Phonological Knowledge in Speech Perception: The Case of Illusory Consonants

Karthik Durvasula[1], Chenchen Xu[1], Mingzhe Zheng[3], Xiaomei Wang[2], and Yen-Hwei Lin[1]

[1]Michigan State University, USA
[2]Tianjin Normal University, China
[3]Earlham College, USA

Corresponding Author Email: durvasul@msu.edu

Abstract

Native listeners perceive illusory sounds, typically when presented with sound sequences that do not respect the phonotactic constraints of their language. Prior work on such illusions has focused on, what we call, auditory ILLUSIONS OF QUANTITY related to vowels (syllabic nuclei) in CC sequences that violate word-internal phonotactic constraints. Here, we focus on CV sequences, which are illicit word-internally in Mandarin Chinese, but not in American English, to show that ILLUSIONS OF QUANTITY related to *consonants* (non-nuclear segments) are indeed possible, provided the phonology of the language supports such illusions. Our results provide further evidence for the viewpoint that sees the task of the perceiver during speech perception as identifying the best parse of the intended underlying (or phonemic) representations of the utterance given the acoustic token; a view that naturally predicts the involvement of phonological knowledge and phonetic factors.

1 Introduction

The study of speech perception has greatly benefitted in recent decades from the identification of the phenomenon of auditory illusions. The phenomenon and the specific patterns of illusions have allowed us to better understand how both phonetic factors and phonological knowledge play a role in speech perception. In this article, we specifically set out to test some predictions that naturally fall out of a view that, during speech perception, the listener attempts to reverse infer the best estimate of the intended underlying representations of the utterance given their phonological/phonetic knowledge and the acoustics of the utterance (Durvasula and Kahng 2015; Gaskell and Marslen-Wilson 1996, 1998; Gow 2003; Mitterer et al. 2013, amongst others).

Most previous research related to auditory illusions has focussed on the issue of illusory vowels or syllabic nuclei[1] (Berent et al. 2008, 2009, 2007; Davidson 2007; Davidson and Shaw 2012; Dupoux et al. 1999, 2011; Durvasula et al. 2018; Kabak and Idsardi 2007; Yun 2016; Zhao and Berent 2016, amongst others). For example, when a Japanese listener is auditorily presented with stimuli such as [ebzo], where the consonant sequence is either naturally produced or created by splicing out the medial vowel completely from productions such as [ebizo] or [ebazo], they may actually perceive /ebuzo/ with an illusory /u/, given that [bz] is an illicit consonant sequence in Japanese, as shown originally by Dupoux et al. (1999).

The most convincing evidence of the involvement of phonological knowledge during auditory illusions is that listeners with different native language backgrounds perceive the same auditory event different, suggesting that the illusion does not stem solely from the acoustics of the auditory input, but instead also has some basis in the language-specific (phonological) knowledge of the listener. Continuing the example from above, Dupoux et al. (1999) show that, in contrast to Japanese speakers, French speakers correctly identified far fewer (if any) illusory vowels for the same stimuli. This suggests that the illusory vowels heard by the Japanese speakers were a language-specific effect and not simply due to fine phonetic detail of the stimuli. Of course, the presence of language-specific perception is not new, and in at least some cases can be traced to different ranges of phonetic parameters associated with categories. For example, if two languages have roughly the same categories, but differ in the amount of category overlap in phonetic space (*e.g.*, if either the means or the variances of the relevant phonetic categories are different), it is

[1] Throughout, we will use 'vowel' to refer to a 'syllabic nucleus'. Similarly, we will use 'consonant' to refer to a 'non-nuclear segment'. We recognise that the terms 'vowel' and 'consonant' are not necessarily isomorphic with syllablehood for all analysts, but we maintain the use of the terms for ease of exposition.

possible for the listeners of those languages to hear the same auditory input in different ways, contingent on the decision thresholds stemming from the differences in the category overlap in phonetic space. What sets the auditory illusions of interest here apart is that they cannot simply be traced back to differences in the decision thresholds over phonetic parameters associated with categories, as they typically involve, what we term in this paper, ILLUSIONS OF QUANTITY, whereby there is a difference in the number of segments perceived, for the same auditory input, based on the native language.

It is important to note that there are some systematic aspects to the contexts used in such experiments probing ILLUSIONS OF QUANTITY: (a) the context of the illusion is typically interconsonantal (CC), (b) the ILLUSION OF QUANTITY involves the perception of an additional *vowel* by one set of native language speakers but not by another. However, nothing in the reverse inference view, briefly mentioned above, being testing in this article suggests that auditory illusions (or more specifically, ILLUSIONS OF QUANTITY) have to involve these two aspects. In line with this expectation, we show in this article that it is indeed possible to get ILLUSIONS OF QUANTITY involving *consonants* in CV contexts, as long as the listener's native language has the appropriate set of phonological conditions for it.

There has been some important prior work observing auditory illusions related to consonants. Such work typically involves observing that a segment is misperceived as another in a phonotactically illicit context. We discuss two strands of such research below.

First, it has been observed that listeners compensate for assimilatory patterns in their native language, by undoing expected patterns of assimilation from the input, and recovering the plausible source segment before the assimilation (Darcy et al. 2009; Gaskell and Marslen-Wilson 1996, 1998; Gow 2003; Mitterer et al. 2013, amongst others). For example, the phrase "garden bench" /gɑːdn bɛntʃ/ often manifests acoustically as something very close to, and sometimes identical to, [gɑːdm bɛntʃ],[2] where the word-final nasal /n/ has (at least, partially) assimilated to the place of articulation of the following segment. It has been observed that listeners are able to compensate for such coarticulatory changes, i.e., when presented with an assimilated variant (e.g., [gɑːdm]), listeners are able to recognize the unassimilated word (e.g., "garden"), but only when the nasal consonant is followed by a word that begins with a bilabial sound (e.g., [gɑːdm bɛntʃ]). Again the most convincing evidence for the language specific nature of this effect comes from across-language comparisons looking at compensations to voicing and place of articulation assim-

[2]Throughout, we notate phonemic representations with /…/, and acoustic output/input with […].

ilations by French and English listeners (Darcy et al. 2009). They show that while English speakers compensate for place of articulation assimilation (which is present in the language), they do not compensate for voicing assimilation (which is not present in the standard varieties). In contrast, French speakers compensate for voicing assimilation (which is present in the language), but not for place of articulation assimilation (which are not present in the language).

Second, it has been observed that in phonotactically illicit consonant sequences, listeners misperceive the quality of one of the consonants (Hallé and Best 2007; Hallé et al. 1998; Moreton 2002). For example, /dl/ and /tl/ are phonotactically illicit word-initial (or syllable-initial) sequences in French, and such sequences are typically perceived as /gl/ and /kl/ by French speakers. As with the above, the most convincing results of the language-specific nature of the effect are across-language comparisons that show that French and English speakers, for whom such sequences are phonotactically illicit, mis-perceive such sequences more so than Hebrew speakers, for whom the same sequences are phonotactically licit (Hallé and Best 2007).

Both the above sets of results can be viewed as consonantal illusions stemming from phonotactically illicit consonantal sequences in the listener's native language. However, it is fair to describe such illusions as ILLUSIONS OF QUALITY, whereby a segment is misperceived as another segment. What has not been observed yet, to our knowledge, for consonantal cases is an ILLUSION OF QUANTITY, whereby there is a difference in the number of consonants perceived based on the native language. In this paper, we argue that consonantal ILLUSIONS OF QUANTITY are possible, provided appropriate phonotactic contexts/patterns are tested.

Furthermore, all the consonantal illusions discussed above are typically in illicit consonantal (CC) sequences. Given that these are the same contexts in which illusory vowels are observed, it raises the question of whether there is something special to such CC contexts. In this paper, we argue that (consonantal) illusions are indeed also possible in phonotactically illicit CV sequences, and that there is nothing necessarily special about CC contexts triggering auditory illusions.

Overall the results presented in this paper further support the reverse inference view of speech perception that was mentioned before and is explicated more carefully below. However, before laying out the expectations for the experiments discussed in this article, it is important to present our conception of the nature of the problem that is being solved during speech perception (following Marr (1982)). We assume that the task of the listener in speech perception is primarily a task of reverse inference - it is to identify the best estimate of the intended underlying (or phonemic) representations of the utterance given the acoustic token (Durvasula et al. 2018; Durvasula and

Kahng 2015; Gaskell and Marslen-Wilson 1996, 1998; Gow 2003; Mitterer et al. 2013, amongst others). This view parallels Bayesian models of speech perception (Bever and Poeppel 2010; Feldman and Griffiths 2007; Poeppel and Monahan 2011; Sonderegger and Yu 2010; Wilson and Davidson 2013), which typically involve perception as reverse inference from acoustic input to surface representations. As a consequence of this viewpoint, knowledge about both phonological alternations and phonotactic constraints is required to reverse-infer the phonemic/underlying representations from the acoustic tokens.

It is important to note that the recruitment of phonological knowledge is a *necessary* aspect of speech perception, if the perceiver is trying to identify the best estimate of the intended 'underlying representations'. This is because, except in trivial cases where the underlying representations 'match' the surface representations and are directly inferable from the acoustics, there are many cases where the underlying representation is not directly inferable from the acoustics, without knowledge of the phonology, *e.g.*, in many dialects of English, a word-final /t/ surfaces as a glottal stop in coda positions (Wells 1982, 1990), but in many dialects of Malay, a word-final /k/ surfaces as a glottal stop in coda positions (Omar 1977/1991; Yanti 2010); therefore, the inference from the acoustic input to the correct intended underlying representation *has* to use the phonology of the language; otherwise, such an inference is not possible.

A type of process that is likely to bias a listener's expectation about the upcoming segment quality is one that involves a phonotactic restriction, whereby $[C_1]$ is allowed only next to the sound $[X_2]$ (*i.e.*, *C_1X, where X $\neq$ X_2).[3] If the listener is auditorily presented with a consonant $[C_1]$ in a context where it is not phonotactically licit, and if the sequence can be repaired phonotactically by a segment, then the best segment to infer to perceptually 'repair' the sequence, provided the acoustics of the stimuli allow for such an illusion, is the segment /X_2/. This would account for the acoustic properties of the illicit consonant [C_1] while satisfying the phonotactic requirement in the language that [C_1] can only appear before the sound [X_2]. This is the type of process that we test in this article.

1.1 Relevant phonological patterns and predictions

In Mandarin, obstruent consonants cannot be codas. Furthermore, while alveolar stops as a group can precede all vowels and glides [✓t^{h}i, ✓t^{h}a, ✓t^{h}u, t^{h}j, ...]),[4] alveo-palatal consonants can only

[3] This can also be extended to all allophonic mappings before particular segments/features.

[4] /t^h/ specifically seems to be absent before /e/.

appear before high front vowels or front/palatal glides [✓tɕʰi, ✓tɕʰj, *tɕʰa, *tɕʰu, …]. These facts allow the front/palatal glides, particularly /j/, to be a good consonant for perceptual repairs in contexts where an alveo-palatal consonant (but not an alveolar/dental consonant) is immediately followed by a vowel which is not a high front vowel. We therefore expect /j/ to be an illusory consonant in Mandarin. (Note, there is a fair amount of prior literature discussing the segmental status of glides in Mandarin; we return to this issue in section 5.1).

Mandarin also has a rounded palatal glide /ɥ/, and alveopalatal consonants are allowed before them. While nothing directly stemming from the viewpoint presented above (and discussed further below) motivates a more specific prediction than a palatal glide, one could argue based on markedness considerations that that palatal glide must be the unrounded /j/. Of course, it is possible that /j/ is preferred for acoustic reasons; if so that is consistent with the reverse inference view that we suggest here. However, more precise acoustic recordings and models are needed to sustain such a claim. Furthermore, word medially, the rounded palatal glide [ɥ] appears only before /ɛ/ (Duanmu 2007). For simplicity, we chose only the [au] and [a] vowels for the stimuli in our experiments, therefore the stimuli were of the template [aC(G)auma]. As a consequence, at least in the current experiments, it is unlikely that /ɥ/ will be perceived in any of the stimuli presented simply because [ɥ] is not allowed before [a] (*[tɕʰɥau], *[tʰɥau]). In contrast, as discussed above, /j/ is allowed in such sequence ([tɕʰjau], [tʰjau]). Finally, testing the perception of the presence/absence of [ɥ] in our stimuli would have been difficult for the current series of experiments where American English speakers were used as controls, as there is no equivalent phoneme that they could map the sound to. For the above reasons, we leave the possibility of /ɥ/ for future work.

In contrast, in American English, both /alveolar stop + j/ and /palato-alveolar + j/ sequences are not possible within words. In fact, there are only 82 such words with the sequence /…tʰj…/ in the CMU Dictionary (Weide 1994), and most of them are either loanwords/names (*e.g.*, Katya, Satya,…), or very uncommon alternative pronunciations in Standard American English (*e.g.*, Tuesday, Tuition,…), or compounds (*e.g.*, boatyard, courtyard,…). Similarly, there are only 9 words with the sequence /…tɕʰj…/ in the CMU Dictionary. Seven of them are foreign loanwords/names (*e.g.*, Altschuler, Cheung,…), one is a very uncommon alternative pronunciation in Standard American English (statutorily), and one compound word (churchyard). However, both such sequences (…tʰj…/ and /…tɕʰj/…) are allowed across words and across members of a compound (*e.g.*, churchyard, boatyard,…). So, while there are legitimate phonemic parses to both sequences in

English, they both involve a word-boundary.[5]

Based on the above patterns in American English and Mandarin and the view of speech perception laid out earlier, we can make predictions about what Mandarin and American English speakers should perceive; these predictions are diagrammatically laid out in Figure 1. We will first discuss the expected perceptual patterns for Mandarin speakers and then American English speakers.

When Mandarin speakers are auditorily presented with tokens containing alveolar stops followed by a vowel or a glide, [at^{h}auma] or [at^{h}jauma], given that both the sequences are possible word-internally in their language, they should perceive them veridically, *i.e.*, as /at^{h}auma/ and /at^{h}jauma/, respectively. Similarly, when Mandarin speakers are auditorily presented with tokens containing an alveo-palatal stop followed by palatal glide, [at$\c{c}^{h}$jauma], they should perceive it veridically, *i.e.*, as /at$\c{c}^{h}$jauma/. However, when they are auditorily presented with a token containing an alveo-palatal stop followed by a non-high front vowel or palatal glide, [at$\c{c}^{h}$auma], given the sequence /t$\c{c}^{h}$au/ is not possible in their language, they are expected to hear a perceptual illusion. The inference of a palatal glide /j/ after the alveo-palatal stop is the smallest 'repair' that can make the sequence licit in their language; therefore, they are expected to perceive the token as /at$\c{c}^{h}$jauma/.

Based on the above, we further expect that Mandarin speakers should confuse [at$\c{c}^{h}$auma] and [at$\c{c}^{h}$jauma], but not [at^{h}auma] and [at^{h}jauma]. It is worth noting that there might be a few instances where Mandarin speakers hear [t$\c{c}^{h}$] as /t$\textsubarch{s}^{h}$/, because it is an acoustically proximal category. In such cases, there is no expectation of any illusions as /t$\textsubarch{s}^{h}$au/ is a licit phonotactic sequence in Mandarin Chinese. However, given that [t$\c{c}^{h}$] is much more likely to be perceived as /t$\c{c}^{h}$/, we expect the above predictions about the differences in the overall rates of consonantal illusions between alveolar and alveopalatal contexts to stand.

When English speakers are presented with the sequences, as we will see below, we expect no illusory consonants, though the relevant sequences are *not* allowed within words, as discussed above. Note, English does not have an alveo-palatal consonant such as [t$\c{c}^{h}$], but there are palato-alveolar consonants such as [t$\int^{h}$].[6] So, when an English speaker is presented with [t$\c{c}^{h}$], they are likely to perceive it as /t$\int^{h}$/ - this is expected based on the results of non-native

[5]Here, we do not distinguish between Morphological word-boundaries, and Phonological/Prosodic word-boundaries (Nespor and Vogel 1986; Peperkamp 1999). These constituents have been argued by many to not be isomorphic. However, the distinction is not relevant for current purposes.

[6]Following Honeybone (2005) and Iverson and Salmons (1995), throughout this article, we assume that English voiceless stops are [+spread glottis], so we notate them with a superscript [^{h}].

perception discussed by Best et al. (2003). As a consequence, when English speakers are auditorily presented with tokens containing an alveolar/alveo-palatal consonant followed by a vowel, [at^{h}auma] or [at$\c{c}^{h}$auma], given the sequences are licit within words, they should perceive them veridically, *i.e.*, as /at^{h}auma/ and /at$\int^{h}$auma/, respectively. When English speakers are presented with corresponding tokens that include palatal glides after the relevant consonants, [at^{h}jauma] or [at$\c{c}^{h}$jauma], they cannot parse the sequences as single-words, since the sequences are disallowed in the language within words. However, such sequences are allowed across words, and therefore an English speaker could still perceive the segments veridically, by inferring a word-boundary between the relevant consonants and the following palatal glide, *i.e.*, /at^{h}#jauma/[7] or /at$\int^{h}$#jauma/ (here, we assume the use of phonological knowledge outside the word; however, this prediction of the reverse inference viewpoint whereby the illusions are conditional on knowledge about higher phonological domains was much more explicitly tested in Durvasula and Kahng (2016)). As a consequence of the above facts, an English speaker is not expected to hear any perceptual illusions in the relevant tokens, even though they contain sequences that are disallowed word-internally in the language. We further expect the English speaker to better discriminate [at$\c{c}^{h}$auma$\sim$at$\c{c}^{h}$jauma], compared to Mandarin speakers.

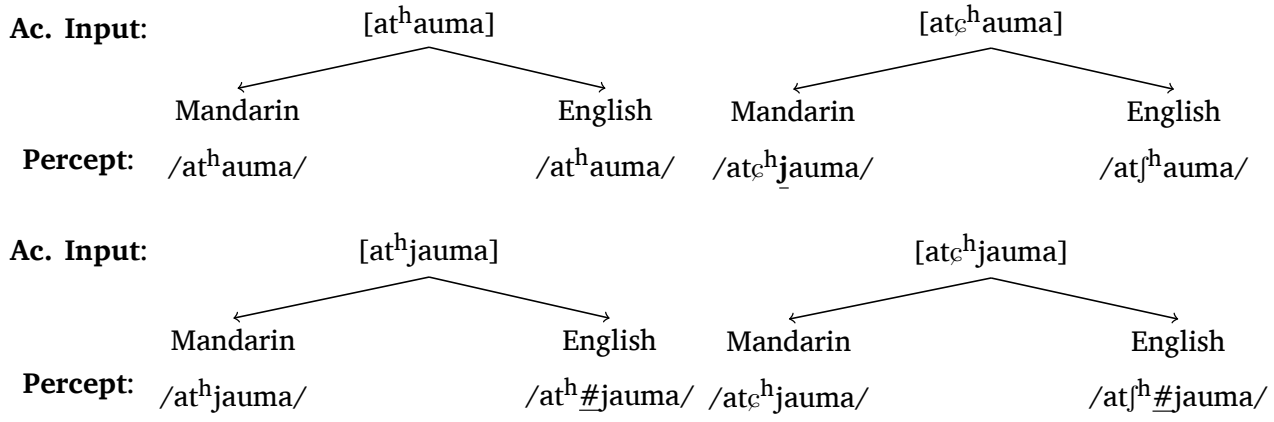

Figure 1: Predicted possible percepts for different acoustic inputs

It is important to note that the above predictions stand when there are no other influences or acoustic artefacts to consider. However, pairs of segments and segment sequences are likely to have some degree of inherent confusability, *i.e.*, some pairs of segments/sequences are likely to be more confusable than others for acoustic reasons inherent to the stimuli. Therefore, the above predictions stemming from the phonological systems of English and Mandarin can only be

[7] We use # to notate a word-boundary.

treated as *mutatis mutandis* predictions. In order to test the predictions, we cannot simply conduct within-language comparisons, as they conflate inherent auditory confusability with confusability stemming from language-specific factors. For this reason, the crucial predictions are between-language predictions. In short, we expect Mandarin speakers to hear *more* illusory /j/ than American English speakers when presented with [at$\c{c}^h$auma], but not when presented with [at^{h}auma]. Similarly, in discrimination tasks, we expect Mandarin speakers to confuse [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma] more than English speakers, but to show a similar level of confusability as the English speakers with [at^{h}auma$\sim$at^{h}jauma].

We test these predictions below in a series of three experiments. In Experiment 1, we use an ABX task to probe the differences in confusability of different pairs of nonce-word between Mandarin and American English speakers. As expected, Mandarin speakers confuse the pair [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma], but not [at^{h}auma$\sim$at^{h}jauma], more than American English speakers. In Experiment 2, we used a Yes/No task to see if Mandarin and American English speakers hear different rates of /j/ in different test items containing alveolar/alveo-palatal consonants followed by vowels. Again, the crucial expectation is met in that Mandarin speakers hear more illusory /j/ than American English speakers when presented with [at$\c{c}^h$auma], but not when presented with [at^{h}auma]. Finally, in Experiment 3, the alveolar consonants in Experiment 2 were replaced by dental consonants, because there is some debate in the phonological literature on Mandarin about the place of articulation of the relevant coronal series; we replicate the crucial results of Experiment 2 in Experiment 3.

2 Experiment 1

Experiment 1 probed for illusory consonants through an ABX task, in which participants heard sets of three stimuli and had to decide if the third stimulus was more like the first or second stimulus. As laid out in the *Introduction*, we expect Mandarin speakers to confuse the pair [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma], but not the pair [at^{h}auma$\sim$at^{h}jauma], more than American English speakers. Therefore, in an ABX paradigm, we expect Mandarin speakers to have lower accuracy rates, than American English speakers, when the first two stimuli are [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma], but not when the first two stimuli are [at^{h}auma$\sim$at^{h}jauma]. The results we present in the following sub-sections match these predictions.

2.1 Methods

2.1.1 Participants

Twenty native Mandarin speakers from Beijing (mean age = 20.5 years, SD = 2.2 years, 7 men and 13 women) and 19 native English speakers from Michigan as a control group (mean age = 21 years, SD = 2.7 years, 6 men and 13 women) participated in the experiment. All the participants were recruited at Michigan State University. The Mandarin participants received no compensation for their participation, but the English participants received extra credit for their participation. The Mandarin participants spent an average of 2.4 years (SD = 1.6 years) in the US before the experiment, primarily as undergraduate students at Michigan State University. Furthermore, the Mandarin speakers also reported an average of 14.1 years (SD = 2.6 years) of exposure to English as a second language in a classroom environment in China.

2.1.2 Materials

All the test items followed the template aC_1V_1ma, in which C_1 was an alveolar or alveo-palatal consonant [t^h / $tɕ^h$]; and V_1 was [au / jau / ∅ (Null)]. None of the stimuli were words in either Mandarin or in English. All the stimuli had stress on the first vowel with a high-high-low tone sequence on the 3-syllable nonce words and a high-low tone sequence on the 2-syllable nonce words. They were natural recordings by a trained male phonetician (the first author), who is a native speaker of Indian English and Telugu, and a second-language speaker of standard Hindi.

There were two reasons for the use of this particular speaker. Firstly, he could naturally produce all the stimuli, as they are phonotactically licit in his dialects of both Hindi and Telugu (particularly, across words). The use of a native Mandarin speaker to record the stimuli would have only been possible if the speaker had neutralised their own linguistic biases, some of the sequences are not licit in the language. We strongly suspect that the use of Mandarin speakers to record stimuli would have introduced biases into the stimuli, especially for those sequences that are not licit in the relevant language, thereby making the interpretation of the results much more challenging. Secondly, the use of an American English speaker to record the stimuli was also avoided, because it would be a challenge for native American English speakers to produce Mandarin alveo-palatals. Furthermore, we did not want to introduce a bias that would help the control group, as the overall phonetic patterns would have been more natural for the American English speakers than for the Mandarin speakers. The interpretation of the crucial between-language

results could therefore have been potentially confounded by this. For these reasons, we used the first author's voice for recording stimuli. (Note: If our objective were trying to understand American English loanwords in Mandarin, then using an American English speaker would have been necessary. However, this is not our focus.)

There are two more issues worth discussing in detail with respect to the stimuli and the speaker. First, the alveo-palatal pronuciations by the speaker were reasonably close to native Mandarin pronunciations. One of the speaker's native languages, Telugu, has sounds which are be reasonably close, acoustically, to alveo-palatal stops (note: it also has dental stops; a fact that is relevant for Experiment 3). However, we depended on four of the co-authors (who are native Mandarin speakers), and four more native Mandarin speaking members of the phonology-phonetics lab in our department to first vet all the stimuli for naturalness, particularly with respect to the crucial consonants' place and manner of articulations. And, only after the relevant members and co-authors were satisfied with the stimuli did we proceed with experimentation. Second, a similar issue arises with the naturalness of the tones, particularly that of the low tone used. Here, It is worth noting that some phonological analyses consider Tone 3 in Mandarin, traditionally transcribed as a low-dipping-rising tone, to be underlyingly low (Duanmu 1999, 2007; Yip 1980). In fact, Duanmu (1999, p. 14) suggests that this tone largely has a low pitch contour, and is best described as either 211 or 11 in the Chao system (where, 5 is the highest pitch, and 1 is the lowest pitch). Furthermore, a Tone 3 in the surface representation is consistently pronounced as a low tone before another tone (Chao 1968; Duanmu 2007; Lin 2007). Finally, in many Mandarin speakers' speech, the final rise of Tone 3 is absent even in final position, hence there is just a low tone in final position (Duanmu 2007; Lin 2007). Therefore, both the low tone and high tone used in this experiment could very well be reasonably natural for Mandarin speakers. Having said the above, our original intention was *not* to use Mandarin tones, but to use tones consistent with both previous/future experiments in our lab. It is possible that the use of non-Mandarin tones might have resulted in confounds for the Mandarin speakers; however, for reasons discussed towards the end of the *Introduction*, the important comparisons to control for acoustic artifacts in the stimuli are between-language comparisons. As a consequence, using perfectly phonetically matched Mandarin tones that have no correspondents in American English, might well have introduced confounds into the control group's (English speakers') responses, given that pitch height difference and contours do play a role in the English stress/intonational system. To us, there is no immediately obvious way of solving this problem, and therefore the best way to proceed is to

use tones in a consistent fashion, and look for tonal interactions in future experiments that are focussed on such effects; we particularly chose high-low and high-high-low tone sequences in order to mirror a natural declination in pitch. Here again, we made sure that the stimuli sounded reasonably natural to the Mandarin-speaking co-authors and other Mandarin (and English speaking) members of our lab before we proceeded to running experiments.

Each item was recorded several times using Praat (Boersma and Weenink 2016) with a microphone (Logitech USB Desktop Microphone; Frequency Response – 100Hz-16KHz) at a 44KHz sampling rate (16-bit resolution; 1-channel). The stimuli were normalized in Praat to have a mean intensity of 70dB SPL. From these recordings, two tokens were selected for each item and they were each presented twice; therefore, there were 12 tokens in the experiment.

2.1.3 Procedure

We used an ABX task to investigate the expected consonantal auditory illusion. There were two recordings used for each item and the order of tokens in an AB sequence was counterbalanced. We explain the combinatorics below using the case of [at^{h}auma$\sim$at^{h}jauma].[8] There were four AB sequences [at^{h}auma$_{1}$-at^{h}jauma$_{1}$], [at^{h}auma$_{1}$-at^{h}jauma$_{2}$], [at^{h}auma$_{2}$-at^{h}jauma$_{1}$], [at^{h}auma$_{2}$-at^{h}jauma$_{2}$], and an additional four word-pairs in reversed order. To each of these AB sequences, either A or B was added as an X. When adding X's, the same token was never repeated in a single trial.

Therefore, in the case of [at^{h}auma$\sim$at^{h}jauma], there were eight ABA triplets , and and an additional eight ABB triplets as shown below (Table 1).

[8] We use '$\sim$' to notate a pair irrespective of order, and '-' to notate a particular ordered sequence. The numeric subscript in what follows represents the stimulus version, since two recordings were used for each item.

Comparison	Triplet
ABA	[at^{h}auma$_{1}$-at^{h}jauma$_{1}$-at^{h}auma$_{2}$],
	[at^{h}auma$_{1}$-at^{h}jauma$_{2}$-at^{h}auma$_{2}$],
	[at^{h}auma$_{2}$-at^{h}jauma$_{1}$-at^{h}auma$_{1}$],
	[at^{h}auma$_{2}$-at^{h}jauma$_{2}$-at^{h}auma$_{1}$],
	[at^{h}jauma$_{1}$-at^{h}auma$_{1}$-at^{h}jauma$_{2}$],
	[at^{h}jauma$_{1}$-at^{h}auma$_{2}$-at^{h}jauma$_{2}$],
	[at^{h}jauma$_{2}$-at^{h}auma$_{1}$-at^{h}jauma$_{1}$],
	[at^{h}jauma$_{2}$-at^{h}auma$_{2}$-at^{h}jauma$_{1}$]
ABB	[at^{h}auma$_{1}$-at^{h}jauma$_{1}$-at^{h}jauma$_{2}$],
	[at^{h}auma$_{1}$-at^{h}jauma$_{2}$-at^{h}jauma$_{1}$],
	[at^{h}auma$_{2}$-at^{h}jauma$_{1}$-at^{h}jauma$_{2}$],
	[at^{h}auma$_{2}$-at^{h}jauma$_{2}$-at^{h}jauma$_{1}$],
	[at^{h}jauma$_{1}$-at^{h}auma$_{1}$-at^{h}auma$_{2}$],
	[at^{h}jauma$_{1}$-at^{h}auma$_{2}$-at^{h}auma$_{2}$],
	[at^{h}jauma$_{2}$-at^{h}auma$_{1}$-at^{h}auma$_{1}$],
	[at^{h}jauma$_{2}$-at^{h}auma$_{2}$-at^{h}auma$_{1}$]

Table 1: Sample stimulus set for [at^{h}auma∼at^{h}jauma]

So, there were a total of 48 triplets constructed from the alveolar stimuli ([at^{h}ma, at^{h}auma, at^{h}jauma]). Similar combinations were used to create another 48 triplets for the alveo-palatal stimuli ([atɕ^{h}ma, atɕ^{h}auma, atɕ^{h}jauma]). This amounted to a total of 96 trials in the experiment, presented in pseudo-randomized order with the added constraint that there be no identical triplets in succession. Note, we included the pairs [at^{h}jauma∼at^{h}ma], [at^{h}auma∼at^{h}ma], [atɕ^{h}jauma∼atɕ^{h}ma] and [atɕ^{h}auma∼atɕ^{h}ma] for experimental reasons to ensure a reasonable number of clearly different stimuli, and as a sanity check. With respect to the latter concern, if either the Mandarin or the English speakers showed low number of correct responses, that would be extremely surprising.

The experiment was conducted in a quiet room with a group of 4-6 participants per session. The stimuli were presented with a low-noise headset (Koss R80 headphones) to each participant through an ABX task scripted in Praat (Boersma and Weenink 2016). The participants were asked

to listen to the triplets of stimuli and determine whether the last sound was more similar to the first or the second and click on the corresponding box (1 or 2) on the screen with a mouse. All the instructions were in English for the English speakers ("Is the last sound more similar to the first or the second?") and in Mandarin for the Mandarin speakers ("第三段音和第一段音还是第二段音更相似?"). The experiment started with a practice session to ensure familiarity with the task. The practice session had 12 trials with another set of nonce words (where the C_1 was [m]). The inter-stimulus interval was 500ms and the inter-trial interval was 1500ms. All 96 trials were randomized for each participant. The experiment took about 7–8 minutes.

2.2 Results

The data munging and plots presented throughout this article were done in R (R Development Core Team 2014) using functions from `tidyverse` packages (Wickham 2017). Furthermore, for each of the experiments presented in this article, the complete set of by-participant averages for each stimulus are presented in the Appendix.

A visual inspection of the mean percentage of correct responses to the stimuli by both the Mandarin and English speakers (Figure 4) suggests the following: (a) the Mandarin speakers appear to be worse at distinguishing [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma] (mean difference in accuracy = 17.7%); (b) the Mandarin and English speakers had similar responses for other test pairs.

In order to confirm the observations made by visual inspection of the results, we followed up with statistical analysis. Note, the crucial comparisons are pairwise between-language comparisons for the same stimuli, and throughout, we use ANOVAs to model the overall patterns instead of more complex mixed-effects models in the interest of simplicity of presentation.

A two-way mixed ANOVA was run, using the `ez` package (Lawrence 2015), to model the percentage of correct responses as the dependent variable with LANGUAGE (English, Mandarin) as a between-subjects factor and COMPARISON (all 6 comparison pairs) as a within-subject factor. Mauchly's test revealed violations of the assumption of sphericity for the main effect of COMPARISON and the interaction LANGUAGE*COMPARISON. Both effects were corrected with Greenhouse-Geisser correction for the degrees of freedom ($\epsilon = 0.45$). There was a main effect of COMPARISON [$F(5,185) = 54.1$, $p < 0.0001$, $\eta^2_{gen} = 0.47$]. Crucially, there was a two-way interaction of LANGUAGE and COMPARISON [$F(5,185) = 7.1$, $p < 0.001$, $\eta^2_{gen} = 0.10$]. The interaction suggests that the Mandarin and English speakers had different responses to different comparison pairs.

To investigate the results from the ANOVA further, we conducted pairwise Mann Whitney

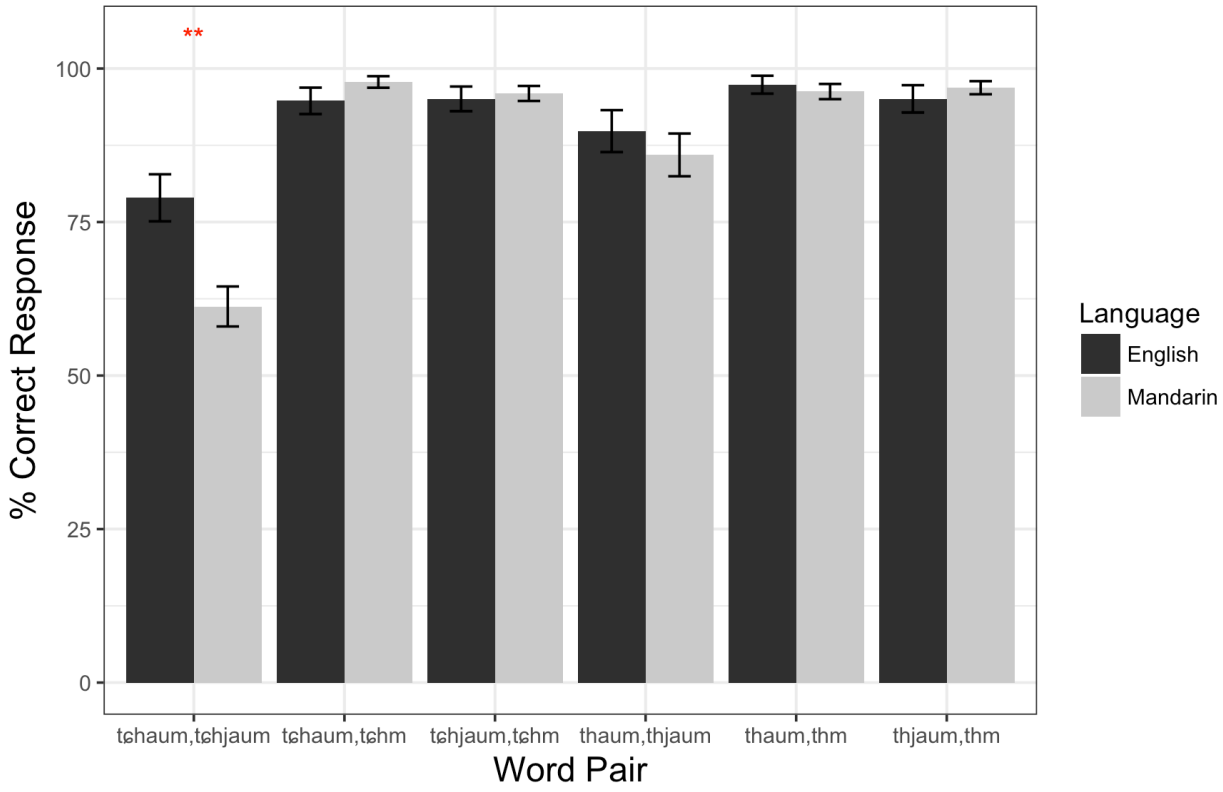

Figure 2: Mean % correct for English and Mandarin speakers in Exp. 1 (error bars = 1 S.E.; * = $p < 0.05$; ** = $p < 0.01$; asterisks represent p-values for between-language comparisons)

U tests (Table 2). These non-parametric tests were conducted as the dependent variable was a proportion, and hence the assumption of normality of errors made by t-tests was violated. Further note, ANOVAs are typically understood to be more robust to such normality violations, so we use them above to maintain ease of interpretability for the reader. As discussed in detail in the *Introduction*, the crucial comparisons are between-language comparisons, as such comparisons control for any artifacts in the stimuli. The pairwise tests suggest that Mandarin speakers were confusing the predicted pair [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma] more than the English speakers. It is also worth noting, for the same pair, that the English speakers while clearly better than the Mandarin speakers were still not a level on par with the other pairs.

Comparison	Mean diff (%) [Eng.-Mand.]	*W*	Pr($>\|z\|$)	
at$\c{c}^h$auma$\sim$at$\c{c}^h$ma	-3.1	161.0	0.33	
at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma	17.7	299.5	0.002	**
at$\c{c}^h$jauma$\sim$at$\c{c}^h$ma	-0.9	203.0	0.69	
at^{h}auma$\sim$at^{h}ma	1.1	223.0	0.26	
at^{h}auma$\sim$at^{h}jauma	3.9	228.0	0.28	
at^{h}jauma$\sim$at^{h}ma	-1.8	195.5	0.86	

Table 2: Mann Whitney U tests for crucial between-language comparisons in Exp. 1

2.3 Discussion

The Mandarin speakers confused the pair [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma] more than the English speakers. Such a result is consistent with the predictions discussed towards the end of the *Introduction*, whereby we expect the Mandarin speakers to hear more illusory [j] in [at$\c{c}^h$auma] than English speakers. Furthermore, the English speakers' performance on the same pair was still not on par with the other pairs (though clearly better than the Mandarin speakers); this suggests that such comparisons between [at$\c{c}^h$auma$\sim$at$\c{c}^h$jauma] are generally difficult independent of language experience. This further reinforces our claim that the relevant comparisons have to be between-language comparisons, and not within-language comparisons; the latter type of comparison cannot distinguish between the contribution of language-specific knowledge to the task, as opposed to general auditory difficulty with some acoustic comparisons, or for that matter subtle artefacts in the stimuli.

Before concluding this section, it is worth noting that the ABX task is an inherently comparative

task. Therefore, a confusion between two stimuli could be either due to perceptual changes in the expected stimulus, or due to perceptual changes in the unexpected stimulus. For example, in the pair [atɕʰauma~atɕʰjauma], it is possible that the reason the Mandarin speakers confused the pair more than the English speakers is not because they heard more /j/ for [atɕʰauma], but because they heard fewer [j] in [atɕʰjauma] (for some reason). As a consequence, though an ABX task is excellent in that it requires far less meta-linguistic knowledge than other tasks, it is complicated by the fact that the interpretation is not direct. Therefore, support from a more direct perceptual task would help the current predictions. This motivates Experiment 2, where we asked participants to judge whether or not a [j] sound is present in the stimuli in what might be called a modified identification task. Note, such a task is clearly more metalinguistic, and therefore there is a stronger likelihood of task-related effects due to response bias, selective attention focused on particular parts of the stimuli, and the effect these have on auditory coding (Caporello Bluvas and Gentner 2013). Despite these concerns, it is useful to run an identification task, as it can give us yet another perspective into what is happening during the perception of the relevant stimuli.

3 Experiment 2

In Experiment 2, we presented Mandarin and American English speakers with the test stimuli, one at a time, and asked them to respond Yes/No to whether there was a glide present in the stimulus. Given the predictions laid out in the *Introduction*, we expect Mandarin speakers to hear more illusory /j/, than American English speakers, when presented with [atɕʰauma], but not when presented with [atʰauma].

3.1 Methods

3.1.1 Participants

Seventeen native Mandarin speakers from close to Beijing (mean age = 20.5 years, SD = 1.7 years, 7 men and 10 women) and 19 native English speakers from Michigan (mean age = 19.7 years, SD = 1.6 years, 5 men and 14 women) participated in the experiment. Neither group of participants participated in Experiment 1. All the participants were recruited at Michigan State University. The Mandarin participants received no compensation for their participation, but the English participants received extra credit for their participation. The Mandarin participants spent

an average of 2.7 years (SD = 1.9 years) in the US before the experiment, primarily as undergraduate students at Michigan State University. Furthermore, the Mandarin speakers also reported an average of 10 years (SD = 4.5 years) of exposure to English as a second language in a classroom environment in China.

3.1.2 Materials

All the test items in Experiment 2 followed the same template as those in Experiment 1, *i.e.*, aC_1V_1ma, in which C_1 was an alveolar or alveo-palatal consonant [t^h / $t\c{c}^h$]; and V_1 was [au / jau / ∅ (Null)]. Furthermore, although the nonce-words used were identical to Experiment 1, the stimuli were re-recorded by the same speaker for Experiment 2. As with Experiment 1, we recorded multiple tokens and chose only those tokens where the co-authors and other relevant members of the lab were satisfied with the quality of the vowels, consonants and tones involved.

There were two recordings used for each test item, and each token was presented 4 times; therefore, there were 8 tokens of each test item, and a total of 48 tokens in the experiment, presented in pseudo-randomized order with the added constraint that there be no identical test items in succession.

3.1.3 Procedure

We used a Yes/No response task (essentially, a modified identification task as discussed below) in Experiment 2 to investigate a perceptual epenthesis effect. The experiment was conducted in a quiet room with a group of 4-6 participants per session. The stimuli were presented with a low-noise headset (Koss R80 headphones) to each participant through an identification task scripted in Praat. The participants were asked to listen to a stimulus and determine whether they heard a [j] sound in the stimuli. All the instructions were in English for the English speakers ("Did you hear a 'y' sound?") and in Mandarin for the Mandarin speakers ("你听到的词中有 'i' 吗?"). We chose to provide the participants with the orthographic description of "i" for the Mandarin speakers because the letter stands for [j] before other vowel letters. However, it is important to note that the letter "i" in Pinyin represents [i] before consonantal letters. Therefore "iama" stands for [jama], but "dima" stands for [dima]. This orthographic confound plays a role below in the interpretation of the results.

Before the actual experiment, each participant completed a practice session to ensure famil-

iarity with the task. The practice session had 12 trials with another set of nonce words, where C_1 was [b]. The inter-trial interval was 1500ms. All 48 trials were randomized for each participant.

3.2 Results

A visual inspection of the mean percentage of Yes-responses to the stimuli by both the Mandarin and English speakers suggests that the following differences were observed between the two language groups (Figure 3): (a) as expected, the Mandarin speakers appear to have more Yes-responses than English speakers for [atς^hauma] (mean difference in Yes-response rate = 33.3%); (b) as expected, the Mandarin and English speakers appear to have similar Yes-response rates for [at^{h}auma] (mean difference in Yes-response rate = -5%); (c) also as expected based on previous findings, the Mandarin speakers appear to have more Yes-responses than English speakers for [atς^hma] (mean difference in Yes-response rate = 57.5%); (d) somewhat unexpectedly, though both Mandarin and English speakers had rather high Yes-responses, the Mandarin speakers gave more Yes-responses than English speakers for [atς^hjauma] (mean difference in Yes-response rate = 6.1%).

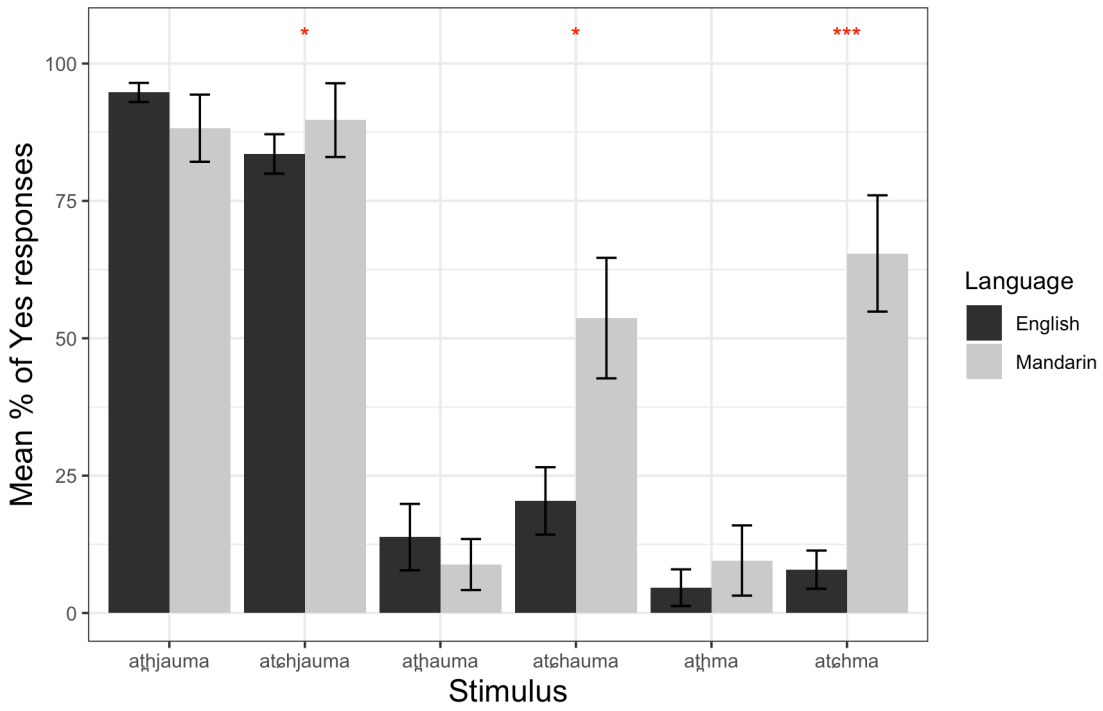

Figure 3: Mean % of Yes-responses of English and Mandarin speakers in Exp. 2 (error bars = 1 S.E.; * = $p < 0.05$; ** = $p < 0.01$; asterisks represent p-values for between-language comparisons)

In order to confirm the observations made by visual inspection of the results, we followed up with statistical analysis. A two-way mixed ANOVA was run to model the percentage of Yes-responses as the dependent variable with LANGUAGE (English, Mandarin) as a between-subjects factor and ITEM (all 6 items) as a within-subject factor. Mauchly's test revealed violations of the assumption of sphericity for the main effect of ITEM and the interaction LANGUAGE*ITEM. Both effects were corrected with Greenhouse-Geisser correction for the degrees of freedom ($\epsilon = 0.76$). There was a main effect of LANGUAGE [$F(1,34) = 12.9$, $p = 0.001$, $\eta^2_{gen} = 0.08$], and a main effect of ITEM [$F(3.8,129.2) = 74.5$, $p < 0.0001$, $\eta^2_{gen} = 0.63$]. Crucially, there was a two-way interaction of LANGUAGE and ITEM [$F(3.8,129.2) = 8.9$, $p < 0.0001$, $\eta^2_{gen} = 0.17$]. The interaction suggests that the Mandarin and English speakers had different responses to different test items.

Following the analytic strategy of Experiment 1, we conducted pairwise Mann Whitney U tests (Table 3). It is worth noting again that the crucial comparisons are between-language comparisons; therefore, there were a total of 6 pairwise comparisons that were conducted. The pairwise tests suggest that Mandarin speakers had more Yes-responses than the English speakers on being presented with [at$\c{c}^h$auma] and [at$\c{c}^h$ma]; the Mandarin speakers all had a few more Yes-responses than the English speakers on being presented with [at$\c{c}^h$jauma].

Stimulus	Mean diff (%) [Mand.-Eng.]	W	Pr($>\|z\|$)	
at^{h}jauma	–6.5	167.5	0.84	
at$\c{c}^h$jauma	6.1	89.0	0.011	*
at^{h}auma	–5.0	170.0	0.74	
at$\c{c}^h$auma	33.3	95.0	0.03	*
at^{h}ma	5.0	157.0	0.85	
at$\c{c}^h$ma	57.5	46.5	0.0001	***

Table 3: Mann Whitney U tests for the crucial between-language group comparisons in Exp. 2

3.3 Discussion

There were three differences observed between the English and Mandarin speakers in Experiment 2. Below, we will go through each case independently. First, we were able to again confirm the crucial prediction that was laid out in the *Introduction*. Mandarin speakers had more Yes-

responses in the [atɕʰauma] than English speakers. Given that pre-vocalically, the Pinyin letter "i" stands for [j] in Mandarin, this result can be interpreted as Mandarin speakers hearing more [j] in [atɕʰauma] than English speakers.

Second, the Mandarin speakers also had more Yes-responses than English speakers to [atɕʰma]. Recall, in Pinyin, the letter "i" represents the vowel [i] pre-consonantally. Therefore, this results suggests that, in this context, Mandarin speakers were very likely to hear an illusory /i/. While not particularly relevant to the illusory consonant case of interest, this result is also consistent with the reverse inference view laid out earlier; as mentioned before, palatalised consonants also appear before /i/, and not just before /j/. Durvasula et al. (2018) observe the same pattern of illusory vowels in Mandarin Chinese, and argue for a similar inference.

Third, Mandarin speakers also had more Yes-responses to [atɕʰjauma] than English speakers. Given the characteristics of Pinyin mentioned above, this suggests they heard more [j] in this context. Though not immediately relevant to the predictions we set out to test, this was an unexpected result as we *a priori* expected a similar proportion of Yes-responses for [atɕʰjauma] by both English and Mandarin speakers; we would like to suggest that it is potentially an experimental artefact for two reasons: (a) the difference between the Mandarin and English Yes-responses is actually quite small (∼6%), and a roughly similar difference, though non-significant, in the opposite direction exists for [atʰjauma], where English speakers seem to be higher than Mandarin speakers, (b) anticipating the results of Experiment 3, we see there that the difference for the same stimulus is in the *opposite* direction, *i.e.*, English speakers responded with higher Yes-responses than Mandarin speakers. Based on these two reasons, it is likely that the differential response between the Mandarin and English speakers for [atɕʰjauma] is not meaningful.

Before concluding this section, there is one aspect of some of the stimuli that is important to note, namely, the alveolar place of articulation used in some of the stimuli above. While in many dialects of American English (including, Michigan English), there is no observed variation with respect to the place of articulation of /tʰ/, there is some debate in the literature on Mandarin. Some Mandarin scholars have described the relevant sounds in Mandarin as alveolar (Kratochvil 1968; Luo and Wang 1981); some others have described them as dental (Chao 1968; Duanmu 2007); and some have described them as potentially varying between alveolar and dental (Lin 2007). It is therefore unclear whether the alveolar consonant [tʰ] used in Experiment 2 was completely natural for Mandarin speakers. This unclarity raised the possibility that the Mandarin participants in the experiment might have responded differently to some of the stimuli, had they been more

natural to them. Given that the Mandarin participants were not purely monolingual Mandarin speakers, and have some experience with American English as undergraduate students, it is possible that their responses to the crucial alveolar stimuli were different from the alveo-palatal stimuli because the alveolarity of the consonants triggered a response based on their American English experience. Such shifts in perception are common in early learners of a second language (Gonzales and Lotto 2013). Therefore, though the lack of difference in responses between the Mandarin and English speakers for the alveolar stimuli, particularly, [at^{h}auma], is very much in favour of the predictions, it would be useful and important to compare the Mandarin/English responses to similar stimuli with dental stops, *i.e.*, [at̪^{h}auma] and [at̪^{h}jauma]. For this reason, in Experiment 3, we replaced the alveolar items with those containing the dental consonant [t̪^{h}], and conducted a very similar Yes/No response task.

4 Experiment 3

As pointed out in the previous section, there is some debate about the place of articulation of the segments called alveolar/dental stop in Mandarin. Therefore, in Experiment 3, we tested Mandarin and American English speakers' responses in a Yes/No task identical to Experiment 2, but where the alveolar stops stimuli were replaced with dental stop stimuli. Following the view laid out earlier, we expect Mandarin speakers to hear more illusory /j/, than American English speakers, when presented with [atɕ^{h}auma], but not when presented with [at̪^{h}auma].

4.1 Methods

4.1.1 Participants

Nineteen Mandarin speakers from Beijing (mean age = 20.3 years, 7 men and 12 women) and 18 native English speakers from Michigan (mean age = 18.9 years, SD = 1.21 years, 3 men and 15 women) participated in the experiment. Neither group of participants participated in Experiments 1 and 2. All the participants were recruited at Michigan State University. The Mandarin participants received no compensation for their participation, but the English participants received extra credit for their participation. The Mandarin participants spent an average of 2.7 years (SD = 1.8 years) in the US before the experiment, primarily as undergraduate students at Michigan State University. Furthermore, the Mandarin speakers also reported an average of 10 years (SD = 4.3

years) of exposure to English as a second language in a classroom environment in China.

4.1.2 Materials

All the test items followed the template aC_1V_1ma, in which C_1 was a dental or alveo-palatal consonant [$t̪^h$ / $tɕ^h$]; and V_1 was [au / jau / ∅ (Null)]. All the stimuli were re-recorded (even if there was an overlap with previous experiments) by the same speaker for Experiment 3. As with Experiments 1 and 2, we recorded multiple tokens and chose only those tokens where the co-authors and other relevant members of the lab were satisfied with the quality of the vowels, consonants and tones involved.

As with Experiment 2, there were two recordings used for each test item, and each token was presented 4 times; therefore, there were 8 tokens of each test item, and a total of 48 tokens in the experiment, presented in pseudo-randomized order with the added constraint that there be no identical test items in succession.

4.1.3 Procedure

The procedure was identical to that of Experiment 2. The experiment was conducted in a quiet room with a group of 4-6 participants per session. The stimuli were presented with a low-noise headset (Koss R80 headphones) to each participant through an identification task scripted in Praat. Before the actual experiment, each participant completed a practice session to ensure familiarity with the task. The practice session had 12 trials with another set of nonce words, where C_1 was [b]. The inter-trial interval was 1500ms. All 48 trials were randomized for each participant.

4.2 Results

A visual inspection of the mean percentage of Yes-responses to the stimuli by both the Mandarin and English speakers suggests that *all* the following differences were found between the two language groups (Figure 4): (a) as expected, the Mandarin speakers appear to have more Yes-responses than English speakers for [at$ɕ^h$auma] (mean difference in Yes-response rate = 37.1%); (b) as expected, the Mandarin and English speakers appear to have similar Yes-response rates for [at^{h}auma] (mean difference in Yes-response rate = -2.9%); (c) also as expected, the Mandarin speakers appear to have more Yes-responses than English speakers for [at$ɕ^h$ma] (mean difference in Yes-response rate = 67.7%); (d) some what unexpectedly, though both Mandarin and

English speakers had rather high Yes-responses, the English speakers gave more Yes-responses than Mandarin speakers for both [atɕʰjauma] (mean difference in Yes-response rate = -22.2%) and [atʰjauma] (mean difference in Yes-response rate = -25.5%).

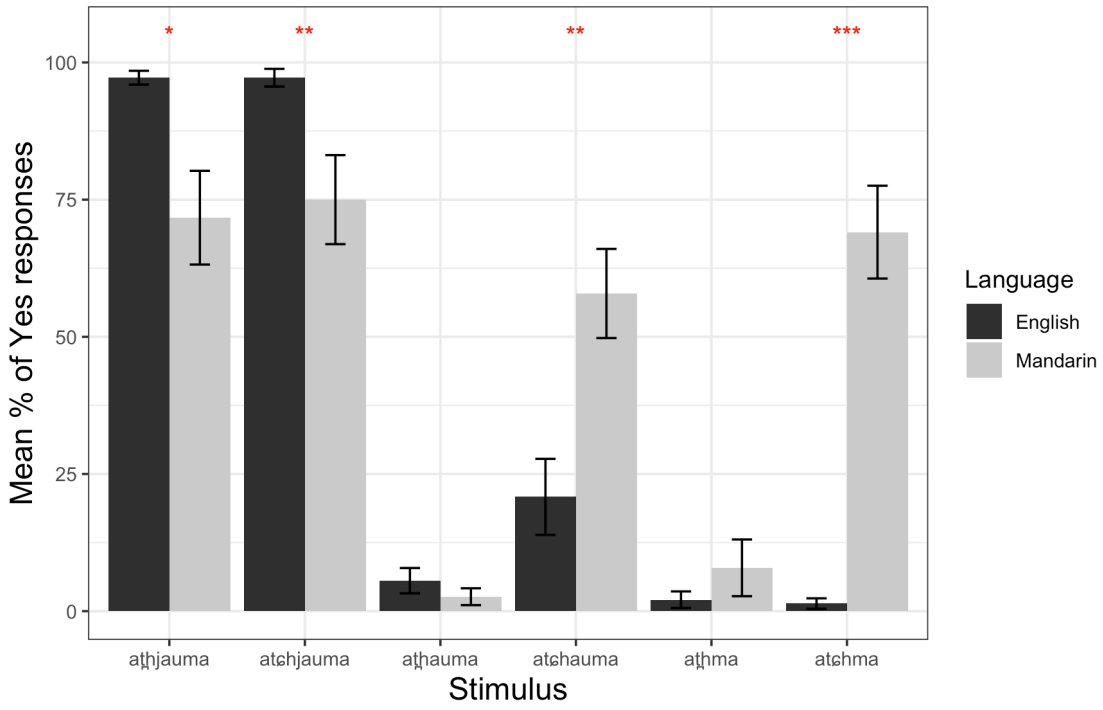

Figure 4: Mean % of Yes-responses of English and Mandarin speakers in Exp. 3 (error bars = 1 S.E.; * = $p<0.05$; ** = $p<0.01$; asterisks represent p-values for between-language comparisons)

In order to confirm the observations made by visual inspection of the results, we followed up with statistical analysis. A two-way mixed ANOVA was run to model the percentage of correct responses as the dependent variable with LANGUAGE (English, Mandarin) as a between-subjects factor and ITEM (all 6 items) as a within-subject factor. Mauchly's test revealed violations of the assumption of sphericity for the main effect of ITEM and the interaction LANGUAGE*ITEM. Both effects were corrected with Greenhouse-Geisser correction for the degrees of freedom ($\epsilon = 0.64$). There was a main effect of LANGUAGE [$F(1,35) = 5.1$, $p = 0.03$, $\eta^2_{gen} = 0.04$], and a main effect of ITEM [$F(3.2,112) = 99.8$, $p < 0.0001$, $\eta^2_{gen} = 0.66$] Crucially, there was a two-way interaction of LANGUAGE and ITEM [$F(3.2,112) = 24.9$, $p < 0.0001$, $\eta^2_{gen} = 0.33$]. The interaction suggests that the Mandarin and English speakers had different responses to different test items.

Following Experiments 1 and 2, we conducted pairwise Mann Whitney U tests (Table 4). For the crucial between-language comparisons, there were a total of 6 comparisons that were conducted. The pairwise tests suggest that Mandarin speakers had more Yes-responses than

the English speakers on being presented with [atɕʰauma] and [atɕʰma]; the Mandarin speakers had roughly similar Yes-response rates to English speakers for [aʈ̪ʰauma] and aʈ̪ʰma; the Mandarin speakers all had fewer Yes-responses than the English speakers on being presented with [atɕʰjauma] and [aʈ̪ʰjauma].

Stimulus	Mean diff (%) [Mand.-Eng.]	W	Pr($>\|z\|$)	
aʈ̪ʰjauma	−25.5	239.0	0.018	*
atɕʰjauma	−22.2	247.5	< 0.01	**
aʈ̪ʰauma	−2.9	200.0	0.25	
atɕʰauma	37.1	78.0	< 0.01	**
aʈ̪ʰma	5.8	161.5	0.65	
atɕʰma	67.7	30.0	< 0.0001	***

Table 4: Mann Whitney U tests for the crucial between-language group comparisons in Exp. 3

4.3 Discussion

There were differences observed between the English and Mandarin speakers in Experiment 3. Below, we will go through each case independently. First, as with Experiment 2, we were able to again confirm the crucial prediction that Mandarin speakers heard more illusory [j] in [atɕʰauma] than English speakers. This prediction manifested as more Yes-responses by Mandarin speakers for the stimuli, compared to the English speakers. Furthermore, there was no similar difference for the stimulus with the dental stop [aʈ̪ʰauma], replicating the results for the stimulus with the alveolar stop ([atʰauma]) in Experiment 2.

Second, again as in Experiment 2, Mandarin speakers heard a lot of /i/ in [atɕʰma]. This result, as pointed out before, while consistent with the reverse inference view laid out in this article, is tangential to our current interests. So, we withhold any further discussion.

Third, for both the [aʈ̪ʰjauma] and [atɕʰjauma], Mandarin speakers responded with fewer Yes-responses than English speakers. A few notes are worth making about this result: (a) as pointed out in Section 3.3, the differences in Yes-responses to at least the alveo-palatal stimuli is inconsistent. It was in the opposite direction in Experiment 2, where Mandarin speakers had higher Yes-responses than English speakers. (b) On looking more carefully at the Mandarin participants'

responses, it looks like at least two participants did not understand the task or choices given. One participant had 0 Yes-responses for all the stimuli, suggesting that they either misunderstood the task, or had trouble hearing the stimuli for some unknown reason. A second participant had 0 Yes-response for all the stimuli except [atɕʰma], where they had an 87.5% Yes-response. This latter pattern suggests that the participant was treating "i" as only the vowel /i/, and therefore they responded with a "yes" only to those stimuli where they heard an illusory vowel /i/. It is instructive to look at the results without these two participants (Figure 5). As can be seen, both the surprising results, while still (marginally) statistically significant, have substantially decreased in size. Furthermore, the size of the differences between the Mandarin and English subjects for [at̪ʰjauma] and [atɕʰjauma] now, are much smaller than those for [atɕʰauma] and [atɕʰma]. The mean differences (Mandarin - English) for the four are as follows: [at̪ʰjauma] = -16.7%, [atɕʰjauma] = -13.05%, [atɕʰauma] = 42.8%, & [atɕʰma] = 70.5%. The above two observations (a-b) suggest that the differences between Mandarin and English speakers in their Yes-response to [at̪ʰjauma] and [atɕʰjauma] are likely to be spurious, and potentially related to the ambiguity of the letter "i" in Pinyin. (Note: While we include the p-values for the relevant comparisons in Figure 5, we do not reproduce the full set of results. The interested reader might want to know that an ANOVA identical to the one described above without the two subjects resulted crucially in a two-way interaction for LANGUAGE and ITEM [$F(3.3,102.3) = 22.3$, $p < 0.0001$, $\eta^2_{gen} = 0.37$]. We do not present these results in more elaborate detail as the actual results in the previous section, because they involve removing a couple of participants, based on what seem to us reasonable post-hoc criteria, and some interested readers might be uncomfortable with the manipulation.)

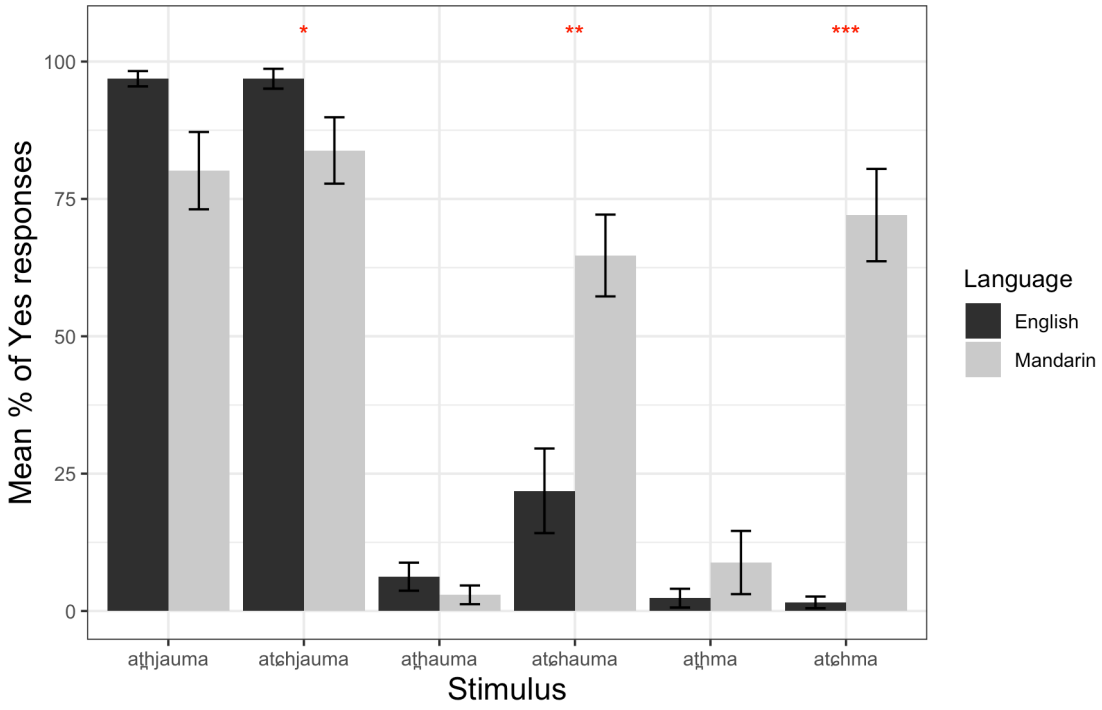

Figure 5: Mean % of Yes-responses of English and Mandarin speakers in Exp. 3, excluding two participants who did not seem to understand the task (error bars = 1 S.E.; * = $p < 0.05$; ** = $p < 0.01$; asterisks represent p-values for between-language comparisons)

5 Conclusion

In this paper, we aimed to further corroborate the view that, during speech perception, the task of the listener is to identify the best estimate of the intended underlying representations of the utterance given the acoustic token. As has been pointed out before, an important consequence of this viewpoint is that both phonological knowledge (therefore, knowledge of phonological alternations and phonotactic constraints) and the acoustics of the relevant tokens make contributions during perception, i.e., perception cannot be solely based on one or the other. In relation to the focus of the current article, the view predicts that, just as there are ILLUSIONS OF QUANTITY involving vowels (syllabic nuclei), there should be ILLUSIONS OF QUANTITY involving consonants (non-nuclear segments). Furthermore, the view also predicts that such illusions should be modulated by the specific phonological patterns of the listener's native language; therefore, one might expect them to appear even outside the phonotactic context of consonantal contact, where most, if not all previous such ILLUSION OF QUANTITY have been observed. A second prediction stemming

from the view is that there is nothing special about word-internal phonotactic restrictions. Though prior work has largely focussed on such phonotactic restrictions, what really matters to the listener as per the viewpoint is the entire phonological system; therefore, if there is a viable parse for the auditory input for a listener that includes a word boundary or prosodic boundary, then there is no expectation of any illusory segments, even if the specific sequence is illicit word-internally.

Given the above predictions stemming from the reverse inference view point, and with the phonological facts of Mandarin and English, we expect that Mandarin speakers but not the English speakers should hear illusory /j/ in [tɕʰau] sequences. In line with this prediction, we showed through three experiments that Mandarin speakers appear to be hearing more illusory glides [j] when presented with auditory input containing [tɕʰau] sequences. In Experiment 1, an ABX task, we observed that native Mandarin speakers were confusing the stimuli [atɕʰauma]∼[atɕʰjauma] far more than native American English speakers. In Experiments 2 and 3 (Yes-No task), native Mandarin speakers reported a higher number of illusory /j/ in [atɕʰauma] than native American English speakers.

There are three further issues that deserve further discussion: (a) The status of glides in Mandarin Chinese, (b) Implications for loanword phonology, (c) Implications for our understanding of hypercorrection. We turn to these issues in the following sub-sections.

5.1 The status of glides in Mandarin Chinese

As mentioned earlier, the syllabic affiliation of the glides is quite a debated issue in Chinese phonology (Z. Bao 1990; Duanmu 2007; Weijer and Zhang 2008; Yip 1980, amongst others). The crux of the issue is that there are no clear segmental alternations that allow one to infer the best syllabic representation for glides in Mandarin Chinese. As a result, researchers have either used language games (as discussed below) or speech errors (Z. Bao 1996; Wan 2003; Weijer and Zhang 2008, amongst others). Here, we present arguments based primarily on language games and distributional restrictions that show: (a) that the glide is part of the onset, not the rhyme, (b) the glide is likely a separate segment from other onset consonants. Based on these two arguments, we deem it reasonable to claim that glides are consonantal, *i.e.*, non-nuclear segments, in Mandarin Chinese.

To support the claim that the glides are part of the onset, and not the rhyme, we present three arguments. The first argument comes from language games called *Fanqie* languages, which have by themselves received quite a lot of attention, and have invariably formed an important part of

the argument related to syllable structure in Chinese languages (Z. Bao 1990, 1996; Chao 1931; Duanmu 2007; Weijer and Zhang 2008; Yip 1980, 1982; Zhang 2011, amongst others).

The argument we present below is from a specific Fanqie language game called Na-ma, which is based on Chengdu, a sister dialect of Beijing Mandarin Chinese (Z. Bao 1996; Chao 1931; Duanmu 2007). In this game, a word is reduplicated, and the first onset of the reduplicant is replaced by an /n/ (1a). As can be observed, in cases where there is an initial glide or an initial consonant-glide sequence, the whole pre-vocalic sequence is replaced by /n/ (1b). Furthermore, [nw] and [nj] are phonotactically licit sequences in Mandarin Chinese; therefore, the replacement of [lj] or [tw] by [n] cannot be due to independent phonotactic requirements. The simplest analysis of this language game is one where the unit being replaced is an onset, thereby suggesting that the glide must be part of the onset.

1. The Na-ma Fanqie language game (lexical tone not represented in original sources)

   (a) Simple onsets

      i. [ma] "mother" → [n̲a-ma]

      ii. [te] "to get" → [n̲e-te]

   (b) Simple glide or consonant-glide sequences

      i. [je] "grandfather" → [n̲e-je]

      ii. [twei] "correct" → [n̲ei-twei], *[n̲wei-twei]

      iii. [ljaŋ] "two" → [n̲aŋ-ljaŋ], *[n̲jaŋ-tjaŋ]

A second argument that shows that the glide is indeed in the onset comes from the fact that vowel-initial words must have an epenthetic consonant in what have been called "zero onset" cases (Chao 1968; Duanmu 2007). The exact consonant inserted appears to depend on the speaker (2a). However, in words that begin with a glide, there is no such insertion, suggesting that there is already a consonant in the onset (2b).

2. Consonant insertion in "zero onset" cases game (lexical tone represented as superscript; 1 = high, 2 = rising, 3 = low, 4 = falling)

   (a) Vowel-initial words

   i. [ʔɤɤ^{2}]/[ɣɤɤ^{2}]/[ŋɤɤ^{2}]/[∅ɤɤ^{2}] "goose"

   ii. [ʔæ̃n^{1}]/[ɣæ̃n^{1}]/[ŋæ̃n^{1}]/[∅æ̃n^{1}] "peace"

   (b) Glide-initial sequences

   i. [waa^{1}] (*[ɦwaa] *[ɣwaa] *[ʔwaa] *[ŋwaa]) "frog"

   ii. [jaa^{1}] (*[ɦjaa] *[ɣjaa] *[ʔjaa] *[ŋjaa]) "crow"

A third argument that shows that the glide is indeed in the onset comes from rhyming patterns in Mandarin. In a study of rhyming patterns in over 800 poems, M. Bao (1978) (as discussed in Z. Bao (1996)) found that the glide always behaves as if it was outside the rhyming constituent. For example, [pe] rhymes with [e], [we], [je], and [twe].

The above arguments suggest that the glides in Mandarin Chinese are indeed in the onset, thereby justifying our use of the term 'consonant', *i.e.*, non-nuclear segment, in describing them.[9]

The claim that the glide in consonant-glide sequences in Mandarin Chinese is a separate segment, as opposed to a secondary articulation, can be made on two arguments. The first argument comes from another Fanqie language game called May-Ka, which is based on Beijing Mandarin Chinese (Chao 1931; Yip 1982). The description of the language game is more complicated than what we present below. However, focussing on the relevant aspects of the language game, the stem is reduplicated; which is followed by the addition of a rhyme [aj] and the replacement of the first consonant with [k] in the reduplicant 3a. What is of most interest to current needs is the fact that the replacement with [k] in words with consonant-glide sequences only targets the first consonant 3b. (Note, the [k] → [tɕ^{h}] changes in the second example follows the phonotactic constraint that only [tɕ^{h}] can precede front glides; furthermore, [aj] → [ɛ] in the first syllable.) Crucially, if indeed the glide were a secondary articulation of the first consonant, then the glide should also have been replaced, contrary to fact. This suggests that in such consonant-glide sequences, the glide is a separate segment from the preceding consonant.

[9]Somewhat tangentially, it is, of course, possible to describe glides and vowels differing *only* in terms of syllabic positions as opposed to the feature [consonantal] (Clements and Keyser 1983; Kaye and Lowenstamm 1984; Rosenthall 1994). This has however been argued against by Nevins (2005). Furthermore, if one still takes this approach, then the difference between the prior illusory vowels research and the current one is not specifically about the feature [consonantal] as defined by some phonologists, as much as it is about syllabic (nuclear vs. non-nuclear) position.

3. The May-ka Fanqie language game (data from (Z. Bao 1990; Yip 1982); tones represented include tone-sandhi changes; post-nuclear glides are transcribed as such following original source; meanings not provided in the original sources; tones not available in the orginal for the last example.)

   (a) Simple onsets

   i. [ma^{3}] → [maj^{2}-k̲a^{3}]

   ii. [pej^{3}] → [paj^{2}-k̲ej^{3}]

   (b) Consonant-glide sequences

   i. [xwaj^{2}] → [xwaj^{2}-k̲wej^{2}], *[xwaj^{2}-k̲ej^{2}]

   ii. [lja] → [ljɛ-t͡ɕ̲ja], *[ljɛ-k̲a]

The second argument that the glide in such consonant-glide sequences is a separate consonant comes from distributional evidence. First, glides can independently appear in an onset (as shown above). Furthermore, the distributions of alveo-palatals can receive a uniform/single description if glides in consonant-glide sequences are separate segments from the preceding consonants; namely, the description can be simplified to a statement that alveo-palatals appear before front segments. If in contrast, the glides were treated as secondary articulations, then the restriction of alveo-palatals before front vowels and glides will need to be stated using very different phonotactic constraints, with the former as sequential (syntagmatic) constraints, while the latter as simultaneous co-occurence (paradigmatic) constraints. The above distributional facts suggest that the simplest hypothesis is one where the glides are separate segments from the preceding consonants in consonant-glide sequences.[10]

Taken together the arguments in this section suggest that the glides in consonant-glide sequences in Mandarin Chinese are in the onset, and are separate consonants from the preceding consonant.

While we use this proposed representation as the basic assumption of this paper, there are others who have claimed that the glide is neither in the onset nor in the rhyme per se (Weijer and Zhang 2008; Yip 1980, 1982). Furthermore, Z. Bao (1996) discusses some morphophonological data from two dialects of Mandarin, Yanggu and Taiyuan. In those dialects, there is some evidence

[10] It is worth pointing out that if this analysis is incorrect, and indeed the glides in consonant-glide sequences are secondary articulations in Mandarin Chinese, then the article would still be a case of an auditory illusion; however, it would be classified as a case of an ILLUSION OF QUALITY, as the sounds [tʃ] is being perceived as /t͡ɕ^{hj}/ by Mandarin listeners, but as /tʃ/ by English speakers.

to suggest asymmetrical behaviour of the glides in post-consonantal positions: In Yanggu, the front glide appears to be part of the onset, and the back glide appears to be part of the rhyme; in Taiyuan, the front glide appears to be part of the rhyme, and the back glide appears to be part of the onset. Crucially for us, based on rhyming patterns, Z. Bao (1996) ultimately concludes that the glides (whatever their affiliation) are not part of the nucleus. For these reasons, if one were committed to one of the above alternative representational claims, it would not alter the main focus of the article as the glide would still be analysed as a consonantal (non-nuclear) segment, and therefore the observed auditory illusion would still be related to non-nuclear segments.

5.2 Implications for loanword phonology

In this section, we briefly discuss how the speech perception results presented in this article help understand loanword patterns in Mandarin Chinese. In a corpus study consisting of 2423 borrowings into Mandarin from English (N = 1177), German (N = 977) and Italian (N = 269), Miao (2005) showed that pre-vocalic palatal affricates from the three languages are typically borrowed into Mandarin Chinese either as alveo-palatal affricates (70-75% of the time) or as retroflex affricates (20-25% of the time), as exemplified below in (4), where the crucial segment has been underlined and italicised.

Crucially, when a pre-vocalic palatal consonant is borrowed into Mandarin Chinese as an alveo-palatal sound, it is always followed by a palatal glide (unless the following vowel is itself a front vowel). These loanword facts fall out naturally from the view of speech perception probed in this paper.

4. Typical loanword patterns in Mandarin Chinese (lexical tone represented as superscript number for each syllable; 1 = high, 2 = rising, 3 = low, 4 = falling)

   (a) Alveolar contexts

   i. [$\underline{t\text{ɕ}^{h}j}$aŋ^{2} ʂəŋ^{1}] “Johnson & Johnson (English)”

   ii. [$\underline{t\text{ɕ}^{h}j}$e^{4} ɚ^{3} ɕi^{1}] “Chelsea (English)”

   (b) Borrowed as retroflex consonants

   i. [$\underline{t\text{ʂ}^{h}}$a^{2} li^{3} pej^{4} ɚ^{3}] “Charlie Bell (English)”

   ii. [$\underline{t\text{ʂ}}$an^{1} mu^{3} sɹ̩^{1} po^{2} k^{h}ɤ^{4}] “James Burke (English)”

Before ending this discussion, we would like to add the caveat that the discussion should not be taken as arguing that all loanword patterns can be reduced to perceptual sources. In fact, we agree with the vast literature on loanwords that there are multiple sources of loanwords into a language (Davidson 2007; Kang 2011; Peperkamp 2005; Smith 2006, *inter alia*). We do however highlight that our results suggest that theorists should be careful about the representation available to the listener through the speech perception process, *i.e.*, the underlying representation inferred from the input token, while modelling loanword patterns through a loanword-specific phonology.

5.3 Implications for our understanding of hypercorrection

The viewpoint probed in this article naturally accounts for cases of what have been called HYPERCORRECTION, in previous research. Although the focus of the article was on a sequential phonotactic restriction (on palatal consonants in Mandarin Chinese), a second type of phonological process that is likely to have an effect in conditioning auditory illusions is *segment deletion*. The presence of a regular process of segment deletion ($/C_1/ \rightarrow [\emptyset]$) in the phonology of the language supports the reverse inference of the same segment in the phonemic representation when the surface representation has nothing (reverse inference: $[\emptyset] \rightarrow /C_1/$). This expectation has been used to explain illusory vowels (Durvasula et al. 2018; Durvasula and Kahng 2015), but is equally applicable to the case of illusory consonants.

As an example, we think the phenomenon or r-intrusion, observed in a variety of languages can be understood to originate during speech perception. We'd like to suggest that the pattern/alternation originates as a hypercorrection *during the speech perception process*, and not just some post-perceptual analogical change; in other words, it is the perceptual inference of an /r/ in positions where such an inference is supported by the phonology of the language. If indeed this is the case, then speakers of such dialects should be less sensitive to differences between words ending in vowels and those ending in /r/, *e.g.*,[aba] vs. [abar].

Particularly interesting evidence for this view comes from the specific patterns of r-intrusion in many dialects of English (Blevins 2004; Halle and Idsardi 1997; McCarthy 1991; Wells 1982). Dialects that have r-intrusion invariably have an r-deletion process. A further interesting fact about many dialects with r-intrusion, discussed in detail by Halle and Idsardi (1997) and McCarthy (1991), is that while the r-deletion process can be generalised to all /r/s in coda position (5a), the r-intrusion process typically is constrained to low vowels (5b). Adopting the perspective that the r-intrusion originates as a product of reverse inference to the best parse of the intended underlying

(or phonemic) representations of the utterance given the acoustic token allows us to explain this disparity. Typically, dialects with restrictions to r-intrusion also have a separate process of pre-rhotic [ə]-insertion after high vowels/glides (5c).

5. Rhotic deletion/insertion patterns in English dialects

(a) R-deletion

i.	/sp^{h}ar/	→	[spa]	“spar”
ii.	/sɔr/	→	[sɔ]	“soar’

(b) R-intrusion

i.	/sp^{h}a # ɪz/	→	[spa$\underline{r}$ɪz]	“spa is”
ii.	/ælʟdʒɪbrə # ɪz/[11]	→	[ælʟdʒɪbrə$\underline{r}$ɪz]	“algebra is”
iii.	/fi # ɪz/	→	[fi(ʔ)ɪz], *[fi$\underline{r}$ɪz]	“fee is”

(c) Pre-rhotic [ə]-insertion after high vowels/glides

i.	/fir/	→	[fiə]	“fear”
ii.	/fir-ɪŋ/	→	[fiərɪŋ]	“fearing”

As a consequence of such patterns in the language, when the listener of such dialects hears an input with a final non-high vowel, such as [ata], both /ata/ and /atar/ are reasonable percepts that can account for the phonetics of the input given the phonology of the language (note, the percept is the underlying representation). However, when the listener of such dialects hears an input with a final high vowel such as [ati], only /ati/ (but not, */atir/) is a reasonable percept that can account for the phonetics of the input given the phonology of the language. For, if the perceiver were to infer /atir/, they should also have heard a [ə] in the input. Therefore, we predict that, when compared to speakers of rhotic dialects, speakers of such non-rhotic dialects should have a decreased perceptual distance between [ata]∼[atar], but not between [ati]∼[atir]. We leave this a study for future work.

Finally, it is worth pointing out that it is important to separate such cases of HYPERCORRECTION which involving undoing phonological patterns, from cases that Ohala (1981, 1993) calls “hypercorrection”, which involve a mis-attribution of some acoustic cues in the input related to a specific segment to a different (proximal) segmental source; this latter phenomenon has also been

called “feature parsing” by others (Gow 2003). Though both types of HYPERCORRECTION are perfectly compatible with the reverse inference viewpoint probed in this article, in this section, we focused on the one related to phonological alternations.

Appendix

A.1 - Experiment 1 Results

Table 5: Correct response rates for each of the participants for each pair in Exp. 1

Lang.	Part.	tɕʰau,tɕʰjau	tɕʰau,tɕʰm	tɕʰjau,tɕʰm	tʰau,tʰjau	tʰau,tʰm	tʰjau,tʰm
Eng	1	93.75	100	100	75	100	100
Eng	2	81.25	87.5	75	81.25	93.75	75
Eng	3	43.75	87.5	87.5	37.5	100	87.5
Eng	4	56.25	87.5	100	87.5	87.5	100
Eng	5	50	87.5	81.25	81.25	100	93.75
Eng	6	87.5	93.75	100	100	100	100
Eng	7	75	62.5	75	100	75	68.75
Eng	8	100	100	100	100	100	100
Eng	9	93.75	100	93.75	100	100	100
Eng	10	87.5	100	93.75	93.75	100	81.25
Eng	11	75	100	100	100	100	100
Eng	12	68.75	100	100	87.5	100	100
Eng	13	93.75	100	100	93.75	100	100
Eng	14	87.5	100	100	100	93.75	100
Eng	15	68.75	100	100	87.5	100	100
Eng	16	62.5	100	100	87.5	100	100
Eng	17	87.5	100	100	100	100	100
Eng	18	100	93.75	100	100	100	100
Eng	19	87.5	100	100	93.75	100	100
Mand	20	68.75	100	100	93.75	100	100
Mand	21	75	100	100	93.75	93.75	100
Mand	22	68.75	87.5	100	100	87.5	87.5
Mand	23	43.75	100	100	87.5	100	100
Mand	24	62.5	100	93.75	81.25	100	100
Mand	25	56.25	100	100	93.75	100	100

Mand	26	43.75	100	100	93.75	100	100
Mand	27	56.25	100	100	81.25	100	100
Mand	28	62.5	87.5	93.75	93.75	100	93.75
Mand	29	68.75	93.75	100	93.75	93.75	100
Mand	30	43.75	100	100	93.75	100	100
Mand	31	50	100	87.5	56.25	81.25	100
Mand	32	68.75	93.75	87.5	56.25	93.75	87.5
Mand	33	56.25	100	93.75	93.75	100	100
Mand	34	87.5	100	93.75	100	100	93.75
Mand	35	62.5	100	93.75	87.5	93.75	93.75
Mand	36	87.5	100	100	93.75	93.75	100
Mand	37	56.25	100	93.75	87.5	87.5	87.5
Mand	38	75	93.75	100	93.75	100	100
Mand	39	31.25	100	81.25	43.75	100	93.75

A.2 - Experiment 2 Results

Table 6: Yes-response rates for each of the participants for each item in Exp. 2

Lang.	Participant	at^{h}jauma	at$\c{c}^{h}$jauma	at^{h}auma	at$\c{c}^{h}$auma	at^{h}ma	at$\c{c}^{h}$ma
Eng	1	87.5	75	0	0	0	0
Eng	2	87.5	87.5	0	0	0	0
Eng	3	87.5	62.5	0	0	62.5	37.5
Eng	4	100	100	0	25	0	0
Eng	5	100	100	0	0	0	0
Eng	6	100	100	50	25	0	0
Eng	7	100	75	12.5	0	0	0
Eng	8	100	50	0	0	0	0
Eng	9	100	62.5	0	0	0	0
Eng	10	100	62.5	0	50	0	0
Eng	11	100	87.5	75	75	12.5	25
Eng	12	87.5	87.5	0	0	0	0

Eng	13	100	100	62.5	37.5	0	25
Eng	14	75	100	0	75	0	50
Eng	15	100	100	0	0	0	0
Eng	16	100	87.5	62.5	62.5	0	0
Eng	17	87.5	75	0	12.5	12.5	0
Eng	18	87.5	87.5	0	12.5	0	12.5
Eng	19	100	87.5	0	12.5	0	0
Mand	20	100	100	25	75	0	0
Mand	21	87.5	0	0	0	100	87.5
Mand	22	100	100	0	100	0	100
Mand	23	100	100	0	0	0	100
Mand	24	100	100	0	12.5	0	100
Mand	25	0	100	0	75	0	100
Mand	26	100	87.5	0	87.5	0	37.5
Mand	27	100	100	0	0	0	0
Mand	28	87.5	100	0	100	0	100
Mand	29	100	100	0	100	0	0
Mand	30	75	37.5	0	0	0	12.5
Mand	31	100	100	12.5	100	0	100
Mand	32	100	100	50	100	12.5	100
Mand	33	87.5	100	0	12.5	0	100
Mand	34	100	100	0	0	0	12.5
Mand	35	62.5	100	62.5	50	50	62.5
Mand	36	100	100	0	100	0	100

A.3 - Experiment 3 Results

Table 7: Yes-response rates for each of the participants for each item in Exp. 3

Lang.	Participant	aʈʰjauma	atɕʰjauma	aʈʰauma	atɕʰauma	aʈʰma	atɕʰma
Eng	1	100	100	0	0	0	0
Eng	2	100	100	12.5	12.5	12.5	0

Eng	3	100	100	0	75	0	12.5
Eng	4	100	100	12.5	0	0	0
Eng	5	100	100	37.5	75	0	0
Eng	6	87.5	100	0	0	0	0
Eng	7	87.5	87.5	12.5	0	0	0
Eng	8	100	100	0	12.5	0	0
Eng	9	100	100	0	0	0	0
Eng	10	87.5	100	0	75	0	0
Eng	11	100	100	12.5	12.5	0	0
Eng	12	100	100	0	25	0	0
Eng	13	100	75	0	0	0	0
Eng	14	87.5	87.5	0	62.5	0	0
Eng	15	100	100	0	0	25	12.5
Eng	16	100	100	12.5	25	0	0
Eng	17	100	100	0	0	0	0
Eng	18	100	100	0	0	0	0
Mand	19	100	100	0	50	0	0
Mand	20	100	100	0	100	12.5	87.5
Mand	21	37.5	12.5	0	0	87.5	87.5
Mand	22	100	100	0	100	0	100
Mand	23	100	100	0	25	0	100
Mand	24	50	37.5	12.5	62.5	0	25
Mand	25	100	87.5	0	62.5	0	100
Mand	26	0	87.5	0	87.5	0	87.5
Mand	27	0	0	0	0	0	0
Mand	28	100	87.5	0	12.5	0	0
Mand	29	62.5	75	12.5	75	0	87.5
Mand	30	87.5	100	0	75	0	37.5
Mand	31	0	0	0	0	0	87.5
Mand	32	100	100	0	75	0	75
Mand	33	100	100	0	100	0	100
Mand	34	75	87.5	0	50	0	100

Mand	35	100	100	0	100	0	87.5
Mand	36	62.5	62.5	25	50	50	62.5
Mand	37	87.5	87.5	0	75	0	87.5

Acknowledgements

References

Bao, Mingwei (1978). “Ren Xiandai Shi Yun (On Rhyming in Modern Poetry).” *Bulletin Nanjing University (Philosophy and Social Science)* 4.

Bao, Zhiming (1990). “Fanqie Languages and Reduplication.” *Linguistic Inquiry* 21.3, pp. 317–350. ISSN: 00243892, 15309150.

Bao, Zhiming (1996). “The Syllable in Chinese / 漢語的音節.” *Journal of Chinese Linguistics* 24.2, pp. 312–354. ISSN: 00913723.

Berent, Iris, Tracy Lennertz, J Jun, M Moreno, and Paul Smolensky (2008). “Language universals in human brains.” 105, pp. 5321–5325.

Berent, Iris, Tracy Lennertz, Paul Smolensky, and Vered Vaknin-Nusbaum (2009). “Listeners’ knowledge of phonological universals: Evidence from nasal clusters.” *Phonology* 26, pp. 75–108.

Berent, Iris, Donca Steriade, Tracy Lennertz, and Vered Vaknin (2007). “What we know about what we have never heard: Evidence from perceptual illusions.” *Cognition* 104, pp. 591–630.

Best, C. T., P. Hallé, O.-S. Bohn, and A. Faber (2003). “Cross-language perception of nonnative vowels: Phonological and phonetic effects of listeners’ native languages.” *Proceedings of the 15th international congress of phonetic sciences*. Barcelona, pp. 2889–2892.

Bever, Thomas G and David Poeppel (2010). “Analysis by synthesis: A (re-)emerging program of research for language and vision.” *Biolinguistics* 4.2-3, pp. 174–200.

Blevins, Juliette (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press. DOI: 10.1017/CBO9780511486357.

Boersma, Paul and David Weenink (2016). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/.

Caporello Bluvas, Emily and Timothy Q. Gentner (2013). “Attention to natural auditory signals.” *Hearing Research* 305, pp. 10–18.

Chao, Yuen Ren (1931). “Fanqie yu ba zhong [Eight types of Fanqie languages].” *Bulletin of the Institute of History and Philology*. Vol. 2. 3. Academia Sinica, pp. 312–354.

Chao, Yuen Ren (1968). *A Grammar of Spoken Chinese*. Berkeley, CA, USA: University of California Press.

Clements, George N. and Samuel Jay Keyser (1983). *CV Phonology: A generative theory of the syllable*. Cambridge, MA, USA: MIT Press.

Darcy, Isabelle, Franck Ramus, Anne Christophe, Katherine Kinzler, and Emmanuel Dupoux (2009). "Phonological knowledge in compensation for native and non-native assimilation." *Variation and Gradience in Phonetics and Phonology*. Ed. by Frank Kügler, Caroline Féry, and Ruben Vijver. Phonology and Phonetics [PP] 14. Berlin, New York: Mouton de Gruyter, pp. 265–309.

Davidson, Lisa (2007). "The relationship between the perception of non-native phonotactics and loanword adaptation." *Phonology* 24.2, pp. 261–286.

Davidson, Lisa and Jason A. Shaw (2012). "Sources of illusion in consonant cluster perception." *Journal of Phonetics* 40.2, pp. 234–248. ISSN: 0095-4470. DOI: http://dx.doi.org/10.1016/j.wocn.2011.11.005.

Duanmu, San (1999). "Metrical structure and tone: evidence from Mandarin and Shanghai." *Journal of East Asian Linguistics* 8.1, pp. 1–38.

Duanmu, San (2007). *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler (1999). "Epenthetic vowels in Japanese: A perceptual illusion?" *Journal of Experimental Psychology: Human Perception and Performance* 25.6, pp. 1568–1578. ISSN: 1939-1277(ELECTRONIC);0096-1523(PRINT). DOI: 10.1037/0096-1523.25.6.1568.

Dupoux, Emmanuel, Erika Parlato, Sonia Frota, Yuki Hirose, and Sharon Peperkamp (2011). "Where do illusory vowels come from?" *Journal of Memory and Language* 64.3, pp. 199–210.

Durvasula, Karthik, Ho-Hsin Huang, Sayako Uehara, Qian Luo, and Yen-Hwei Lin (2018). "Phonology modulates the illusory vowels in perceptual illusions: Evidence from Mandarin and English." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9.1, p. 7. DOI: http://doi.org/10.5334/labphon.57.

Durvasula, Karthik and Jimin Kahng (Dec. 2015). "Illusory vowels in perceptual epenthesis: the role of phonological alternations." *Phonology* 32 (03), pp. 385–416. ISSN: 1469-8188. DOI: 10.1017/S0952675715000263.

Durvasula, Karthik and Jimin Kahng (2016). "The role of phrasal phonology in speech perception: What perceptual epenthesis shows us." *Journal of Phonetics* 54, pp. 15–34. ISSN: 0095-4470. DOI: http://dx.doi.org/10.1016/j.wocn.2015.08.002.

Feldman, Naomi H and Thomas L Griffiths (2007). "A Rational Account of the Perceptual Magnet Effect." *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Austin, TX, pp. 257–262.

Gaskell, M. Gareth. and William D. Marslen-Wilson (1996). "Phonological variation and inference in lexical access." *Journal of Experimental Psychology: Human Perception and Performance* 22, pp. 144–158.

Gaskell, M. Gareth. and William D. Marslen-Wilson (1998). "Mechanisms of phonological inference in speech perception." *Journal of Experimental Psychology: Human Perception and Performance* 24.2, pp. 380–396.

Gonzales, Kalim and Andrew J. Lotto (2013). "A Bafri, un Pafri." *Psychological Science* 24.11. PMID: 24022652, pp. 2135–2142. DOI: 10.1177/0956797613486485. eprint: http://dx.doi.org/10.1177/0956797613486485.

Gow, David (2003). "Feature parsing: Feature cue mapping in spoken word recognition." *Perception & Psychophysics* 65.4, pp. 575–590.

Halle, Morris and William James Idsardi (1997). "/r/, Hypercorrection and the Elsewhere Condition." *Derivations and constraints in phonology*. Ed. by Iggy Roca. Oxford: Oxford University Press, pp. 367–392.

Hallé, Pierre A. and Catherine T. Best (May 2007). "Dental-to-velar perceptual assimilation: a cross-linguistic study of the perception of dental stop + /l/ clusters." *Journal of the Acoustical Society of America* 121.5 Pt1. PMC2773260[pmcid], pp. 2899–2914. ISSN: 0001-4966.

Hallé, Pierre A., Juan Segui, Uli Frauenfelder, and Christine Meunier (1998). "Processing of illegal consonant clusters: a case of perceptual assimilation?" *Journal of Experimental Psychology: Human Perception and Performance* 24, pp. 592–608.

Honeybone, Patrick (2005). "Diachronic evidence in segmental phonology: the case of obstruent laryngeal specifications." *The Internal Organization of Phonological Segments*. Ed. by M. van Oostendorp and J. van de Weijer. Berlin: Mouton de Gruyter, pp. 319–354.

Iverson, G K and J C Salmons (1995). "Aspiration and laryngeal representation in Germanic." *Phonology* 12, pp. 369–396.

Kabak, Baris and William James Idsardi (2007). "Perceptual distortions in the adaptation of English consonant clusters: syllable structure or consonantal contact constraints?" *Language and Speech* 50.1, pp. 23–52.

Kang, Yoonjung (2011). “Loanword Phonology.” *The Blackwell Companion to Phonology*. Ed. by Marc van Oostendorp, Colin Ewen, Elizabeth Hume, and Keren Rice. Vol. IV. Hoboken, N.J.: Wiley-Blackwell, pp. 2258–2281.

Kaye, Jonathan D. and Jean Lowenstamm (1984). “De la syllabicité.” *Forme sonore du language: structure des représentations en phonologie*. Ed. by D. Hirst F. Dell and J.-R. Vergnaud. Hermann, pp. 123–159.

Kratochvil, Paul (1968). *The Chinese Language Today*. London, China: Hutchinson.

Lawrence, Michael A. (2015). *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.3.

Lin, Yen-Hwei (2007). *The Sounds of Chinese*. Cambridge, UK: Cambridge University Press.

Luo, Chang-Pei and Jun Wang (1981). *Putong Yuyinxue Gangyao [Outline of General Phonetics]*. Beijing, China: Shangwu Yinshuguan.

Marr, David (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

McCarthy, John J. (1991). “Synchronic Rule Inversion.” *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Grammar of Event Structure*, pp. 192–207.

Miao, Ruiqin (2005). “Loanword Adaptation in Mandarin Chinese: Perceptual, Phonological and Sociolinguistic Factors.” Ph.D. Dissertation. Stony Brook, NY, USA: Stony Brook University.

Mitterer, Holger, Sahyang Kim, and Taehong Cho (2013). “Compensation for complete assimilation in speech perception: The case of Korean labial-to-velar assimilation.” *Journal of Memory and Language* 69, pp. 59–83.

Moreton, Elliott (2002). “Structural constraints in the perception of English stop-sonorant clusters.” *Cognition* 84, pp. 55–71.

Nespor, Marina and Irene Vogel (1986). *Prosodic Phonology*. Dordrecht: Foris.

Nevins, Andrew Ira (2005). “Conditions on (dis)harmony.” Ph.D. Dissertation. Cambridge, MA, USA: MIT.

Ohala, John J. (1981). “The listener as a source of sound change.” *Papers from the parasession on language and behavior*. Ed. by Robert A. Hendrick Carrie S. Masek and Mary Frances Miller. Chicago: Chicago Linguistic Society, pp. 178–203.

Ohala, John J. (1993). “The phonetics of sound change.” *Historical linguistics: problems and perspectives*. Ed. by Charles Jones. London: Longman, pp. 237–278.

Omar, Asmah H. (1977/1991). *Kepelbagaian fonologi dialek-dialek Melayu (2nd Ed)*. Kuala Lumpur: Dewan Bahasa dan Pustaka Kementerian Pendidikan Malaysia.

Peperkamp, Sharon (1999). "Prosodic Words." *GLOT International* 4.4, pp. 15–16.

Peperkamp, Sharon (2005). "A psycholinguistic theory of loanword adaptations." *Proceedings of the Berkeley Linguistics Society 30*, pp. 342–352.

Poeppel, David and Phillip J Monahan (2011). "Feedforward and feedback in speech perception: Revisiting analysis by synthesis." *Language and Cognitive Processes* 26.7, pp. 935–951.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Rosenthall, Samuel (1994). "Vowel/glide alternation in a theory of constraint interaction." Ph.D. Dissertation. Amherst, MA, USA: UMass Amherst.

Smith, Jennifer L. (2006). "Loan phonology is not all perception: evidence from Japanese loan doublets." *Proceedings of the 14th Annual Japanese/Korean Linguistics Conference*, pp. 63–74.

Sonderegger, Morgan and Alan Yu (2010). "A rational account of perceptual compensation for coarticulation." *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci10)*, pp. 375–380.

Wan, I-Ping (2003). *Alignments of prenuclear glides in Mandarin*. Taipei: Crane Publishing.

Weide, Robert L. (1994). *CMU Pronouncing Dictionary*. Available from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Weijer, Jeroen van de and Jisheng Zhang (2008). "An X-bar approach to the syllable structure of Mandarin." *Lingua* 118.9. Trends in prosodic phonology, pp. 1416–1428. ISSN: 0024-3841. DOI: https://doi.org/10.1016/j.lingua.2007.09.006.

Wells, Jonathan C. (1982). *Accents of English. Three volumes + cassette*. Cambridge, UK: Cambridge University Press.

Wells, Jonathan C. (1990). "Syllabification and allophony." *Studies in the Pronunciation of English, a Commemorative Volume in Honour of A. C. Gimson*. Ed. by S. Ramsaran. London and New York: Cambridge University Press, pp. 76–86.

Wickham, Hadley (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wilson, Colin and Lisa Davidson (2013). "Bayesian analysis of non-native cluster production." *Proceedings of NELS*. Ed. by S. Kan, C. Moore-Cantwell, and R. Staubs. Vol. 40, pp. 265–278.

Yanti (2010). "A Reference Grammar of Jambi Malay." Ph.D. Dissertation. Newark, DE, USA: University of Delaware.

Yip, Moira (1980). “The Tonal Phonology of Chinese.” Ph.D. Dissertation. Cambridge, MA, USA: MIT.

Yip, Moira (1982). “Reduplication and C-V Skeleta in Chinese Secret Languages.” *Linguistic Inquiry* 13.4, pp. 637–661. ISSN: 00243892, 15309150.

Yun, Suyeon (2016). “A Theory of Consonant Cluster Perception and Vowel Epenthesis.” Ph.D. Dissertation. Cambridge, MA, USA: Massachusetts Institute of Technology.

Zhang, Wei (2011). “THE LINGUISTIC SECRET OF FANQIE LANGUAGES IN CHINESE / 反切语的语言秘密.” *Journal of Chinese Linguistics* 39.2, pp. 345–369. ISSN: 00913723.

Zhao, Xu and Iris Berent (2016). “Universal Restrictions on Syllable Structure: Evidence From Mandarin Chinese.” *Journal of Psycholinguistic Research* 45.4, pp. 795–811. ISSN: 1573-6555. DOI: 10.1007/s10936-015-9375-1.