Phonology modulates the illusory vowels in perceptual illusions: evidence from Mandarin & English

Karthik Durvasula, Ho-Hsin Huang, Sayako Uehara, Qian Luo, and Yen-Hwei Lin

Michigan State University, USA

Corresponding Author Email: durvasul@msu.edu

Abstract

Native speakers perceive illusory vowels when presented with sound sequences that do not respect the phonotactic constraints of their language (Dupoux et al. 1999; Kabak and Idsardi 2007). There is, however, less work on the quality of the illusory vowel. Recently, it has been claimed that the quality of the illusory vowel is also modulated by the phonology of the language, and that the phenomenon of illusory vowels can be understood as a result of the listener reverse inferring the best parse of the underlying representation given their native language phonology and the acoustics of the input stream (Durvasula and Kahng 2015). The view predicts that listeners are likely to hear different illusory vowels in different phonological contexts. In support of this prediction, we show through two perceptual experiments that Mandarin Chinese speakers (but not American English speakers) perceive different illusory vowels in different phonotactic contexts. Specifically, when presented with phonotactically illegal alveo-palatal coda consonants, Mandarin speakers perceived an illusory /i/, but in illegal alveolar stop coda contexts, they perceived a /ə/.

Keywords: Speech Perception, Illusory Vowels, Mandarin, American English.

1. Introduction

The phenomenon of perceptual illusions, particularly that of illusory vowels, has become a very useful probe to understand both loanword patterns and the speech perception mechanism. In this article, we will argue, in line with some recent work, that we can use such perceptual illusions to get a better understanding of not just if, but also how, phonological knowledge is utilized during speech perception. More specifically, we suggest that, during speech perception, the listener attempts to identify 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 1998; Gow 2003; Mitterer et al. 2013, amongst others); this makes specific predictions about the quality of the illusory vowels; we test and corroborate one such consequence - that there can be different illusory vowels in different illicit phonotactic contexts.

Previous research has unearthed a variety of factors that affect where illusory vowels are perceived. They are perceived in stimuli that contain: (a) consonantal sequences that are phonotactically illicit in the native language of the speaker (Dupoux et al. 1999; Durvasula and Kahng 2016; Kabak and Idsardi 2007), (b) consonantal sequences that violate putative language universals such as the Sonority Sequencing Principle (Berent et al. 2008, 2009, 2007; Zhao and Berent 2016),[1] and (c) specific acoustic cues in the stimuli, such as strong consonantal release bursts, increased voicing amplitudes, *etc.* (Davidson 2007; Davidson and Shaw 2012).

In contrast, there is far less work and agreement in the prior literature on what modulates the quality of the illusory vowel perceived. Some have claimed that there can be only *one* illusory vowel which is either the *default* epenthetic vowel in the listener's native language (Uffman 2006) or "the phonetically minimal element of the language" (Dupoux et al. 2011). Others have argued that there can be multiple illusory vowels within the same language (Mattingley et al. 2015), even for the same phonotactic violation (Durvasula and Kahng 2015). There are five prior results that have a direct bearing on this issue. First, Dupoux et al. (2011) show that for the stimuli that contain consonantal sequences that are phonotactically illicit in both Japanese and Brazilian Portuguese, Japanese speakers perceive the illusory vowel [ɯ],[2] while Brazilian Portuguese speakers perceive the illusory vowel [i]. They note that the vowels [ɯ] and [i] are the shortest or "minimal" vowels in Japanese and Brazilian Portuguese, respectively, and suggest that perhaps only such "minimal" vowels can be genuine illusory vowels in a language. Here, it is important to note that, in some

of the tokens where the illicit consonantal sequence was created by splicing out the medial vowel (*e.g.*, [abda] from [abida]), they showed that Japanese speakers primarily perceived the spliced out vowel, and not the typical vowel [ɯ] in the same consonantal contexts when the sequence is naturally produced. They suggest that remnant co-articulatory traces in the spliced stimuli led to this particular result. Therefore, the vowels perceived in the spliced contexts are not really illusory vowels. This case should be kept separate from cases where there is no coarticulated information due to a spliced-out vowel to aid the listener, which was the case in their stimuli that were produced naturally with the consonant sequence violation (for e.g., [abda]). And in such items, consistent with the claim of participants perceiving the shortest or "minimal" vowel, the Japanese speakers primarily perceived an illusory [ɯ].

Second, similar to the above article, Guevara-Rukoz et al. (2017) recently probed Japanese speakers' perception of illusory vowels in [VhpV] and [VkpV] contexts, where the consonantal contexts were recorded naturally, or were from [VhVpV] and [VkVpV] contexts with the medial vowel (V) spliced out.[3] They showed that the co-articulatory traces of the spliced out vowel are a stronger predictor of the quality of the illusory vowel than the flanking vowels are, particularly for [hp] contexts. Their results suggest that co-articulatory cues play a role in the identification of the illusory vowel. It is important to note that it is difficult to infer anything stronger about co-articulatory cues being the *only* predictor of the quality of the illusory vowel, as a crucial cross-linguistic comparison language group was missing. The cross-linguistic comparison is important because if the results are completely due to co-articulatory cues, then there should be similar effects cross-linguistically. However, as discussed above, Dupoux et al. (2011) already showed that the same tokens can lead to *different* illusory vowels in different languages, in the case of naturally recorded stimuli.

Third, Monahan et al. (2009) attempted to obtain more than one illusory vowel for Japanese speakers. Based on loanword patterns in Japanese such as [makɯdonarɯdo] 'McDonald's', they conjectured that, for Japanese speakers, the illusory vowel next to non-coronal consonants (e.g., [k], [g]...) would be the vowel [ɯ], and that next to coronal consonants (e.g., [t], [d]...) would be the vowel [o]. However, they found that though Japanese speakers seem to perceive more illusory [ɯ] vowels than English speakers in non-coronal contexts, *e.g.*, [egma], the Japanese speakers did not seem to show any increased levels of illusory [o] in coronal contexts, *e.g.*, [etma]. They therefore were not able to get the second illusory vowel that they expected for Japanese speakers.

Fourth, Mattingley et al. (2015) also probed the same issue in Japanese speakers. As Mona-

han et al. (2009) did before them, they expected different illusory vowels in different consonantal contexts based on the loanword patterns in Japanese, but they explored a larger set of phonotactic contexts than Monahan et al. (2009). Interestingly, though their Japanese participants perceived different vowels in different illicit consonantal sequences, the illusory vowels that they perceived were not always isomorphic with those expected based on the loanword patterns. Specifically, they found that in non-palatal contexts, [ɯ] was by far the most common illusory vowel perceived, and in palatal contexts, [i] was the illusory vowel perceived. One thing to note though is the fact that they did not include a control group, therefore it is not immediately clear if the Japanese speakers were really hearing illusory vowels, or if there were confounds in the stimuli that encouraged these patterns of perception. This is possible given that the speaker who produced the relevant stimuli was a native speaker of Japanese, who was also fluent in English. In which case, the speaker produced sequences that are illicit in their native language, and might have therefore articulated them with very short excrescent vowels, which could have given rise to the observed patterns in the experiment.

Finally, Durvasula and Kahng (2015) tested Korean speakers' perception of illicit consonantal sequences with English speakers as controls. They showed that, unlike the English speakers presented with the same stimuli, the Korean speakers were perceiving different illusory vowels in different consonantal contexts, and different illusory vowels in some cases for the same context too. More specifically, they showed that in alveolar contexts, Korean speakers perceived the illusory vowel [ɨ]; adjacent to the palatal fricative [ɕ], they perceived the illusory vowel [i]; and adjacent to palatal stops [c], they perceived both [ɨ] and [i] (on different trials of the same stimuli).

It is also important to note that while Guevara-Rukoz et al. (2017), Monahan et al. (2009) and Mattingley et al. (2015) were basing their prior experimental expectations on loanword patterns observed in Japanese, Dupoux et al. (2011) and Durvasula and Kahng (2015) were basing their expectations on predictions made by their particular views of speech perception. As pointed out by many authors (Jong and Cho 2012; LaCharité and Paradis 2000; Peperkamp et al. 2008; Smith 2006; Vendelin and Peperkamp 2006, amongst others), there can be multiple sources through which loanwords can enter a language, with normal auditory input as just one possible source. Therefore, to use loanwords to understand speech perception, one needs to first factor out the loanword patterns that have entered the language through other channels, and then use the remaining corpus to identify the relevant patterns. However, such historical sources are often not

available, and thus, using loanwords to make predictions about illusory vowels is not straightforward. In contrast, the predictions by Dupoux et al. (2011) and Durvasula and Kahng (2015) stem from proposals about the speech perception mechanism, and are therefore more relevant to our understanding of the phenomenon of illusory vowels. It is for this reason that, in what follows, we first lay out a certain view of speech perception and then test out the consequences of the view for illusory vowel patterns.

The phenomenon of illusory vowels generally and the quality of the illusory vowel itself have also received an Optimality-Theoretic analysis by Boersma and Hamann (2009). While they were primarily interested in understanding loanword patterns through the lens of first-language speech perception, they suggest, in a discussion of the illusory vowel phenomenon in Korean speakers, that the illusory vowel is likely to be /ɨ/, because it is involved in optional deletion processes. A similar observation relating specific language-specific processes in Japanese (namely, vowel shortening and devoicing) to the illusory vowel /ɯ/ was made as far back as Dupoux et al. (1999). This is an issue that we will return to below, when we present our own viewpoint and predictions.

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 (Gaskell and Marslen-Wilson 1996, 1998; Gow 2003; Mitterer et al. 2013, amongst others). Previously, the viewpoint has largely been employed in understanding perceptual compensation for assimilatory changes, particularly, at the edges of words. For example, the phrase 'garden bench' /gɑːdn bɛntʃ/ is often pronounced as [gɑːdm bɛntʃ],[4] where the word-final nasal /n/ has assimilated to the place of articulation of the following segment. It has been shown 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]). Durvasula and Kahng (2015) extend such views to understand the phenomenon of illusory vowels.

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. More specifically, in regard to the quality of the illusory vowel, the perceiver's task is an attempt to repair the illicit phonotactic sequence with a (vowel) phoneme that best maps to the phonetic characteristics of the acoustic token. When no relevant phonological alternations bias listeners towards a certain vowel in the particular segmental context, the best vowel guess that repairs the particular phonotactic violation could either be based on co-articulatory cues present in the relevant content or be the phonetically minimal/shortest vowel in the inventory, *à la* Dupoux et al. (2011). However, when relevant phonological alternations do bias listeners towards particular vowel percepts in specific segmental context, the best guess depends on both the phonetics of the acoustic token and also the phonological alternations themselves. The types of phonological processes that are likely to play a role are those that bias the listener's expectations about the quality of the illusory vowel. One such process is a consistent/regular vowel deletion process that targets a particular vowel. The presence of a regular process of vowel deletion that targets a particular vowel ($/V_1/ \rightarrow [\emptyset]$) in the phonology of the language supports the reverse inference of the same vowel in the phonemic representation when the surface representation has nothing (reverse inference: $[\emptyset] \rightarrow /V_1/$). For these reasons, in a phonotactically illicit consonantal context, where the condition can be perceptually repaired by a vowel, the best vowel to repair the situation is the phoneme $/V_1/$ that maps to [$\emptyset$] in the surface/acoustic representations. As mentioned above, similar or related views, linking language-specific deletion/reduction processes to the quality of the illusory vowel, were presented by Boersma and Hamann (2009) and Dupoux et al. (1999).

A second type of process that is likely to bias a listener's expectations about the vowel quality of the illusory vowel is one that involves a phonotactic restriction ($[C_1]$ is allowed only next to $[V_2]$).[5] 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 vowel, then the best vowel to perceptually repair the sequence is the vowel $/V_2/$, as this would also account for the acoustic properties of the illicit consonant; this is so because the consonant [C_1] can only appear before a vowel [V_2] in the language.

1.1. Relevant phonological patterns and predictions

In what follows, we briefly describe some phonological patterns in Mandarin and English that are relevant for the phonological contexts tested in this paper. We then present specific predic-

tions that stem from the relevant language-specific facts and the viewpoint of speech perception presented above.

In Mandarin, there are no obstruent coda consonants. Therefore, consonantal sequences where the first consonant is an obstruent consonant are not possible [*atɕʰma, *atʰma]. This constraint is true both within words (so across syllables) and across higher phonological domains: *i.e.*, a Mandarin speaker is not likely to hear such sequences at all, in a random speech stream. Following the prior work on the topic of illusory vowels, word-medial obstruent-nasal clusters are thus contexts that are ripe for illusory vowel perception in Mandarin Chinese.

Some of the phonological patterns in Mandarin exhibit exactly the characteristics needed to bias the perception of the illusory vowels in the relevant contexts. Weinberger (1996) discussed one such phonological process of vowel deletion that targets the vowel /ə/ in certain fast conversational-speech contexts: an unstressed word-final /ə/ is deleted when preceded by a nasal consonant (1).[6] This allows /ə/ to be a good vowel for perceptual repairs in most contexts because it already varies with ∅ (nothing) in the surface phonetic representations. Furthermore, /ə/ is often toneless (and stressless) in Mandarin (Weinberger 1996). Therefore, inferring such a vowel phoneme allows for the listener to straightforwardly account for the lack of tone (or stress) between two consonants that form an illicit sequence. For these reasons, /ə/ is an excellent candidate for an illusory vowel - we call it ***illusory vowel 1***.[7] This can also be thought of as the *default* illusory vowel in Mandarin, as this is the illusory vowel expected in contexts where there are no other influencing phonological processes.

1. Fast-speech vowel elision in Mandarin Chinese (Weinberger 1996)

    - /ʂə́mə/ → [ʂə́m] ‘what’
    - /tˢə̌mə/ → [tˢə̌m] ‘how’

A second pattern in Mandarin that can bias the perception of illusory vowels has to do with consonant-vowel phonotactics. While alveolar stops as a group can precede all vowels [✓tʰi, ✓tʰa, ✓tʰu, …], alveo-palatal consonants can only appear before high front vowels or glides [✓tɕʰi, *tɕʰa, *tɕʰu, …] (note: /tʰ/ specifically seems to be absent before /e/ based on the corpus results presented in Tables 4-5). These facts allow front vowels, particularly /i/, to be a good vowel for perceptual repairs in illicit alveo-palatal coda contexts (but not in illicit alveolar coda contexts). We call the high front vowel /i/ ***illusory vowel 2***.

Now, we turn to the relevant phonological and phonetic facts in English. In English, there is no alveo-palatal consonant such as [tɕʰ], but there are palato-alveolar consonants such as [tʃʰ].[8] So, when an English speaker is presented with [tɕʰ], they are likely to perceive it as /tʃʰ/ - this is expected based on the results of non-native perception discussed by Best et al. (2003). Furthermore, both alveolar and palato-alveolar consonants are possible in coda positions in English. Therefore, both [tʃʰm] and [tʰm] are phonotactically valid sequences. One might object that /tʰ/ in coda positions typically glottalize, but they do appear variably as [tʰ] (Kreidler 1989; Vaux 2002), particularly across compound members and across words in careful or enunciated speech in our own experience. So, [tʰm] is a possible sequence/pronunciation in English. As a consequence, sequences such as [atɕʰma], [atʰma] have perfectly licit phonemic parses in the language. Though not pursued here, the viewpoint presented in this article predicts that /ə/ should be the main illusory vowel for English speakers in illicit phonotactic contexts, as it regularly deletes in fast/casual-speech processes (Hooper 1978).

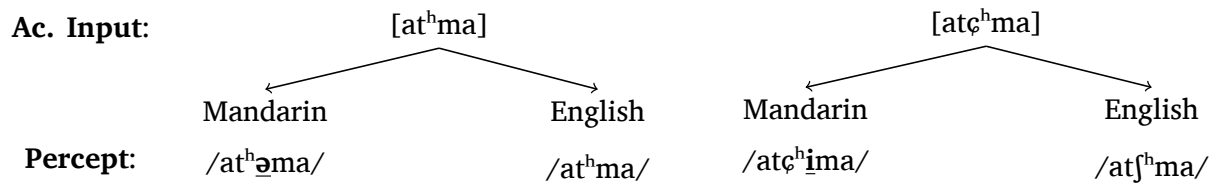

Figure 1: Predicted percepts for different acoustic inputs

In light of the above phonological facts from Mandarin and English, we present the predictions that the viewpoint presented earlier makes for particular illicit sequences. We expect that when Mandarin speakers are presented with an illicit consonantal sequence involving an alveo-palatal affricate (e.g., [tɕʰm]), the illusory vowel perceived by the listener should be biased by the phonology; as in this context, only high front vowels are possible. While nothing directly stemming from the viewpoint presented above motivates a more specific prediction than a high front vowel, one could argue based on markedness considerations that that high front vowel must be /i/ ***illusory vowel 2*** (note: apart from /i/, there is a second high front vowel in Mandarin, namely, the rounded counterpart /y/. It is possible that /i/ is preferred for acoustic reasons; if so that is consistent with the reverse inference view that we suggest here. However, without precise acoustic recordings and models, it is difficult to sustain such a claim currently, so we haven't explicitly argued for it.). Furthermore, when Mandarin speakers are presented with an illicit consonantal sequence involving an alveolar stop (e.g., [tʰm]), no particular phonological pattern in

Mandarin has a direct influence on the identification of the illusory vowel. In such a context, they are expected to perceive the more general ***illusory vowel 1***, which happens to be /ə/.

In contrast, when an English speaker is presented with either of the above sequences, *i.e.*, [tɕʰm] or [tʰm], no illusory vowel is expected due to the phonology of the language, as both sequences are possible (albeit, across higher prosodic domains than just a syllable boundary). However, the presence of confounding acoustic factors in the stimuli, such as extremely strong burst release or aspiration, might induce some illusory vowels even for English speakers (along with Mandarin speakers).

There are two important caveats to bear in mind about the above predictions. First, the prediction that Mandarin speakers should hear an illusory /i/ in alveo-palatal contexts, but a /ə/ in alveolar contexts is a prediction that holds true if only phonological knowledge were deployed (essentially, it is a *mutatis mutandis* prediction). However, all naturalistic stimuli are likely to carry subtle artifacts. For example, it is possible that the Mandarin speakers will have some /ə/ illusions in illicit alveo-palatal contexts for reasons related purely to the acoustic properties of the particular stimuli in the experiment. Therefore, the above prediction can also be seen as saying that, for Mandarin speakers, the /i/ illusions are going to be higher in illicit alveo-palatal contexts than illicit alveolar contexts; or conversely, the /ə/ illusions are going to be higher in illicit alveolar contexts than illicit alveo-palatal contexts.[9] Second, the possibility that artifactual acoustic effects in the stimuli could trigger illusory vowels even for English speakers entails that the appropriate predictions are not *within-language* predictions, but *between-language* predictions. Such predictions would control for any effects due to acoustic artifacts in the stimuli. Therefore, the appropriate predictions for us are that Mandarin speakers, compared to English speakers, should perceive *more* /i/ in [tɕʰm] contexts, and more /ə/ in [tʰm] contexts. Consequently, Mandarin speakers should confuse [tɕʰm]∼[tɕʰim] and [tʰm]∼[tʰəm] more than the English speakers.

In the following sections, we test these predictions out using two perceptual experiments, an ABX task (Experiment 1), and an identification task (Experiment 2).

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty native Mandarin speakers from Beijing (mean age = 20.5 years, 7 men and 13 women) and 19 native English speakers from Michigan (mean age = 21 years, 6 men and 13 women) participated in the experiment. All the participants were recruited at Michigan State University. The Mandarin participants spent an average of 2.4 years (SD = 1.6 years)[10] 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. The Mandarin participants received no compensation for their participation, but the English participants received extra credit for their participation.

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 [i / ə / ∅ (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. The speaker was chosen to avoid providing either linguistic group with stimuli from their native dialect/language.

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, as many 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 (in the form of very short excrescent vowels), 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 those that we tried had difficulty in producing unstressed medial vowels that were unreduced (*i.e.*, they could not block the vowel-reduction process in their dialect), and 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 results would therefore have been 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. Instead, we are interested more generally in speech perception, and particularly, in how Mandarin speakers, but not American English speakers, are expected to hear more illusory vowels under different acoustic circumstances. In this latter case, the use of an American English speaker to record the stimuli is not obvious and has to be considered in the context of what confounds are raised. It is also not our intention to probe any language-neutral perception. However, the use of an American English speaker would have asymmetrically made the task easier for the English participants, as the acoustics in the non-crucial portion of the tokens would have been easier to interpret, thereby allowing them to potentially focus more on the crucial environment. Such a confound makes it difficult to interpret any differences between the English and Mandarin speakers' perceptual responses.)

There are three more issues worth discussing in detail with respect to the stimuli and the speaker.[11] 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 can be reasonably close, acoustically, to alveo-palatal stops; However, we depended on three 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 T3 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, T3 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 T3 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. Third, it is also reasonable to ask how natural the medial vowel in the stimuli were for two language groups. Again, we initially depended on impressions of the group of Mandarin linguists we consulted (mentioned above) to establish the naturalness of the vowel quality. Anticipating the results of Experiment 2 here, we note that the results presented in Figure 5 suggest that both the English and Mandarin speakers were at reasonably high levels (and not much different from each other) at identifying the medial vowels.

Each item was recorded several times using the software 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 perceptual epenthesis effect. We tested all combinations of vowels [i, ə, ∅]. For example, for the cluster [t^hm], the AB word-pairs were [at^hima∼at^hma], [at^həma∼at^hma], and [at^hima∼at^həma].[12] There were two recordings used for each item and the order of tokens in an AB sequence was counterbalanced. For instance, in the case of [at^hima∼at^hma], there were four AB sequences [$at^h ima_1$-$at^h ma_1$], [$at^h ima_1$-$at^h ma_2$], [$at^h ima_2$-$at^h ma_1$], [$at^h ima_2$-$at^h ma_2$], 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^hima∼at^hma], there were eight ABA triplets [$at^h ima_1$-$at^h ma_1$-$at^h ima_2$], [$at^h ima_1$-$at^h ma_2$-$at^h ima_2$], [$at^h ima_2$-$at^h ma_1$-$at^h ima_1$], [$at^h ima_2$-$at^h ma_2$-$at^h ima_1$], [$at^h ma_1$-$at^h ima_1$-$at^h ma_2$], [$at^h ma_1$-$at^h ima_2$-$at^h ma_2$], [$at^h ma_2$-$at^h ima_1$-$at^h ma_1$], [$at^h ma_2$-$at^h ima_2$-$at^h ma_1$], and an additional eight ABB triplets [$at^h ima_1$-$at^h ma_1$-$at^h ma_2$], [$at^h ima_1$-$at^h ma_2$-$at^h ma_1$], [$at^h ima_2$-$at^h ma_1$-$at^h ma_2$], [$at^h ima_2$-$at^h ma_2$-$at^h ma_1$], [$at^h ma_1$-$at^h ima_1$-$at^h ima_2$], [$at^h ma_1$-$at^h ima_2$-$at^h ima_1$], [$at^h ma_2$-$at^h ima_1$-$at^h ima_2$], [$at^h ma_2$-$at^h ima_2$-$at^h ima_1$]. So, there were a total of 48 triplets constructed from the alveolar stimuli ([at^hma, at^həma, at^hima]). Similar combinations were used to create another 48 triplets for the alveo-palatal stimuli ([at$ɕ^h$ma, at$ɕ^h$əma, at$ɕ^h$ima]). 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^hima∼at^həma] and [at$ɕ^h$ima∼at$ɕ^h$əma] for experimental reasons to ensure a reasonable number of clearly different stimuli, and as a sanity check. With respect to the latter concern, if the Mandarin speakers confused the vowel pairs at much higher rates than the English speakers, then it would be difficult to see if the Mandarin speakers really heard a particular vowel in a certain context. As will be obvious later in Figure 2, the Mandarin speakers were at least as good as the English speakers with these pairs.)

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

A visual inspection of the mean percentage of correct responses to the stimuli by both the Mandarin and English speakers (Figure 2) suggests the following: (a) the Mandarin speakers appear to be worse at distinguishing [atɕʰima~atɕʰma], [atɕʰəma~atɕʰma], and [atʰəma~atʰma]; (b) they also appear to be no different from the English speakers with respect to [atɕʰəma~atɕʰima]; (c) the Mandarin speakers appear to be slightly worse than the English speakers for the pair [atʰima~atʰma]; (d) the English speakers appear to be slightly worse than the Mandarin speakers with [atʰima~atʰəma].

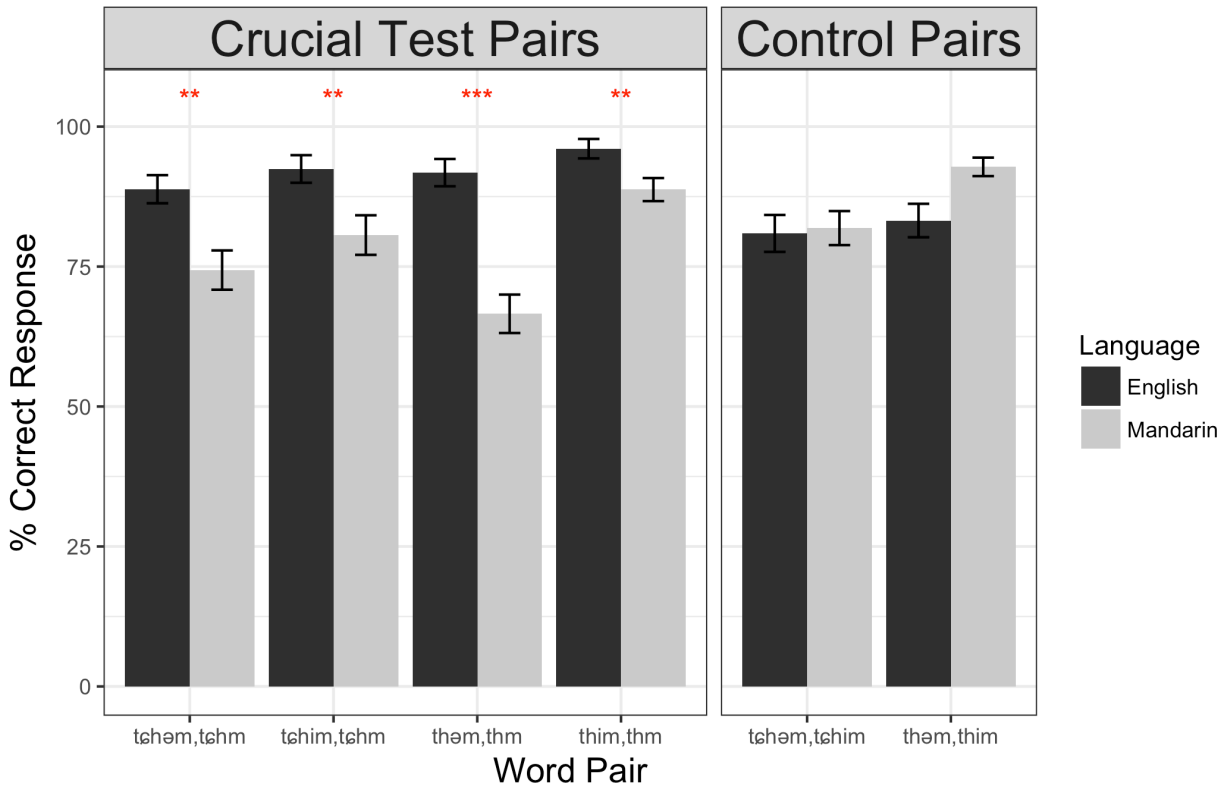

Figure 2: Mean % correct for English and Mandarin speakers in Exp. 1 (error bars = 1 S.E.; * = $p<0.05$; ** = $p<0.01$; *** = $p<0.001$; asterisks represent p-values for between-language comparisons of only the crucial word-pairs)

In order to confirm the observations made by visual inspection of the results, we followed up with statistical analysis in R (R Development Core Team 2014). A three-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 two within-subjects factors, CONSONANT (tɕʰ, tʰ) and VOWELPAIR (i∼nothing, ə∼nothing).[13] Mauchly's test revealed no violations of the assumption of sphericity. There was a main effect of LANGUAGE [$F(1,37) = 23.87$, $p < 0.0001$]. There was also a main effect of VOWELPAIR [$F(1,37) = 42.01$, $p < 0.0001$]. There was a two-way interaction of LANGUAGE and VOWELPAIR [$F(1,37) = 13.04$, $p < 0.001$]. There was a second two-way interaction of CONSONANT and VOWELPAIR [$F(1,37) = 8.01$, $p < 0.01$]. Finally, there was a three-way interaction of LANGUAGE*CONSONANT*VOWELPAIR [$F(1,37) = 6.48$, $p = 0.015$]. The interaction suggests that the Mandarin and English speakers had different responses to different vowel pairs after different consonants.

To investigate the results from the ANOVA further, we conducted pairwise Mann Whitney U tests (Table 1). These non-parametric tests were conducted as the dependent variable was a proportion, and hence the assumption of normality made by t-tests was violated. Further note, ANOVAs are typically understood to be more robust to such normality violations, so we use them here to maintain ease of interpretability for the reader. We present only the crucial between-language comparisons, namely, those involving [vowel∼no vowel] pairs. The pairwise tests suggest that Mandarin speakers were confusing all the cruial pairs more than the English speakers: [atɕʰəma∼atɕʰma], [atɕʰima∼atɕʰma], [atʰəma∼atʰma], [atʰima∼atʰma].

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

Comparison	Mean diff (%) [Mand.-Eng.]	W	Pr(>|z|)	
atɕʰəma∼atɕʰma	14.44	297.5	0.002	**
atɕʰima∼atɕʰma	11.81	282.5	0.007	**
atʰəma∼atʰma	25.21	346.5	< 0.0001	***
atʰima∼atʰma	7.30	286.5	0.004	**

It is important to recognize that the between-language comparisons are indeed the crucial comparisons. For example, it is possible to observe Figure 2 and infer that there is a tendency for Mandarin speakers to confuse [atɕʰəma∼atɕʰma] more than [atɕʰima∼atɕʰma];[14] however, to make sure that this difference doesn't arise from stimulus artifacts, it is important to compare the

results directly with the English speakers. Looking at the English speakers' responses in Figure 2 suggests they have a similar increase in the degree of confusion for [atɕʰəma∼atɕʰma] compared to [atɕʰima∼atɕʰma], though they were not expected to show any such difference. This suggests the increased Mandarin confusion for [atɕʰəma∼atɕʰma] over [atɕʰima∼atɕʰma] should not be interpreted as how the Mandarin speaker's phonological knowledge affects perception. More generally, we want to reiterate that, in order to probe truly phonological knowledge in speech perception, one needs to guard against stimulus artifacts, which practically all naturally-recorded stimuli are bound to have, by conducting between-language comparisons. Within-language comparisons, in our opinion, cannot guard against such issues.

2.3. Discussion

In Experiment 1, Mandarin speakers confused [atɕʰima∼atɕʰma] and [atʰəma∼atʰma] more than English speakers. This is in line with our expectation that Mandarin speakers, but not English speakers, are biased towards some illusory vowels due to the phonology of their native language.

There were also a few unexpected results. First, Mandarin speakers also confused [atɕʰəma∼atɕʰma] and [atʰima∼atʰma] more than English speakers. In understanding these unexpected results, it is important to note that the ABX task is not a direct task looking at illusory vowels in the crucial test stimuli, but is instead fundamentally comparative. For example, it is possible that for Mandarin speakers [atɕʰəma] might have been perceived as /atɕʰima/ due to the phonotactic patterns of Mandarin; [atɕʰəma] was therefore confusable with [atɕʰma], which was also perceived as /atɕʰima/ due to an illusory vowel. Such a possibility while not initially considered by us is consistent with the overall focus of the explanation residing in the listener's knowledge of Mandarin phonology. The [atʰima∼atʰma] difference on the other hand is not amenable to such an explanation. Here, it is worth highlighting that the Mandarin speakers' correct responses are not as dramatically different from the English speakers. Furthermore, the percentage of correct responses is quite high for Mandarin speakers compared to other stimulus pairs including the stimulus pairs with different vowels, so it is possible that the Mandarin responses have hit some sort of ceiling effect related to the non-nativeness of the stimuli. However, we acknowledge that this is not entirely clear. Crucially, Experiment 2, to be discussed below, confirms that Mandarin listeners do not hear an /i/ in [atʰma]. Therefore, Mandarin listeners are confusing [atʰima∼atʰma] because of some other perceptual reason than perceiving an /i/ in [atʰma]. It is important to note that our prediction of Mandarin listeners not confusing [atʰima∼atʰma] was

based on the prediction that, if no other perceptual changes are there, then [at^{h}ima∼at^{h}ma] will not be confused because Mandarin listeners do not infer an /i/ in [at^{h}ma]. Experiment 2's results confirm this crucial aspect of the prediction.

Quite unrelated to our immediate question, there was also a tendency for the English speakers to confuse the [at^{h}əma∼at^{h}ima] more than the Mandarin speakers. This again might be due to phonological reasons; in this case, reasons related to English phonology. The phenomenon of unstressed vowel neutralization that has been extensively observed in English (Burzio 2007; Chomsky and Halle 1968) might have negatively influenced the English speakers' performance. As with the first unexpected result, it is worth pointing out that such an explanation is consistent with our broad view of the role of phonological knowledge, and deserves further scrutiny of its own in future work.

The main issue raised by the unexpected results in Experiment 1 is that ABX tasks are fundamentally indirect (or comparative), and perceptual illusions not only in the expected token, but also in the comparison tokens, might lead to incorrect responses. It is therefore important to attempt an experiment that focusses directly on the illusory vowels in the crucial test items, *i.e.*, the items with the illicit consonantal sequences [at^{h}ma] and [atɕ^{h}ma]. Experiment 2 addresses this need with a more direct identification task.

3. Experiment 2

3.1. Methods

3.1.1. Participants

The participants in Experiment 2 were the same as in Experiment 1. Experiment 2 was an identification task that drew participants' attention to the medial vowel in the stimuli. Therefore, it was conducted after Experiment 1 (after a short break) so as not to have the participants to focus on only the vowel in Experiment 1.

3.1.2. Materials

The stimuli were the same 6 test items used in Experiment 1 as described in Section 2.1.2. There were two recordings used for each item, as in Experiment 1, and they were each presented thrice; therefore, there were 6 tokens of each test item, and a total of 36 tokens in the experiment, presented in pseudo-randomized order with the added constraint that there be no identical test

items in succession.

An issue that is important to acknowledge here is that of experimental power, in both Exp. 1 & 2, related to the number of repetitions of test items (or pairs, in Exp. 1).[15] The issue of power is always a concern in any experiment, especially when null results are involved, and it is possible that the power in the current experiments might have increased if we had increased the repetitions. But, when we planned the experiments, given that the participants would take part in both the identification task and the ABX task, we were concerned about learning effects of many presentations, whereby participants might adopt complex response strategies in response to the stimuli in the experiment. To us, this is still a largely underexplored topic, and we wanted to err on the side of caution with respect to learning effects. Furthermore, given that the responses to the repetitions were very likely not independent of each other, it is not necessary that adding more repetitions would have increased the power of the experiments –if anything, a much more substantial increase in power in such experiments is likely to come from increasing the number of participants, we think.

3.1.3. Procedure

In Experiment 2, we used an identification task 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 the medial vowel was [i], another vowel, or nothing and click on the corresponding box on the screen with a mouse (the actual choices were [i], [other vowel], or [no vowel] for English speakers, and [拼音韵母 i], [其他韵母], or [没有韵母] for Mandarin speakers). We chose to provide the participants with the option [other vowel] because both in English and Mandarin, there is no unique letter to identify a schwa. The letter "e" could have been used, but it is still minimally ambiguous between [ə] and [e]. Furthermore, though not immediately relevant to us, the Pinyin character "i" can also stand for the [ɻ̩] after some fricatives/affricates.

All the instructions were in English for the English speakers ("What's between the two consonants?") and in Mandarin for the Mandarin speakers ("两个声母中间有什么?"). 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 [m/r]. The inter-trial interval was 1500ms. All 36 trials were randomized for each participant.

3.2. Results

A visual inspection of the mean percentage of responses to the stimuli by both the Mandarin and English speakers suggests that *all* and *only* the expected differences were found between the two language groups (Figure 3). As expected: (a) the Mandarin speakers appear to choose more [i] responses and fewer 'no vowel' responses for [atɕʰma]; (b) the Mandarin speakers appear to choose more 'other vowel' responses and fewer 'no vowel' responses for [atʰma].

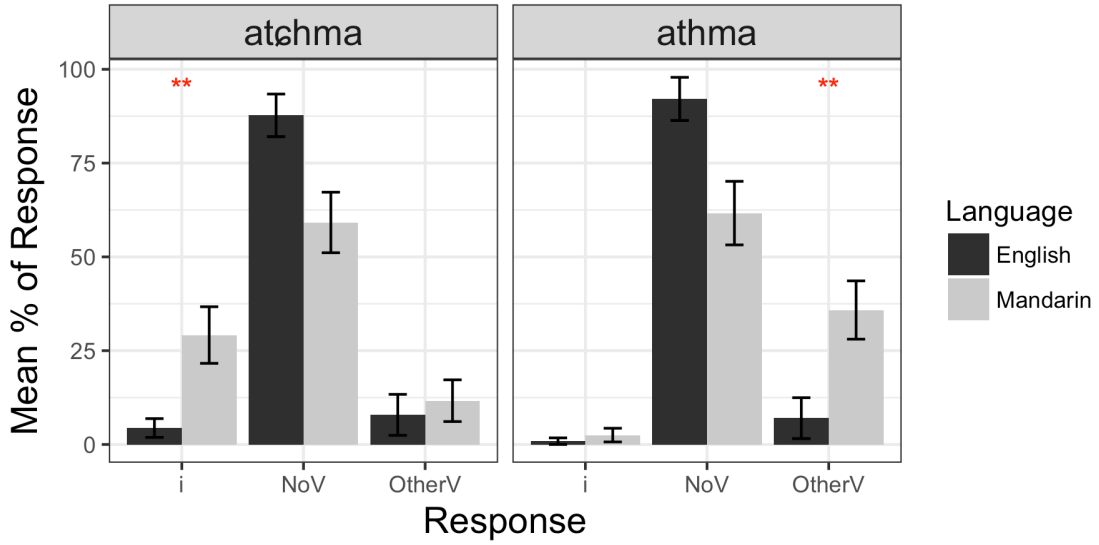

Figure 3: Mean % of responses of English and Mandarin speakers for the crucial test items in Exp. 2 (error bars = 1 S.E.; * = $p<0.05$; ** = $p<0.01$; asterisks represent p-values for between-language comparisons of only the vowel responses)

In order to confirm the observations made by visual inspection of the results, we followed up with statistical analysis. A three-way mixed ANOVA was run to model the percentage of responses as the dependent variable with LANGUAGE (English, Mandarin) as a between-subjects factor; and STIMULUS ([atʰma], [atɕʰma]) and RESPONSE ([i], [other vowel], [no vowel]) as within-subject factors. Mauchly's test revealed there were violations of the assumption of sphericity for the main effect of RESPONSE and the interaction LANGUAGE*RESPONSE. Both effects were corrected with Greenhouse-Geisser correction for the degrees of freedom ($\epsilon = 0.66$). There was a main effect of RESPONSE [$F(0.66,24.4) = 68.68$, $p < 0.0001$]. There were two-way interactions for LANGUAGE*RESPONSE [$F(1.32,48.8) = 8.64$, $p < 0.01$] and RESPONSE*STIMULUS [$F(2,74) = 6.24$, $p < 0.01$]. Finally, there was a three-way interaction effect for LANGUAGE*STIMULUS*RESPONSE [$F(2,74) = 4.64$, $p < 0.05$]. The three-way interaction suggests that the English and Mandarin

speakers were giving different percentages of responses for different stimuli.

To check if, overall, the Mandarin Speakers hear more illusory vowels than English speakers regardless of their quality, we conducted a simple Mann Whitney U test with the percentage of vowel responses for the stimuli with no medial vowel as the dependent variable, and the LANGUAGE as the independent variable.[16] The results suggested, in accordance with visual observation, that there was indeed a considerable difference between the two language groups with respect to illusory vowels heard in the crucial stimuli [Mean diff (Mand.-Eng.) = 29.50, $W = 393$, $p < 0.0001$]. A closer look at the responses of the Mandarin and English participants revealed the following descriptive statistics. For the [atɕʰma] stimuli, only 4 English speakers had any vowel responses (a majority of these were [other vowel] responses), while 13 Mandarin speakers did so too. For the [atʰma] stimuli, only 1 English speaker had any vowel responses, while 12 Mandarin speakers did so too. (Note, the vowel responses by each Mandarin speaker was generally larger in number than most of the English speakers). Furthermore, a look at the participants revealed that 9 of the Mandarin participants overlapped for the [atʰma] and [atɕʰma] stimuli; Therefore, overall 16 of the Mandarin participants gave some vowel responses to the crucial stimuli. In contrast, only overall 4 English participants gave such responses (the one participant who gave some vowel responses for [atʰma] also gave them for [atɕʰma]). We also tried to probe if there was any common element to the four Mandarin speakers who didn't give any vowel responses, and didn't find them to be exceptional in either the amount of time spent in the US, or in the duration of exposure to English as a second language in a classroom environment in China. Finally, we tried to see if the [no vowel] responses by the Mandarin speakers were related to the duration of their English exposure in China or the time they spent in the US. Figure 4 shows there is no observable trend.

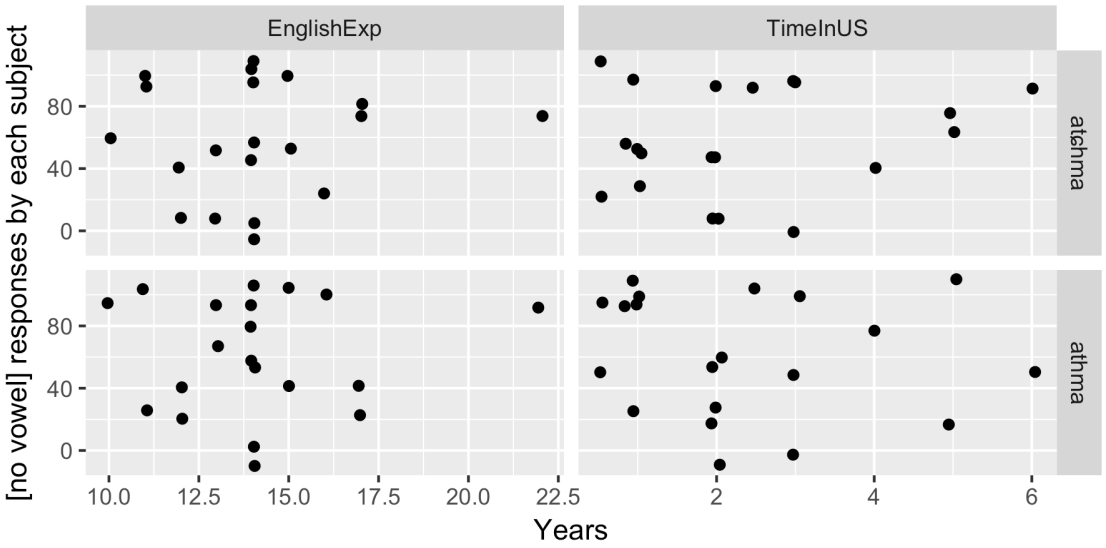

Figure 4: Scatterplots of the proportion of [no vowel] responses by the Mandarin speakers (y-axis) with either duration of exposure to English as a second language in a classroom environment in China (EnglishExp) or time in the US (TimeInUS) as the x-axis. (note: jitter has been added to the plot to show overlapping values.)

As mentioned in the *Introduction,* the crucial planned comparisons are the ones that compare the responses by the two language groups to each type of stimulus without a medial vowel, as such a between-language comparison controls for any subtle artifacts in the stimuli. Therefore, we conducted pairwise Mann Whitney U tests (Table 2). Since there are three possible responses, and two different crucial test stimuli, there are 6 total pairwise comparisons across language groups. However, it is meaningless to simultaneously test the identification rates across all three responses; this is because the responses to the [no vowel] response are not independent of those to the other two responses, *i.e.*, [i] and [other vowel]. Therefore, we conducted only four pairwise comparisons - two for each of the vowel responses. The pairwise tests suggest that Mandarin speakers were identifying more [i] vowels in [atɕ^{h}ma], and more of another vowel in [at^{h}ma].

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

Stimulus	Response	Mean diff (%) [Mand.-Eng.]	W	Pr($>\|z\|$)	
atɕʰma	i	24.78	106.0	0.006	**
	other vowel	3.77	157.0	0.200	
atʰma	i	2.50	180.5	0.584	
	other vowel	35.80	97.0	0.002	**

We also include further pairwise analyses looking at a different question: Is it the case that Mandarin speakers perceive more /i/ than other vowels in the alveo-palatal context, and more other vowels in the alveolar context, while there are no such differences for English speakers?[17] However, it is important to note that this question, which we include for the sake of completeness, requires within-language comparisons; therefore, unlike the results in (Table 2), this way of asking the question does not control for stimulus artifacts directly.

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

Language	Stimulus	Mean diff (%) [i-other vowel]	W	Pr($>\|z\|$)	
Mandarin	atɕʰma	17.5	261	0.068	.
	atʰma	–33.33	90	< 0.0001	***
English	atɕʰma	–3.51	187	0.77	
	atʰma	–6.14	170	0.53	

The within-language comparisons (Table 3) could be interpreted as showing that the Mandarin speakers respond with many more [other vowel] responses (so, more [ə]) than [i] when presented with [atʰma], and with a statistically marginally-significant difference between [other vowel] and [i] responses in the case of [atɕʰma]. For the latter case, given that the difference of 17.5% (favouring the [i] responses) is in our opinion quite sizeable, we interpret the result as the Mandarin speakers favouring [i] responses. However, there do seem to be some unexpected [other vowel] responses by the Mandarin speakers in [atɕʰma]; this could be due to subtle phonetic aspects of the stimuli, and only a direct comparison with the responses by control English group allows us to see if this is the case. When the comparison between the English speakers (who are

expected to have very low levels of vowel responses) and the Mandarin speakers is made for the [i] and [other vowel] responses (Table 2), it is clear that, for the [atɕʰma] stimuli, the Mandarin speakers responded with more [i] responses than the English speakers (Mean Diff = 24.78%), but they didn't respond with more [other vowel] responses than the English speakers. We take this as evidence that the Mandarin speakers typically hear [i] in alveo-palatal contexts [tɕʰm].

Finally, just as a final sanity check, we also looked at the participants' responses to the stimuli with medial vowels.[18] The results are shown below in Figure 5. We conducted statistical analyses, paralleling the crucial between-language comparisons for the test items without medial vowels. We ran a three-way mixed ANOVA where the dependent variable was the percentage of responses, and the independent variables were LANGUAGE (English, Mandarin) as a between-subjects factor; and STIMULUS ([atʰima], [atʰəma], [atɕʰma], [atɕʰəma]) and RESPONSE ([i], [other vowel], [no vowel]) as a within-subjects factors. Mauchly's test revealed there were violations of the assumption of sphericity for the two-way interaction of RESPONSE*STIMULUS and the three-way interaction of LANGUAGE*RESPONSE*STIMULUS; both effects were corrected with Greenhouse-Geisser correction for the degrees of freedom ($\epsilon = 0.57$). There was a main effect of Response [$F(2,74) = 43.75$, $p < 0.0001$]. There were two-way interactions for LANGUAGE*RESPONSE [$F(2,74) = 4.19$, $p = 0.02$] and RESPONSE*STIMULUS [$F(3.42,126.54) = 87.4$, $p < 0.0001$]. Most importantly, there was no three-way interaction effect; this suggests that the Mandarin speakers' responses did not differ substantially from the English speakers differentially depending on the stimuli. As with the crucial comparisons above, we followed up by conducting Mann Whitney U tests comparing the Mandarin and English speakers' responses to each type of stimulus, but only for the vowel responses (for reasons discussed further above in the context of the crucial stimuli comparisons). The only statistically significant test was for the [i] responses for the stimulus [atʰəma]; but, it was the English speakers who showed slightly higher responses [Mean diff (Mand.-Eng.) = -13.24, $W = 240.5$, $p < 0.05$]. The results of the analysis on the stimuli with medial vowels suggest that the Mandarin speaker's differing responses to the crucial alveolar and alveo-palatal stimuli without medial vowels (as observed in Figure 3) were not due to differing amounts of confusion with the relevant vowels in the same consonantal contexts.

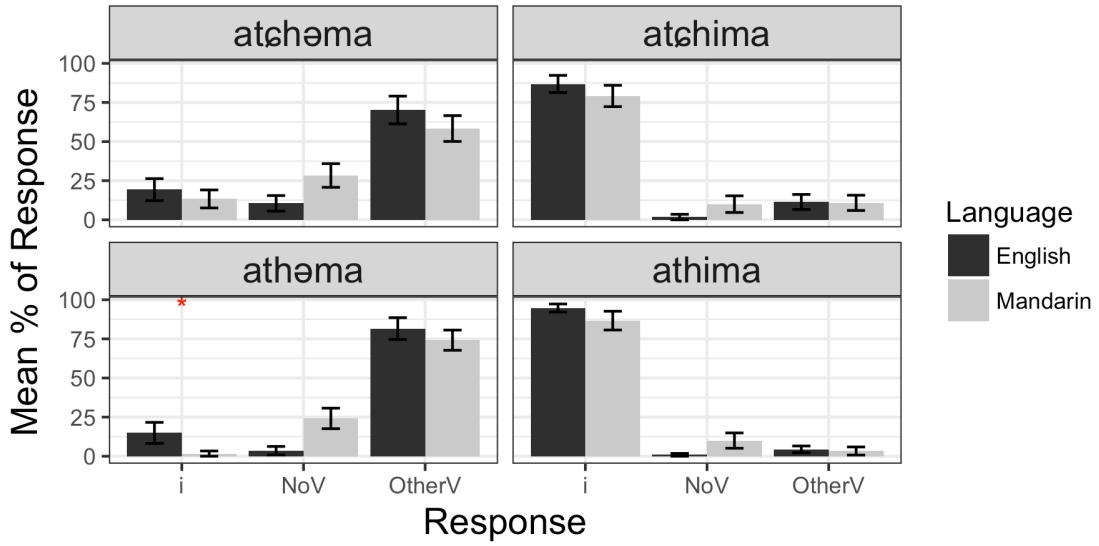

Figure 5: Mean % of responses of English and Mandarin speakers for the items with medial vowels in Exp. 2 (error bars = 1 S.E.; * = $p < 0.05$; ** = $p < 0.01$; asterisks represent p-values for between-language comparisons)

3.3. Discussion

While the results of Experiment 1 were in line with the predictions, the experiment itself was not a direct test of illusory vowels in the crucial test items, as ABX tasks are inherently comparative. This issue was addressed in Experiment 2, where an identification task was employed. The results of Experiment 2 suggest that: (a) [atɕ^{h}ma] triggered only /i/ illusions and no other vowel, (b) [at^{h}ma] did not trigger an /i/ illusions; instead the Mandarin participants heard another vowel, presumably /ə/.

The responses of the English participants suggests that the Mandarin speakers' performance was not due to acoustic artifacts in the recordings.

An interesting issue that arises from the results of both Exp. 1 and Exp. 2 is related to the somewhat small differences in accuracy rates between Mandarin and English speakers even for the crucial pairs (Exp. 1), and the somewhat low illusory vowel responses by Mandarin speakers in the crucial stimuli without medial vowels (Exp. 2).[19] For example, in Exp. 2, the [other vowel] illusion rate (which, note, is the measure of illusory /ə/) in the [at^{h}ma] context for the Mandarin speakers was about 36%. It is important to note that there is quite a large variation in illusory vowel responses in the previous literature; compare the above results to those of similar experiments in previous published work:[20] (a) There was a 65-70% identification of illusory /ɯ/ responses

by Japanese speakers (Dupoux et al. 1999), (b) there was a roughly 58% identification of /ɯ/ for Japanese speakers, and roughly 65% identification of /i/ by Brazilian Portuguese speakers (Dupoux et al. 2011), (c) there was a varying number of illusory /ɨ/ perception for Korean speakers based on the consonantal sequence involved (range: 25%-50%), (d) finally, Davidson and Shaw (2012), in a speeded AX task, report different levels of accuracy (proportions of hits) for English speakers for stimulus pairs (that differed in presence of absence of medial vowels) for different types of consonantal sequences (range: 0.51-0.88). While the differences in illusory vowel rates (and accuracy rates) between experiments appear quite sizeable, we don't yet have a complete understanding of what causes such differences. There are of course many possible reasons for the differences. First, as has been very well documented, illusory vowels are not the only way to "repair" incoming phonotactically illicit sequences; for example, there can be consonant deletions, consonant changes too (Davidson 2007; Davidson and Shaw 2012; Hallé et al. 1998). Note, such repairs are also generally consistent with the reverse inference view presented in this article: listeners might have inferred that the underlying consonants are different (but, still those that can account for a lot of the acoustic properties of the incoming stimuli), leading to what might be termed a consonant change in the percept; or listeners might have inferred that some of the acoustic characteristics perceived might be "noise" of some sort, and therefore didn't infer any underlying consonantal counterpart, leading to what might be termed consonant deletion in the percept. Given that there are other perceptual repairs, it is quite unclear generally how and why participants infer different possibilities. Therefore, it is difficult to predict the exact proportion of illusory vowels listeners' of a language should hear in relevant contexts.

Second, even if the above were disregarded, though there is some understanding of how aspects of the acoustic input proximate to the illicit consonant cluster have an effect on illusory vowels (Davidson and Shaw 2012; Wilson and Davidson 2013, amongst others), there are still many other aspects that researchers, to our knowledge, have not yet studied systematically. For example, what is the influence of the durational (and more generally, phonetic properties) of the other vowels in the stimuli and do illusory vowel percepts increase/decrease depending on whether the surrounding vowels are shorter/longer? There is some reason to believe this should happen based on speech-rate effects on the parsing of ambiguous stimuli (Dilley et al. 2013; Dilley and Pitt 2010; Heffner et al. 2013). As Dilley, Pitt and colleagues have consistently found, whole syllables can "vanish" in ambiguous stimuli due to the distal/proximal speech rates. As a very reasonable extension, one can also conjecture that there should be similar speech rate effects on

illusory vowel effects, which would modulate the rates of perception of illusory vowels. Another reasonable source of differences in such experiments is the acoustic properties of the stimuli *with* the medial vowels. Conducting an identification experiment, where such stimuli with medial vowels have longer (or shorter) vowels should affect the inference of a [ø] in the acoustic input as a vowel.[21]

Third, many of the previous results on illusory vowels are based on non-tonal languages. But, in Mandarin, there is a general relationship between tone and syllable count, and the stimuli with no medial vowels only had two tones. So it is possible that the presence of just two tones that already corresponded to other acoustically present vowels/syllables in the stimuli negatively affected the Mandarin speakers perception of an illusory vowel in the medial position, thereby increasing their [no vowel] responses.

Finally, one can never discount that listeners can also use the general auditory perception system (not just the speech perception system) in these tasks. This is particularly true given that the stimuli in the experiment, though reasonably natural, still might have sounded non-native to the Mandarin speakers. Therefore, it is possible that the Mandarin speakers were influenced by general auditory processing and therefore more pre-disposed (than if only the speech-perception mechanism had been in play) to recognising that there were no vowel-like elements in the medial position for the stimuli without medial vowels.

It is for the above reasons, we have repeated throughout that the clearest test of explanations of illusory vowels, based on a particular view of speech perception that we advocate here, is in the direction of such effects in between-language comparisons; but expectations of the magnitude of the effect are still somewhat premature. However, we raise the above possibilities as both interesting and necessary avenues for further research.

Before concluding this discussion, it is important to mention that an aspect of the illusory vowel phenomenon that Experiment 2 could not clearly isolate was the exact quality of the illusory vowel for Mandarin speakers in [at^{h}ma]. As mentioned in Section 3.1.3, one of the possible responses given to the participants was [other vowel] because both English orthography and Pinyin, which is the official romanization for Standard Chinese, use the same letter “e” appears in graphs for [ə], [e] and [ɤ] (note: in Pinyin, the “e” in the orthographic sequence “ei” represents [e], but the letter “e” by itself can represent [ə] before a nasal coda or [ɤ] in an open syllable). Therefore, giving the participants the orthographic option of “e” would not have been completely sufficient to identify the exact quality of the vowel. Despite the orthographic confound, the experiment, by

itself, does allow us to eliminate some potential counter-hypotheses to account for the patterns. It is not possible to maintain that it is simple phonotactics that determine the vowel. As mentioned in the *Introduction*, alveolar stops are possible before *all* vowels in Mandarin except [*t^{h}e], which means [t^{h}i] is a phonotactically possible sequence. Since the [i] response was at floor levels, it is not possible to say that the illegal sequence present in [at^{h}ma] is fixed by just any phonotactic sequence allowed in the language. The consistent selection of the [other vowel] response by the Mandarin speakers suggests that only one (or a few phonotactically legal vowels) are possible illusory vowels in that context. Again, this cannot be purely due to the acoustic factors present in the stimuli either, since that would predict the English speakers should also have a similar rate of [other vowel] responses. Instead, the English speakers largely chose [no vowel]. If the Mandarin speakers' responses for [other vowel] were purely due to the presence of acoustic factors, then it would remain unexplained why the English speakers did not use the same cues to respond in a similar fashion.

For a similar set of reasons, it is not possible to maintain a straightforward frequency based account for the illusory vowels, particularly those that are adjacent to the alveolar context. In Tables 4-5, we present the bigram and trigram frequencies for [t^{h}V] or [t^{h}Vm] (where V = any vowel represented in Pinyin) that were obtained from the Lancaster Corpus of Mandarin (McEnery and Xiao 2004). The frequencies are for Pinyin orthography, and not directly for phonological segments in the surface representations. Therefore, the earlier caveat of each letter potentially representing more than one segment remains here. Having said that, Pinyin orthography is a much more regular/consistent orthographic system, than say English, therefore the frequencies are still instructive.

The tables contain two sets of bigram and trigram counts/percentages: by *Token*, we mean the total number of such occurences in the corpus, counting those words that are repeated multiple times as separate instances; and by *Type*, we mean the number of occurences where each unique word/Pinyin sequence is considered only once. As can be seen, both *Type* and *Token* counts suggest that [t^{h}i] and [t^{h}im] are in the top two/three most frequent bigram and trigram sequences, respectively, in Mandarin (the bigram *Type* frequency of the [t^{h}i] is the second highest). Despite the clear preponderance of [t^{h}i] and [t^{h}im], Mandarin speakers were unwilling to choose [i] in the [at^{h}ma] items. This suggests that the illusory vowels that the Mandarin speakers were perceiving were not primarily driven by such frequency information.

As an additional note for those interested in further pursuing a frequency-based account, we

Table 4: Bigram frequencies in Mandarin (collapsing across all tones)

	Pinyin	ta	tai	tao	te	tei	ti	tie	tou	tu
	IPA	[t^ha]	[t^hai]	[t^hau]	[t^hei] or [t^hə]	[t^hei]	[t^hi]	[t^hie]	[t^ho]	[t^hu]
Token	Count	12586	2415	865	1478	0	6088	544	2892	2259
	Percentage	43.21	8.29	2.97	5.07	0	20.90	1.87	9.93	7.76
Type	Count	92	237	135	126	0	312	91	396	286
	Percentage	5.49	14.15	8.06	7.52	0	18.62	5.43	23.64	17.07

Table 5: Trigram frequencies in Mandarin (collapsing across all tones)

	Pinyin	tam	taim	taom	tem	teim	tim	tiem	toum	tum
	IPA	[t^ham]	[t^haim]	[t^haum]	[t^həm]	[t^heim]	[t^him]	[t^hiem]	[t^hom]	[t^hum]
Token	Count	1948	3	2	2	0	43	6	46	25
	Percentage	93.88	0.14	0.09	0.09	0	2.07	0.29	2.22	1.20
Type	Count	5	3	2	2	0	5	1	10	8
	Percentage	13.89	8.33	5.56	5.56	0	13.89	2.78	27.78	22.22

would also like to mention that even if another frequency measure were found to correlate with the identification rates in these experiments, it would only be meaningful if there was a clear (theoretical) explanation of why that particular frequency measure is the appropriate measure, and not bigram and trigram frequencies. In short, what is needed to accept a frequency-based explanation is a clear theoretical viewpoint that clearly explains why some types of frequencies, and not others, are relevant to the process.

4. Conclusion

In this paper, we set out to corroborate a particular view that states that the task of the perceiver during speech perception is to identify the best estimate of the intended underlying representations of the utterance given the acoustic token. A 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. 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.*, a language where word-final /t/

becomes a glottal stop - here, inference of the correct intended underlying representation *has* to use the phonology of the language.

The above view makes testable predictions about the quality of the illusory vowels that a native speaker of a particular language might perceive in different illicit phonotactic contexts. Based on Mandarin phonological and phonetic patterns, the view predicts that, compared to native English speakers, native Mandarin speakers should perceive more illusory vowel /i/ in illicit consonantal sequences where the first consonant is alveo-palatal, but more illusory /ə/ in illicit consonantal sequences where the first consonant is alveolar. The results of Experiment 1 & 2 were in line with these expectations. We further argued that the illusory vowels perceived cannot simply be due to phonotactics or (bigram/trigram) frequencies of segments.

Finally, we would like to note that, while the research is in line with previous work that shows a general modulation of speech perception by phonological knowledge (Boomershine et al. 2008; Huang 2001; Hume and Johnson 2003; Johnson and Babel 2010), the article presents results that suggest a particular viewpoint on how exactly that modulation takes place that can naturally account for such phonological sensitivity in speech perception.

4.1. Implications for loanword phonology

As pointed out in the *Introduction*, while loanword patterns cannot be assumed to be purely based on perceptual processes (Jong and Cho 2012; Kang 2011; LaCharité and Paradis 2000; Peperkamp et al. 2008; Smith 2006; Vendelin and Peperkamp 2006), a better understanding of speech perception lends itself naturally to a better understanding of loanword patterns.

The current viewpoint naturally accounts for at least some of the loanword patterns observed in Mandarin Chinese. A few typical examples of loanword borrowings in Mandarin Chinese are presented in (2), where the relevant vowel is boldfaced and underlined (note: lexical tone is not marked in the transcriptions). For the words in (2), the source language is English, and the original words contain alveolar/palato-alveolar consonants in coda positions. As can be seen, when the original sourceword contains an alveolar consonant in a coda position, it is borrowed with an alveolar consonant followed by a [ɤ] vowel (22a); and when the original sourceword contains a palato-alveolar consonant in a coda position, it is largely borrowed with an alveo-palatal consonant followed by an [i] vowel (22b).

2. Typical loanword patterns in Mandarin Chinese (Some of the data is from Lin (2007))

(a) Alveolar contexts

- [ja:t^{h}ɤ:lanta:] ‘Atlanta’
- [atɤ:lɤ:] ‘Adler’
- [ʂɻk^{h}ɑut^{h}ɤ:] ‘Scott’

(b) Alveo-palatal contexts

- [litɕ^{h}i:məŋ] ‘Richmond’
- [k^{h}ai:tɕi:] ‘Cage’
- [fi:tɕ^{h}i:] ‘Fitch’

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 the epenthetic vowels in loanwords containing illicit coda consonants show very specific patterns (note: by “epenthetic”, we do not imply an analysis that the vowel is not present in the underlying representation and surfaces in the surface representation. Instead, we just mean it descriptively to indicate it is not present in the sourceword). In alveolar contexts [t, t^{h}], the epenthetic vowel is [ɤ] in 98% of the cases; and in alveo-palatal contexts, the vowel [i] is the epenthetic vowel in 83% of the cases. Given that the [ɤ] is an allophone of /ə/ in Mandarin Chinese (Lin 2007), both these patterns are consistent with the perceptual claims presented in this paper.

As can be observed, the results have an immediate bearing on the theoretical literature of loanword adaptations, where there is a debate on the factors affecting loanword adaptations (Davidson 2007; Kang 2011; Peperkamp 2005; Smith 2006, *inter alia*). Whereas some claim that perceptual factors are perhaps the primary factor influencing loanword adaptation patterns (Peperkamp 2005; Peperkamp and Dupoux 2003), others claim that phonological factors play a big role, and perception is at best a minor factor in such patterns (Jacobs and Gussenhoven 2000; LaCharité and Paradis 2005; Paradis and LaCharité 1997; Uffman 2006). The proposed account in the current article suggests that the dichotomy between phonology and speech perception is perhaps a false one, given that phonological knowledge is argued to be crucially recruited during speech perception. The question, as we see it, is not whether a particular loanword pattern is due to phonology or speech perception, but instead *how* phonological knowledge is used during speech perception to infer the relevant underlying representations for the incoming acoustic input from the loanword source language. This is of course not to deny the role of non-auditory sources of loanwords.

A second issue that is brought to the fore is due to the claim that the illusory vowel is related, not to any vowel insertion pattern in the native phonology, but to vowel deletion patterns that exist in the native language. This is so, because as per the current viewpoint a native phonological process such as ($/V_1/ \rightarrow [\emptyset]$) can be used to infer a vowel in the underlying representation when no vowel is present in the acoustic signal, but a native phonological process such as ($/\emptyset/ \rightarrow [V_1]$) is not helpful in identifying the underlying representation when no vowel is present in the acoustic signal. This prediction does in fact bear out in many of the languages probed for illusory vowels - in Japanese, the illusory vowel in neutral contexts is /ɯ/[22] (Dupoux et al. 2011); in Korean, in similar contexts, it is /ɨ/ (Durvasula and Kahng 2015); and as seen above, in Mandarin Chinese, the illusory vowel in neutral contexts is /ə/. All of these vowels are involved in native vowel deletion processes in the respective languages, but not necessarily in any native vowel insertion processes. These are all also vowels that are the typical (or default) epenthetic vowels in the loanwords in the respective languages. Therefore, if a loanword pattern can be traced back to perceptual factors, then it is likely that the epenthetic vowel found in such loanwords is actually a vowel that is deleted, and not inserted, in the native phonology.

Acknowledgements

This article was made possible due to the help and support of many individuals. First and foremost, we would like to thank the Associate Editor and and the reviewers for valuable criticism that helped make this article much better. Second, we would like to thank Phil Monahan, Alan Beretta, and the members of the Phonology-Phonetics group at Michigan State University for many helpful comments. Fourth, we would like to thank Suzanne Wagner and Mike Kramizeh for help with experimental equipment and lab-space. Finally, we would like to thank the audiences of LabPhon 2016 for probing questions, and helpful discussion.

Notes

[1] *cf.* Yun 2016.

[2] This vowel is sometimes transcribed as [u] for Japanese. But given its usual description as a high, back unrounded vowel, it is more appropriately transcribed as [ɯ].

[3] We have simplified the description of the stimulus creation a bit here in the interest of presentational clarity. However, we have retained the essential aspects.

[4] We notate phonemic representations with /…/ and acoustic output/input with […].

[5] This can also be extended to allophonic mappings before particular vowels.

[6]Cheng (1973) mentions another optional fast-speech /ə/ deletion process targeting vowels between an /i/ or /u/ and an [ŋ], the /ə/. However, he also mentions that this process is conditioned by "performance factors" and is not accepted by everyone.

[7]Note, we are talking about the phoneme /ə/ here as the illusory vowel, not the surface allophone. Furthermore, as mentioned later in the article, the phoneme has two phonetically close allophones, namely, [ə] and [ɤ]; the former surfaces before a coda consonant [n/ŋ], and the latter in open CV syllables.

[8]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}].

[9]Thanks to a reviewer for suggesting this to us.

[10]The Mean and the SD were driven up largely by two participants who were in the US for 5-6 years.

[11]Thanks to the Associate Editor and a reviewer for raising these issues.

[12]We use '$\sim$' to notate a pair irrespective of order, and '-' to notate a particular ordered sequence.

[13]We excluded the comparison pairs involving the two vowels based on the recommendation of a reviewer. Originally, we ran a two-way mixed ANOVA, where LANGUAGE and COMPARISONS (all 6 comparison pairs) were two separate factors. The analysis crucially revealed an interaction of LANGUAGE*COMPARISON [$F(5,185) = 15.36$, $p < 0.0001$], which again suggests that the Mandarin and English speakers had different responses to different comparison pairs.

[14]Thanks to a reviewer for making this observation in our results.

[15]Thanks to a reviewer for raising this issue.

[16]Thanks to a reviewer for suggesting this to us.

[17]Thanks to a reviewer for suggesting this.

[18]Thanks to a reviewer who requested we include the analysis in the article.

[19]Thanks to the Associate Editor and a reviewer for raising this issue.

[20]Note, the list is representative, not comprehensive.

[21]In fact, we are currently pursuing these lines of inquiry in current experiments.

[22]While there is some debate about whether the vowel is really deleted, or just devoiced, recent research suggests that at least in some environments, there can be complete deletion *i.e.*, there does not seem to be any "reduction". (Shaw and Kawahara 2018; Tsuchida 1997; Varden 1998)

References

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

Berent, Iris, T Lennertz, Paul Smolensky, and V Vaknin-Nusbaum (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology* 26, pp. 75–108.

Berent, Iris, Donca Steriade, T Lennertz, and V 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.

Boersma, Paul and Silke Hamann (2009). Loanword adaptation as first-language phonological perception. *Loan Phonology*. Ed. by Andrea Calabrese and W. Leo Wetzels. Amsterdam/Philadelphia: John Benjamins, pp. 11–58. DOI: 10.1075/cilt.307.02boe.

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/.

Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume, and Keith Johnson (2008). The impact of allophony versus contrast on speech perception. *Contrast in phonology: theory, perception, acquisition*. Ed. by Peter Avery, B. Elan Dresher, and Keren Rice. Berlin & New York: Mouton de Gruyter, pp. 145–171.

Burzio, Luigi (2007). Phonology and phonetics of English stress and vowel reduction. *Language Sciences* 29.2–3. Issues in English phonology, pp. 154–176. ISSN: 0388-0001. DOI: http://dx.doi.org/10.1016/j.langsci.2006.12.019.

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

Cheng, Chin-chuan (1973). A Synchronic Phonology of Mandarin Chinese. Monographs on Linguistic Analysis, No. 4. The Hague: Mouton.

Chomsky, Noam and Morris Halle (1968). The Sound Pattern of English. New York, Evanston, and London: Harper and Row.

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.

Dilley, L. C., T. H. Morrill, and E. Banzina (2013). New tests of the distal speech rate effect: examining cross-linguistic generalization. *Frontiers in Psychology* 4.1002. DOI: 10.3389/fpsyg.2013.01002.

Dilley, Laura C. and Mark A. Pitt (2010). Altering Context Speech Rate Can Cause Words to Appear or Disappear. *Psychological Science* 21.11. PMID: 20876883, pp. 1664–1670. DOI: 10.1177/0956797610384743. eprint: http://dx.doi.org/10.1177/0956797610384743.

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, E Parlato, S Frota, Y Hirose, and Sharon Peperkamp (2011). Where do illusory vowels come from? *Journal of Memory and Language* 64.3, pp. 199–210.

Durvasula, Karthik and Jimin Kahng (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, N H and T 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. G. and W. 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. G. and W. D. Marslen-Wilson (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 24.2, pp. 380–396.

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

Guevara-Rukoz, Adriana, Isabelle Lin, Masahiro Morii, Yasuyo Minagawa, Emmanuel Dupoux, and Sharon Peperkamp (2017). Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis. *The Journal of the Acoustical Society of America* 142.2, EL211–EL217. DOI: 10.1121/1.4998138. eprint: https://doi.org/10.1121/1.4998138.

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.

Heffner, Christopher C., Laura C. Dilley, J. Devin McAuley, and Mark A. Pitt (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes* 28.9, pp. 1275–1302. DOI: 10.1080/01690965.2012.672229. eprint: http://dx.doi.org/10.1080/01690965.2012.672229.

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.

Hooper, Joan B. (1978). Recent Developments in Historical Linguisitics. *Constraints on schwa-deletion in American English*. Ed. by Jacek Fisiak. The Hague: Mouton, pp. 183–207.

Huang, Tsan (2001). The interplay of perception and phonology in Tone 3 sandhi in Chinese Putonghua. *OSU Working Papers in Linguistics* 55, pp. 23–42.

Hume, Elizabeth and Keith Johnson (2003). The impact of partial phonological contrast on speech perception. *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 2385–2388.

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

Jacobs, Haike and Carlos Gussenhoven (2000). Loan phonology: perception, salience, the lexicon and OT. *Optimality Theory: phonology, syntax, and acquisition*. Ed. by J. Dekkers and J. M. van de Weijer F. R. H. van der Leeuw. Oxford: Oxford University Press, pp. 193–210.

Johnson, Keith and Molly Babel (2010). On the perceptual basis of distinctive features: evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics* 38, pp. 127–136.

Jong, Kenneth de and Mi-Hui Cho (2012). Loanword phonology and perceptual mapping: Comparing two corpora of Korean contact with English. *Language* 88, pp. 341–368.

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.

Kreidler, Charles (1989). The Pronunciation of English. Oxford: Blackwell.

LaCharité, Darlene and Carole Paradis (2000). Phonological evidence for the bilingualsim of borrowers. *Proceedings of the 2000 Annual Conference of the Canadian Linguistic Association, Ottawa*, pp. 221–232.

LaCharité, Darlene and Carole Paradis (2005). Category preservation and proximity versus phonetic approximation in loanword adaptation. *Linguistic Inquiry* 36.2, pp. 223–258.

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.

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

Mattingley, Wakayo, Elizabeth Hume, and Kathleen Currie Hall (2015). The influence of preceding consonant on perceptual epenthesis in Japanese. Paper presented at the 18th International Congress of Phonetic, Sciences (ICPhS 2015).

McEnery, Anthony and Zhonghua Xiao (2004). The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*. Lisbon, pp. 1175–1178.

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

Mitterer, H., S. Kim, and T. 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.

Monahan, Philip J., Eri Takahashi, Chizuru Nakao, and William Idsardi (2009). Not All Epenthetic Contexts are Equal: Differential Effects in Japanese Illusory. *Proceedings of the 17th Annual Japanese/Korean Linguistics Conference*. Ed. by I. Shoichi, H. Hoji, P. M. Clancy, and Sohn S.-O., pp. 391–405.

Paradis, Carole and Darlene LaCharité (1997). Preservation and minimality in loanword adaptation. *Journal of Linguistics* 33.2, pp. 379–430.

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

Peperkamp, Sharon and Emmanuel Dupoux (2003). Reinterpreting loanword adaptations: the role of perception. *Proceedings of the 15th International Congress of Phonetic, Sciences (ICPhS 2015)*, pp. 367–370.

Peperkamp, Sharon, Inga Vendelin, and Kimihiro Nakamura (2008). On the perceptual origin of loanword adaptations: Experimental evidence from Japanese. *Phonology* 25, pp. 129–164.

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.

Shaw, Jason A. and Shigeto Kawahara (2018). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics* 66.Supplement C, pp. 100–119. ISSN: 0095-4470. DOI: https://doi.org/10.1016/j.wocn.2017.09.007.

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.

Tsuchida, A. (1997). Phonetics and phonology of Japanese vowel devoicing. Ph.D. Dissertation. Ithaca, NY, USA: University of Cornell.

Uffman, Christian (2006). Epenthetic vowel quality in loanwords: empirical and formal issues. *Lingua* 116, pp. 1079–1111.

Varden, J. K. (1998). On high vowel devoicing in standard modern Japanese: implications for current phonological theory. Ph.D. Dissertation. Seattle, WA, USA: University of Washington.

Vaux, Bert (2002). *Aspiration in English*. Ms. retrieved from http://people.ds.cam.ac.uk/bv230/li8/aspiration-uwm.pdf.

Vendelin, Inga and Sharon Peperkamp (2006). The influence of orthography on loanword adaptations. *Lingua* 116, pp. 996–1007.

Weinberger, Steven H. (1996). Minimal segments in second language phonology. *Second-language speech: structure and process*. Ed. by James A & Leather J. Berlin: Mouton de Gruyter, pp. 263–311.

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.

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

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

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.