and TARGETBASEDD were standardized by centring and dividing by 2 standard deviations using the *rescale()* function from the *arm* package in R. SPECTRUM and DURATION examined the effects of perceptual cue weights on imitation of vowel spectral quality and duration, and TARGETBASEED and TARGETBASEDD were motivated by the prediction that greater distance between target and baseline increases likelihood of imitation, as observed in previous studies (Babel, 2012; Walker & Campbell-Kibler, 2015). DIALECT was motivated by previous findings that participant

dialect plays a role in phonetic imitation (Babel, 2012; Phillips & Clopper, 2011; Walker & Campbell-Kibler, 2015), which was centred (–0.5 and 0.5) and examined the effect of dialects on imitation comparing US (as the reference level, coded 0.5) with Canada. The predictor variables also included within-participant variables target spectral step (TARGETSS) and target duration step (TARGETDS). TARGETSS was Helmert-coded and examined the effects of natural and ambiguous vowel tokens. The first contrast compared /ɛ/ with /æ/ ("ɛ" as the reference level), and the second contrast compared the ambiguous vowel with the average of the two natural vowels ("ambiguous" as the reference level). TARGETDS examined the effects of differential target vowel duration on imitation of vowel quality. TARGETDS was treated as an ordered factor using polynomial contrasts coding to test the linear components of duration steps. For the model of vowel spectral imitation, possible two-way interactions between TARGETSS and TARGETDS to examine whether imitation of different vowel spectral targets was modulated by vowel duration. For the model of vowel duration imitation, SALIENCE was included to examine the effect of perceptual salience on phonetic imitation, which was centred (–0.5 and 0.5) and tested whether salient target tokens (i.e., extended and shortened vowels, coded –0.5) lead to greater likelihood of imitation than the other target tokens (as the reference level, coded 0.5). Both models included random intercepts for participants to account for participant specific variability in their degree of imitation. By-participant random slopes were also included for TARGETSS and TARGETDS for the spectral imitation model and for TARGETSS and SALIENCE for the duration imitation model. The converging model with the maximal random-effects structure was used for both spectral and duration imitation models (Barr, Levy, Scheeper, & Tily, 2013). The *p*-values in the models were calculated based on the Scatterthwaite approximation, implemented in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2015) in R.

3.2.3 Descriptive plots of imitation

Before the statistical analyses in the following sections, descriptive plots of vowel spectral quality and duration imitation are first shown in Figure 5. Figure 5A displays formant frequencies of all the imitated vowel tokens, using a scatter plot and overlaid contour plot which shows the distributional peaks of imitation of each vowel spectral quality. In Figure 5A, imitation of the ambiguous target vowel showed more variation and a wider distribution than that of the other two unambiguous vowel targets. Figure 5B illustrates imitated vowel durations depending on target vowel durations, using a box plot with a violin plot to show the distribution of imitated vowel durations. Unlike imitation of vowel spectral quality in which phonologically ambiguous tokens were not accurately imitated, Figure 5B indicates that all the vowel duration steps were relatively accurately imitated (except perhaps the shortest duration).

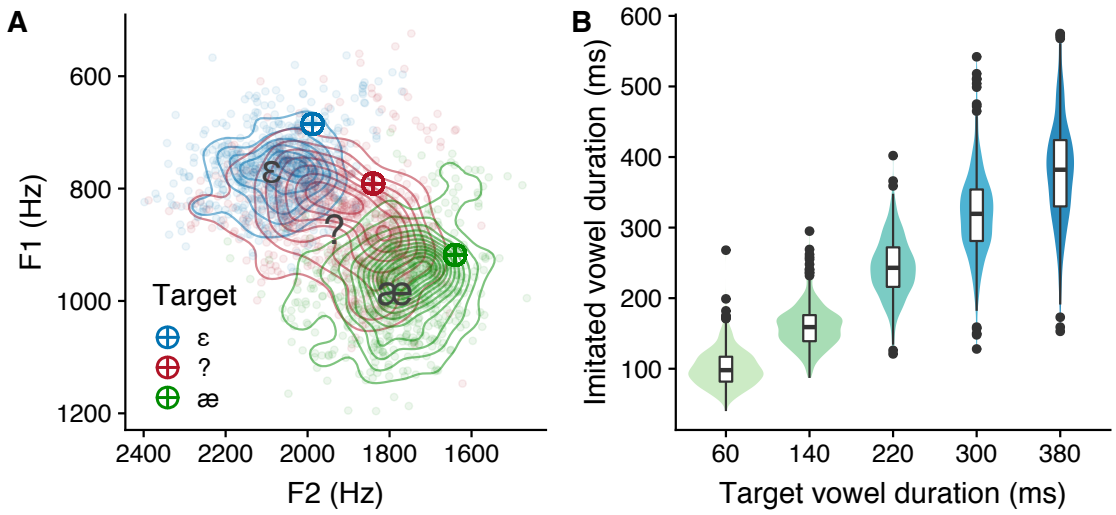

Figure 5. Formant plots of all the imitated vowel tokens (A). The black symbols overlaid on the density plot in A (i.e., "ɛ", "?", and "æ") indicate mean baseline values. Imitated vowel durations depending on target vowel durations (B).

3.2.4 Vowel quality imitation results

Table 4 summarises the estimated values for fixed-effect coefficients (β), along with

their standard errors, *t* statistic, and corresponding significance values (*p*). Figure 6 illustrates the empirical plots of the factors with significant effects on imitation of vowel quality, which will be described below.

Table 4. Summary of vowel quality imitation model. Model coefficient estimates (*β*), standard errors, corresponding *t*-values, and *p*-values.

Predictor	Estimate (*β*)	Std. Error	*t*	*p*
Intercept	–6.73	11.27	–0.60	0.56
TARGETSS (ε)	–6.73	3.79	–1.78	0.09
TARGETSS (ambiguous: A)	–4.00	1.65	–2.43	0.02
TARGETDS	27.49	5.69	4.83	< 0.001
TARGETBASEED	99.94	9.88	10.12	< 0.001
SPECTRUM	35.45	20.70	1.71	0.10
DURATION	–6.86	21.40	–0.32	0.75
DIALECT	44.41	19.60	2.11	0.05
TARGETSS (ε) × TARGETDS	22.52	22.10	3.87	< 0.001
TARGETSS (A) × TARGETDS	–1.69	2.40	–0.71	0.49

The intercept of the model (negative) indicates that on average the model estimates that participants moved away from the target relative to their baseline although this is not statistically significant. This is probably due to the ambiguous vowel quality. The biggest predictor of imitation degree was spectral distance between target and baseline, as shown in Figure 6C ($\beta = 99.94$, $t = 10.12$, $p < 0.001$). Imitation scores for spectral quality increased as spectral distance between target speech and baseline production increased. This is expected in part because for a given participant and target vowel quality, baseline values that are identical to the target (target to

baseline difference = 0) cannot move closer to the target during imitation while participants whose baseline for a given target vowel quality is most different can move up to 326 Hz closer to the target during imitation. Importantly, Figure 6C also shows that participants whose baseline was not close to the target vowel quality did move closer to the target when asked to imitate.

As shown in 6A, there was a marginal difference between /ɛ/ and /æ/ (β = –6.73, t = –1.89, p = 0.09) indicating potentially greater imitation of /æ/ than /ɛ/ and significantly greater imitation scores for spectrally unambiguous vowels than ambiguous tokens (β = –4.00, t = –2.43, p = 0.02). Figure 5A above and the mostly negative imitation scores for the ambiguous vowels indicate that the imitations of these vowels rarely reached an ambiguous value and instead were more towards the imitation values for /æ/ or /ɛ/ targets. Figure 6B shows that vowel spectral imitation was influenced by target vowel duration with a significant positive linear trend with increasing vowel duration (β = 27.49, t = 4.83, p < 0.001). This indicates that participants showed greater imitation with increasing vowel duration. This may also suggest that longer vowels provide better information about the target spectral quality than shorter vowels, especially for /æ/ targets. Participants' dialectal background also had a significant effect (β = 44.41, t = 2.11, p = 0.05). Figure 6D shows that participants from the US had higher vowel quality imitation scores than those from Canada. As the model talker was Canadian, we predicted that this would be due to smaller mean target-baseline distance for participants from Canada versus the US, however target-baseline distance was quite similar across the two groups (156 Hz vs. 152 Hz). Instead, this may indicate that other socio-dialectal factors are at work in this case.

In addition to these main effects on vowel quality imitation, the model found two-way interactions between target vowel quality and duration shown in Figure 6E. The interaction of TARGETSS (ɛ) and TARGETDS indicates that the effects of target

duration on spectral imitation were greater for imitation of /æ/ than /ɛ/ ($\beta = 22.52$, $t = 3.87$, $p < 0.001$).

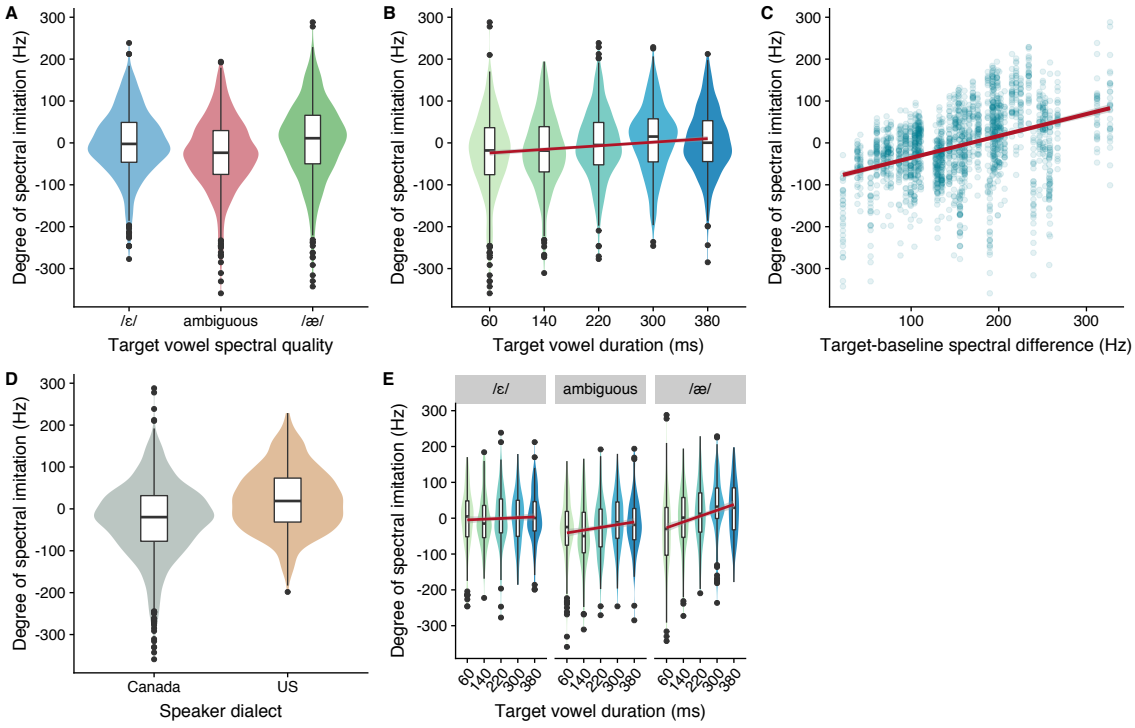

Figure 6. Empirical plots of significant predictors in vowel quality imitation (A–E).

3.2.5 Vowel duration imitation results

Results from the imitation of vowel duration model are summarised in Table 5. Figure 7 illustrates the empirical plots of the factors with significant effects on imitation of vowel duration in the model, which are described below.

Table 5. Summary of vowel duration imitation model. Model coefficient estimates (β), standard errors, corresponding *t*-values, and *p*-values.

Predictor	Estimate (β)	Std. Error	*t*	*p*
Intercept	0.998	0.05	19.17	< 0.001
TARGETSS (ε)	–0.019	0.04	–0.54	0.59
TARGETSS (ambiguous)	–0.028	0.02	–1.30	0.20
SALIENCE	–0.235	0.10	–2.27	0.03
TARGETBASEDD	1.854	0.06	33.49	< 0.001
SPECTRUM	0.234	0.11	2.08	0.05
DURATION	0.003	0.12	0.02	0.98
DIALECT	–0.084	0.11	–0.78	0.44

As seen in the previous section for spectral quality imitation, there is a significant effect of duration differences between target and baseline on imitation of vowel duration, as in Figure 7B ($\beta = 1.854$, $t = 33.49$, $p < 0.001$) indicating that participants had higher imitation scores when there was a large duration difference between target and baseline. No effect of target spectral quality was found on duration imitation, however the model estimated a significant difference between salient vowel tokens (i.e., the extended and shortened vowel stimuli) and the other non-salient vowel tokens as shown in Figure 7A ($\beta = –0.235$, $t = –2.27$, $p = 0.03$). This indicates that higher imitation scores were obtained for perceptually salient target vowel durations than for the other moderate vowel durations.

Notably, as shown in Figure 7C, spectral cue weights in perception predict degree of duration imitation ($\beta = 0.234$, $t = 2.08$, $p = 0.05$), indicating that listeners with higher spectral cue weights in perception also show greater duration imitation. In contrast, duration cue weights in perception did not have a significant effect on the

degree of duration imitation. This might be because the duration cue plays a much smaller role relative to spectral cues in vowel categorisation in English.

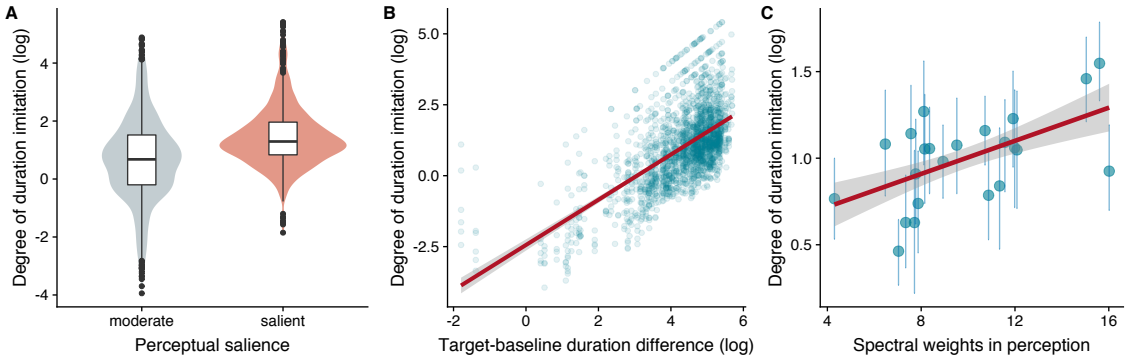

Figure 7. Empirical plots of significant predictors in vowel duration imitation (A–C). Error bars in C indicate 95% confidence intervals.

In our last analysis, we compared overall patterns of imitation between vowel spectral quality and duration. That is, we probed each individual's degrees of spectral and duration imitation using the random effect intercepts fitted for each participant from the spectral and duration imitation models. Each individual's spectral and duration cue weights in perception, as shown in Section 3.1.2, were juxtaposed to compare them with the overall patterns of imitation. First, a correlation analysis was conducted to examine whether spectral imitation and duration imitation are related for an individual speaker. The results showed that there was no significant correlation between individuals' spectral and duration imitation ($r = 0.26$, $p = 0.21$), suggesting that they might involve different underlying mechanisms. Figures 5A and 6A demonstrate that individuals exhibited enormous variability in their imitation of spectral quality, even varying in the direction of spectral imitation. That is, some participants on average modified their baseline productions in response to target stimuli by aligning their speech with the spectral quality of the targets while others diverged from the target speech and still others did not change on average. In contrast, all participants exhibited accommodation

of duration to the target stimuli although there were individual differences in the magnitude of duration imitation. This shows that adjusting vowel duration to the target stimuli is relatively more flexible than vowel spectral quality and thus all participants demonstrated convergence to the target duration stimuli. This flexibility might explain why duration imitation and not spectral imitation is predicted by perceptual cue weights (although it is spectral cue weights and not duration cue weights that predict). On the other hand, it might be the case that variation of spectral imitation was not captured by individual differences in their patterns of perceptual cue weighting because phonological or dialectal factors were also at work in imitability of vowel spectral quality. Crucially, comparison of the imitation patterns (8A and 8B) with perceptual cue weights (8C and 8D) indicates that the dimensions that are linked across perception and production are, in fact, those that best represent their perceptual performance in vowel categorisation (i.e., spectral weights) and their imitation performance in vowel production (i.e., duration imitation) at the individual level.

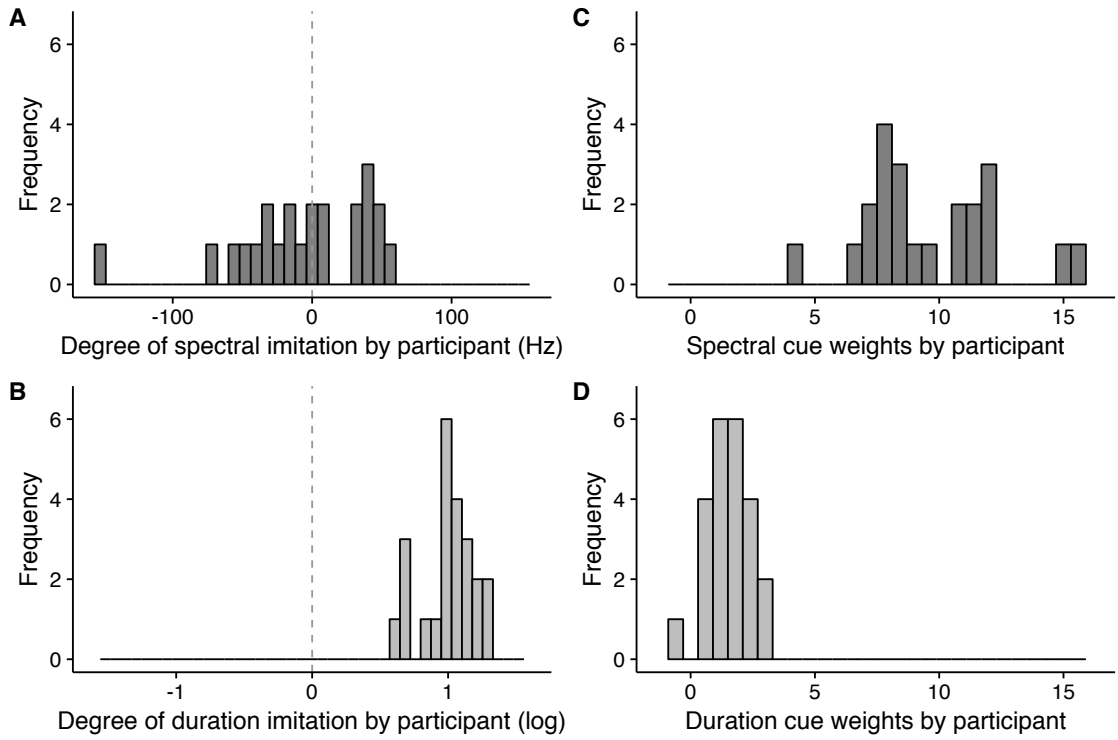

Figure 8. A histogram of each participant's degree of imitation based on the participant random intercepts from the spectral and duration imitation models (A and B), and each participant's spectral and duration cue weights in perception (C and D). Degree of spectral imitation is on the top left panel (A) and degree of duration imitation is on the bottom left panel (B). A positive value indicates convergence to the target speech whereas a negative value indicates divergence. The dashed line at 0 indicates no change. Individuals' spectral weights are on the top right panel (C) and individuals' duration weights are on the bottom right panel (D).

4 Discussion

The overarching goal of this study was to examine the nature of the relationship between speech perception and production at an individual level using phonetic imitation. More specifically, it was hypothesised that production patterns in phonetic imitation would be more reliably predicted by individuals' use of acoustic-phonetic cues in perception than in disparate perception and production tasks as in previous studies. We hypothesized a closer link because phonetic imitation may reflect listeners' capability to transform the phonetic forms they hear into their subsequent production. Furthermore, we examined the underlying mechanisms of phonetic imitation by using

manipulated speech stimuli, to determine whether and to what extent imitability of speech sounds is affected by phonological categorisation and/or perceptual salience. If phonological categorisation constrains imitation (Flege & Eefting, 1988; Mitterer & Ernestus, 2008; Mitterer & Müsseler, 2013), we should see less imitation of spectrally ambiguous vowels compared to spectrally unambiguous vowels, and less imitation of phonologically unnatural target vowel durations (i.e., extended /ɛ/-like vowel tokens and shortened /æ/-like vowel tokens) than natural target durations. On the other hand, if phonetic imitation is conditioned by perceptual salience, we should still see imitation of both extended and shortened durations separately from phonological categorisation.

In this study, we first examined the relation between individuals' perceptual cue weights and their baseline productions to confirm previous findings in which there is no (or at best weak) correlation between individual cue weights in perception and those in production (e.g., Shultz et al., 2012). Our results showed that there were substantial individual differences in the use of acoustic cues (i.e., vowel spectral quality and vowel duration) in both perception and production, but these individual differences in listeners' use of acoustic cues were not correlated with individual differences in their baseline productions, as in previous findings. However, in our subsequent phonetic imitation task we found that individuals' perceptual cue weights and production patterns were more tightly linked, confirming our hypothesis that a link between perception and production can be found when the production task is constrained to be more like the perception task and when both perception and production are part of the same task. The different findings for different tasks could be hypothesized to reflect differences in 'in the moment' versus longer-term representations, perhaps with more flexibility in the moment. However, we feel a better explanation is that our tasks capture participants' ability to use fine-grained acoustic-phonetic information to resolve ambiguity that might simply not be reflected in a word reading task.

Specifically, we found that individuals with greater spectral weights in perception showed better performance in vowel duration imitation. One interpretation is that some listeners are better able to use acoustic-phonetic information in perception than others as shown in the positive correlation between individual cue weights in Figure 2, and this superior perceptual acuity might be reflected in their imitation of vowel duration. This finding is consistent with a view of the perception-production link in which better perceptual acuity relates to better production performance in speech at the individual level (Brunner et al., 2011; Perkell et al., 2004). In the present study, however, we found that when multiple cues to a phonological contrast are taken into account, perceptual acuity might be more closely related to a particular acoustic cue (e.g., vowel duration).

This finding is inconsistent with our prediction that use of specific cues (e.g., vowel duration) would be linked in perception and production. That is, the results indicated no evidence for a link between individuals' perceptual cue weights for spectral quality and their subsequent imitation of vowel spectral quality even though vowel quality is the primary cue our participants used in vowel categorisation. One reason may be that imitation of spectral quality was overall poor and failed to closely correspond to the target speech (as shown in Figure 8A). This might indicate that the imitation performance of some individuals (i.e., especially those who have negative values in Figure 8A) was too poor to be reliably predicted. This interpretation is consistent with Ainsworth and Palatal (1984) in which they failed to find any systematic relationship between perception and production of English glides due to extreme variability within an individual as well as across individuals for their perception and production tasks. It is also possible that some other factors such as dialectal or social influences played a complicated role in their spectral imitation, which requires further research.

Another account for why the perception-production link is mediated in imitation of vowel spectral quality is related to our second question, which is to what extent imitation reflects phonological categorisation. In the present study, participants' poor and variable imitation of vowel spectral quality might be related to the influence of phonological categorisation on imitation. As discussed in the introduction section, previous studies on phonetic imitation have also found mediation of the perception-production link by phonological categorisation (Flege & Eefting, 1988; Mitterer & Ernestus, 2008; Mitterer & Müsseler, 2013). For example, Flege and Eefting (1988) found that phonologically ambiguous tokens along a VOT continuum were not closely imitated by native speakers of English and Spanish. Our results for vowel spectral quality are in line with these previous studies in that we found limited imitation for ambiguous targets.

In contrast, in the imitation of vowel duration we found that all the participants showed good imitation performance, indicating the translation of acoustic details into the imitation of vowel duration. One explanation for this asymmetry may be that phonological categorisation plays a role in imitation, but more strongly for the primary cue, namely vowel spectral quality. There are many possible mechanisms whereby this could occur. For example, Gambi and Pickering's (2013) Simulation Theory predicts that aspects of speech that are important for phonological categorisation and thus are associated with higher-level linguistic structures, require cognitive resources to be devoted to this level of processing. This in turn leads to less resources being available for bottom-up acoustic processing. Hence, vowel duration may be associated with relatively more bottom-up acoustic processing than vowel spectral quality and thus individuals' perceptual abilities may be more directly reflected in their imitation of vowel duration than of vowel spectral quality.

Other possible accounts can also be provided for this asymmetry. For example, speakers may accurately perceive an ambiguous target but be reluctant to imitate because it would threaten the contrast (Neilson, 2011) and ambiguous duration may have been less affected because it was less important to the categorisation. Perceptual magnet effects (Kuhl, 1991) suggest that the ambiguous tokens might be warped and perceived as more prototypical vowel productions and Iverson and Kuhl (1996) showed that more important dimensions are more strongly warped. Another possible explanation is that listeners detect subtle acoustic details of ambiguous spectral quality targets, but they have difficulty in storing and transforming them into phonetic imitation due to less robust memory traces without recourse to clear lexical representations (Savill, Ellis, & Jefferies, 2017). It is also possible that imitation is mediated by activation of motor plans rather than perception or storage of information in memory (Honorof, Weihing, & Fowler, 2011). Importantly, the present result extends previous findings by demonstrating that not all cues are affected by phonological categorisation equally and models of imitation will need to take this into account.

In contrast to accounts that appeal to the limited role of duration in categorisation, duration may have been better imitated because it does play an important role, just not in this vowel contrast. Better duration imitation may derive from participants' experience of utilizing duration cues in different linguistic and non-linguistic domains (e.g., speech rate, rhythm, and musical beat).[2] For example, vowel duration is manipulated for word final obstruent voicing in English as well as for prosodic purposes. Participants may have been able to imitate duration well because they are more easily able to imitate suprasegmental aspects of speech (e.g., Zetterholm, 2006 and D'Imperio & German, 2015 both demonstrate good imitation of suprasegmental aspects of speech in explicit imitation tasks).

[2] We thank two anonymous reviewers for this suggestion.

In addition to the effect of phonological categorisation on imitation, the present study also suggests that perceptual salience plays a role in phonetic imitation. Our results showed that both extended and shortened durations were still imitated even though they included target stimuli which might undermine the phonological contrast (i.e., extended /ɛ/-like vowel tokens and shortened /æ/-like vowel tokens). Further, perceptually salient tokens (i.e., extended and shortened vowel durations) induced relatively more imitation than non-salient tokens (i.e., intermediate durations). These findings indicate that preservation of phonological contrast does not necessarily constrain imitation, and imitation of these extended and shortened durations may instead be conditioned by their perceptual salience (Podlipský & Simáčková, 2015). Using manipulated vowel durations in Czech, Podlipský and Simáčková (2015) also found that salient vowel duration differences promote imitation although they impair the phonological vowel length distinction in Czech. They interpreted their findings as an indication that perceptual salience may exert more influence on phonetic imitation than contrast maintenance. It is also possible that listeners interpreted these not as shortened and extended vowels but as vowels spoken with a fast or slow speaking rate. If that were the case, then the durations would not necessarily threaten the contrast. Participants would need to be asked to imitate vowels within a sentence to rule out this possibility.

The present finding is also partly in line with previous work in which perceptually salient properties such as increased vowel nasality and VOT facilitate imitation of these properties (Nielsen & Scarborough, 2015; Zellou et al., 2013). Furthermore, the acoustic distance account (e.g., Babel, 2012; Walker & Campbell-Kibler, 2015) in which phonetic imitation is facilitated by greater baseline-target acoustic distance (as long as it is not socially salient) could also be interpreted in terms of perceptual salience. That is, as acoustic distance between baseline and target

increases, this increasing distance becomes more salient to listeners and in turn it leads to more likelihood of imitation.

It should be noted that the target vowel /ɛ/ and /æ/ in English differed in degree of spectral quality imitation, depending on the vowel duration. That is, participants generally showed more imitation of /æ/ than /ɛ/, though the effect was marginal. This is consistent with previous studies in which the /æ/ vowel in English generally facilitated more imitation than other vowels, perhaps because it is more variable (Babel, 2012; Walker & Campbell-Kibler, 2015). Also, imitation of /ɛ/ was not modulated by vowel duration differences whereas, we found that imitation of /æ/ was poor when it was temporally reduced but imitation increased as its duration increased. This might be because participants had difficulty extracting sufficient spectral information from the target stimuli when it was short. Research on acoustic characteristics on English vowels shows that /æ/ is one of the longest vowels in English (Hillenbrand et al., 2000), and listeners might require time to decode the identity of the vowel /æ/ from the speech signal. On the other hand, /ɛ/ is typically much shorter than /æ/ in English and speakers seemed to receive sufficient cueing even with reduced vowel duration. Another possible account for the differential pattern of spectral imitation between /ɛ/ and /æ/ might be due to the articulatory characteristics of /æ/. Low vowels are intrinsically longer than mid or high vowels due to the requirements of greater articulatory movement, in particular, degree of jaw lowering (House & Fairbanks, 1953; Peterson & Lehiste, 1960). Thus, 60 ms might not be sufficient for effective motor execution for the low vowel /æ/, and imitation of vowel duration may have interfered with imitation of vowel spectral quality for these short vowels.

Finally, our results confirmed previous findings (Babel, 2012; Phillips & Clopper, 2012; Walker & Campbell-Kibler, 2015) that greater phonetic distance between the target speech and individual's own productions at baseline increased

imitation scores for imitation of both spectral quality and duration. Previous work made cross-dialectal comparisons of distance (e.g., dialectal background), but the present results further confirm that phonetic imitation is indeed considerably influenced by fine-grained acoustic differences between target and baseline productions. Some of the effect of target-baseline distance on imitation may be due to our definition of degree of imitation rather than imitation itself. Based on our imitation metric, degree of imitation is smaller when baseline and imitation values are very close even if imitations are also very close to the target. With this caveat in mind, however, the present study is generally consistent with previous work in terms of the phonetic distance account, but confirmed it with manipulated target stimuli in a more controlled way. In this context, it is particularly notable that phonologically ambiguous vowel spectral quality does not promote phonetic imitation even though these stimuli are most likely to be acoustically distant from participants' own productions. It should also be noted that US participants showed greater spectral imitation than Canadian participants although target-baseline distance was quite similar between the two groups. This suggests that in addition to the phonetic distance other factors may also play a role in phonetic imitation.

5 Conclusion

The present study explored the link between speech perception and production at an individual level. To that end, this study investigated (1) whether individual listeners' perceptual cue weights are related to their patterns of phonetic imitation and (2) the underlying mechanisms of phonetic imitation. Results indicate that individuals with greater perceptual acuity, as measured by greater spectral weights in perception (which were positively correlated with duration cue weights) showed better imitation of vowel duration, the dimension that was imitated best. This is consistent with previous work showing that higher perceptual acuity is linked to more precise productions (e.g.,

Perkell et al., 2004) and not consistent with work showing a link between use of particular dimensions in perception and production (e.g., Coetzee et al., 2018).

In terms of the mechanisms of phonetic imitation, the results showed an asymmetry between imitation of spectral quality and duration. Ambiguous vowel spectral quality in the target speech did not promote imitation as much as unambiguous vowels, supporting an effect of phonological categorisation on phonetic imitation. In contrast, the findings from imitation of vowel duration were not restricted by phonological categorisation. This may be because these values were perceptually salient; because duration was a secondary cue and therefore threatened the contrast less; or because listeners can more easily imitate a suprasegmental cue (e.g., duration) which they vary in their speech on a regular basis. Further work is needed to separate out these possibilities. Overall, this study suggests that multiple factors modulate phonetic imitation including higher-level linguistic processes (i.e., phonological categorisation) and that imitation is a selective rather than an automatic process (Nguyen & Delvaux, 2015).

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Ainsworth, W. A., & Paliwal, K. K. (1984). Correlation between the production and perception of the English glides /w, r, l, j/. *Journal of Phonetics*, *12*(3), 237–243.

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189.

Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, *55*(2), 231–248.

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology, 5*(1), 123–150.

Bailey, P. J., & Haggard, M. P. (1980). Perception-production relations in the voicing contrast for initial stops in 3-year-olds. *Phonetica*, *37*(5-6), 377–396.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Beddor, P. S., Coetzee, A. W., Styler, W., McGowan, K. B., & Boland, J. E. (2018). The time course of individuals' perception of coarticulatory information is linked to their production: Implications for sound change. *Language*, *94*(4), 931–968.

Boberg, C. (2008). Regional phonetic differentiation in Standard Canadian English. *Journal of English Linguistics, 36*(2), 129–154.

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer (Version 6.0.03).

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R. R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., & Perkell, J. (2011). The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research*, *54*(3), 727–739.

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*(6), 633–644.

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., & Wissing, D. (2018). Plosive voicing in Afrikaans: Differential cue weighting and tonogenesis. *Journal of Phonetics*, *66*, 185–216.

Delaney, M., Savji, S., & Babel, M. (2010). An acoustic and auditory comparison of implicit and explicit phonetic imitation. *Canadian Acoustics*, *38*(3), 132–133.

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2-3), 145–173.

D'Imperio, M. & German, J. S. (2015). Phonetic detail and the role of exposure in dialect imitation. *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK.

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*.

Escudero, P. (2000). Developmental patterns in the adult L2 acquisition of new contrasts: The acoustic cue weighting in the perception of Scottish tense/lax vowels in Spanish speakers. Unpublished M. Sc. thesis, University of Edinburgh

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *Journal of the Acoustical Society of America*, *83*(2), 729–740.

Fox, R. A. (1982). Individual variation in the perception of vowels: Implications for a perception-production link. *Phonetica*, *39*(1), 1–22.

Frieda, E. M., Walley, A. C., Flege, J. E., & Sloane, M. E. (2000). Adults' perception and production of the English vowel /i/. *Journal of Speech, Language, and Hearing Research*, *43*(1), 129–143.

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*, 1–9.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Hay, J., Drager, K., & Gibson, A. (2018). Hearing *r*-sandhi: The role of past experience. *Language*, *94*(2), 360–404.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America*, *108*(6), 3013–3022.

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

Honeybone, P., & Watson, K. (2013). Salience and the sociolinguistics of Scouse spelling: Exploring the phonology of the Contemporary Humorous Localised Dialect Literature of Liverpool. *English World-Wide*, *34*(3), 305–340.

Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics*, *39*(1), 18–38.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, *25*(1), 105–113.

Iverson, P., & Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America*, *99*(2), 1130–1140.

Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, *7*.

Kawahara, H., Takahashi, T., Morise, M., & Banno, H. (2009). Development of exploratory research tools based on TANDEM-STRAIGHT. *Proceedings of Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference* (pp. 111–120).

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *Journal of the Acoustical Society of America*, *124*(6), 3959–3971.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2015). *lmerTest*. URL http://cran.r-project.org/web/packages/lmerTest/index.html.

Kwon, H. (2015). Spontaneous speech imitation and cue primacy. *Proceedings of the 18th International Congress of the Phonetic Sciences*, Glasgow, UK.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Walter de Gruyter.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798.

MacLeod, B. (2015). A critical evaluation of two approaches to defining perceptual salience. *Ampersand*, *2*, 83–92.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173.

Mitterer, H., & Müsseler, J. (2013). Regional accent variation in the shadowing task: Evidence for a loose perception-action coupling in speech. *Attention, Perception, & Psychophysics, 75*(3), 557–575.

Morrison, G. S. (2005). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero and Boersma (2004). *Studies in Second Language Acquisition*, *27*(4), 597–606.

Morrison, G. S. (2007). Logistic regression modelling for first- and second-language perception data. In P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 219–236). Amsterdam: John Benjamins.

Morrison, G. S., & Kondaurova, M. V. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis (L). *Journal of the Acoustical Society of America*, *126*(5), 2159–2162.

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report. *Journal of the Acoustical Society of America*, *113*(5), 2850–2860.

Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, *53*, 46–54.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142.

Nielsen, K. (2014). Phonetic imitation by young children and its developmental changes. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2065–2075.

Nielsen, K., & Scarborough, R. (2015). Perceptual asymmetry between greater and lesser vowel nasality and VOT. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*(1), 94–140.

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*, 63–79.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*(3), 183–195.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*, *116*(4), 2338–2344.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*(6), 693–703.

Phillips, S., & Clopper, C. G. (2011). Perceived imitation of regional dialects. *Proceedings of Meetings on Acoustics,12*(1).

Podlipský, V. J., & Simáčková, S. (2015). Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK.

Postma-Nilsenová, M., & Postma, E. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology*, *4*.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rácz, P. (2013). *Salience in Sociolinguistics*. Berlin, Boston: Walter de Gruyter.

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research* (pp. 379–410). Baltimore: York Press.

Savill, N., Ellis, A. W., & Jefferies, E. (2017). Newly-acquired words are more phonologically robust in verbal short-term memory when they have associated semantic representations. *Neuropsychologia*, *98*, 85–97.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183– 204.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429.

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America*, *132*(2), EL95–EL101.

Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, *37*(3), 276–296.

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, *6*.

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PloS ONE*, *8*(9), e74746.

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13–29.

Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience*, *32*(6), 776–791.

Zellou, G., Scarborough, R., & Nielsen, K. (2013). Imitability of contextual vowel nasalization and interactions with lexical neighbourhood density. *Proceedings of Meetings on Acoustics, 19*(1).

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *Journal of the Acoustical Society of America*, *140*(5), 3560–3575.

Zetterholm, E. (2006). Detection of speaker characteristics using voice imitation. In C. Müller and S. Schötz (eds.). *Speaker Classification*. Springer LNCS/LNAI series.