How Much do Visual Cues Help Listeners in Perceiving Accented Speech?

Yi Zheng [1] and Arthur G. Samuel [1,2,3]

1. Department of Psychology, Stony Brook University
2. Basque Center on Cognition, Brain, and Language
3. Ikerbasque, Basque Foundation for Science

Contact Information:

Yi Zheng

Dept of Psychology

Stony Brook University

Stony Brook, NY 11794-2500

Email: yizheng.psychology@gmail.com

Abstract

It has been documented that lip reading facilitates the understanding of difficult speech, such as noisy speech and time-compressed speech. However, relatively little work has addressed the role of visual information in perceiving accented speech, another type of difficult speech. In this study, we had native English speakers make lexical decision judgments on English words or non-words produced by speakers with Chinese accents. The stimuli were presented as videos that were either of a relatively far speaker, or as videos in which we zoomed in on the speaker's head. Consistent with studies of degraded speech, listeners were more accurate at recognizing accented speech when they saw lip-movements from the closer apparent distance. The effect of apparent distance tended to be larger under non-optimal conditions: when stimuli were non-words than words, and when stimuli were produced by a speaker who had a relatively strong accent. Perhaps surprisingly, we did not find any influence of listeners' prior experience with Chinese accented speech, suggesting that cross-talker generalization is limited. The current study provides practical suggestions for effective communication between native and non-native speakers: Having access to visual information is useful, and is more useful in some circumstances than others.

Keywords: visual cues; lip-reading; distance; accented speech; recognition

How Much do Visual Cues Help Listeners in Perceiving Accented Speech?

Communication between native and non-native speakers is widespread, particularly in countries like the United States which have significant numbers of immigrants. This imposes challenges for both speakers and listeners in various contexts, including college classrooms. In a CGS/GRE survey, there were over 800,000 international graduate students entering U.S. colleges for studies in 2012. (http://www.cgsnet.org/ckfinder/userfiles/files/GEDReport_2012.pdf). At Stony Brook University, for example, 60% of the new graduate students during the 2013/2014 academic year were international students, and over half of these were Chinese native speakers. These international students are often assigned to be instructors or teaching assistants in U.S. colleges. However, American undergraduate students frequently complain about not being able to understand their speech well (Borjas, 2000; Finder, 2005; Fitch & Morgan, 2003; Zhou, 2014). Studies suggest that these non-native English speakers are usually perceived as poorer communicators than native speakers (Grossman, 2011; Hosoda, Stone-Romero, & Walter, 2007).

To enhance successful communication between international TAs (ITAs) and American undergraduate students, there are two logical alternatives: improve the speakers' accent/ pronunciation in order for them to be better understood, or improve the listeners' understanding without much change on the speakers' side. In the latter case, for example, one could focus on the use of visual cues during accented speech perception. If lip movements can substantially enhance the understanding of accented speech, they have the potential to provide an important way to enhance communication between accented speakers and native listeners (Banks, Gowen, Munro, & Adank, 2015a; Janse & Adank, 2012). Specifically, one potential solution could be to encourage students to sit close to the speakers if they are having trouble understanding the

speech. Although this seems intuitive, there is actually no direct evidence to suggest that this would improve listeners' ability to understand accented speech.

Many prior studies have examined the role of visual information in processing other types of difficult speech. It has been known since Sumby and Pollack (1954) that access to lip-reading information improves perception of speech in noise, and the benefits of viewing articulatory movements in a noisy environment have been confirmed in later studies (e.g., Erber, 1969, 1971; MacLeod & Summerfield, 1987; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Moreover, visual information has been shown to improve perception of noise-vocoded speech, which has been studied due to its similarity to the signal produced by a cochlear implant. Several of these studies reported that lip-movement information can enhance perceptual learning of noise-vocoded speech (Bernstein, Auer Jr, Jiang, & Eberhardt, 2013; Kawase et al., 2009; Pilling & Thomas, 2011; Wayne & Johnsrude, 2012). Lip-reading has also been shown to facilitate the comprehension of another type of challenging speech -- time-compressed speech (Adank & Devlin, 2010; Banai & Lavner, 2012). Collectively, there is thus substantial evidence that visual speech cues can help listeners to understand sub-optimal speech input. However, relatively little work has examined the possible use of visual cues in non-native speech perception. We will review the relevant work in the following sections.

Lip-reading and non-native phonemes

Hazan, Kim, and Chen (2010) had participants identify /ba/, /da/ and /ga/ produced by native or non-native speakers in audio-visual (AV), audio-only (AO), and visual-only (VO) conditions, and found that listeners gave greater weight to visual information when listening to non-native speech than when listening to native speech. Wang, Behne, and Jiang (2008) presented native Mandarin speakers with syllables that contained English fricatives in AV, AO,

and VO conditions, and found that visual cues facilitated non-native fricative identification. Moreover, Chinese participants who had resided for a short time in Canada showed more reliance on visual information than those who had been there longer, indicating an effect of linguistic experience on non-native phoneme identification. These results converge with the finding that the visual contribution to non-native fricative identification is modulated by the listener's first language (Wang, Behne, & Jiang, 2009). Hazan et al. (2006) asked English learners to identify visually salient contrasts (e.g., labial/labiodental consonant contrasts) and visually less salient contrasts (e.g., /l/-/ɹ/). They found that both Japanese and Spanish learners of English performed better in audiovisual than in audio-only conditions on visually salient contrasts, but neither showed an audiovisual benefit for /l/-/ɹ/ contrasts. Thus, the visual salience of non-native sounds played an important role in the use of visual information.

In addition to these studies that focused on the use of visual cues for consonants, Navarra and Soto-Faraco (2007) demonstrated that visual cues facilitated the recognition of vowel contrasts in non-native speech. In their study, Spanish-dominant bilinguals who spoke Catalan as a second language failed to distinguish the Catalan sounds /ɛ/ and /e/ in an audio-only condition, but could successfully do so using additional visual information. However, Kawase, Hannah, and Wang (2014) suggested that visual cues are not always helpful in non-native phoneme recognition. They presented native English listeners with three English phonemic consonants produced by Japanese native speakers, and found that the presence of visual information could positively or negatively affect the recognition of phonemes. For instance, an inaccurate articulation configuration of /ɹ/ by Japanese speakers provided native listeners with misleading information in the identification task, lowering recognition. Thus, although the literature shows

that in general visual information helps to understand non-native phonemes, sometimes misleading information from visual cues can be detrimental.

Lip-reading and accented speech

The relationship between lip-reading and accented speech has not been extensively studied yet. Yi, Phelps, Smiljanic, and Chandrasekaran (2013) asked participants to transcribe sentences in native- or Korean-accented speech presented in noise, in an audio-only or an audio-visual condition. They found that lip-movements facilitated speech recognition, but the visual enhancement was greater for native speech than for the Korean-accented speech. In addition, Korean speakers were rated as more accented in the AV than in the AO condition, whereas native speakers were rated as producing less accented speech in the AV than in the AO conditions (see Rubin, 1992, and Zheng & Samuel, 2017, for related findings).

Other studies have reported a positive role for visual cues, though in general the effects have been modest. Barros (2010) found that access to visual cues enhanced the intelligibility of Brazilian-accented English slightly and non-significantly for native English listeners. Banks and colleagues (2015a) presented Japanese-accented speech in noise to native English listeners, and found that recognition accuracy was significantly better in an audiovisual condition than in an audio-only condition. However, they found that visual information did not facilitate the perceptual learning of accented speech: Participants improved at the same rate when presented with accented speech in AO versus AV conditions. To reconcile this finding with prior studies showing that visual cues enhanced the perceptual learning of noise-vocoded speech, Banks et al. (2015a) suggested that the results depend on the characteristics of the speech signal: "variation in noise-vocoded speech stems from degrading the acoustical composition of the entire speech signal, whereas accented speech varies in terms of its phonemic patterns, is acoustically intact

and only affects certain speech sounds" (p.2). In a related study, Janse and Adank (2012) showed that native Dutch older adults showed marginally higher accuracy in understanding artificially accented Dutch in an AV condition than in an AO condition, with no difference in reaction times. They found that the initial adaptation to accented speech was faster in the AV condition compared to the AO condition, but the overall improvement ultimately was the same for the two conditions. Collectively, these studies suggest that visual cues can aid perception but not perceptual learning of accented speech (Adank & Janse, 2010; Banks, Gowen, Munro, & Adank, 2015a, 2015b).

The current study

As the preceding review indicates, lip reading can be helpful in processing difficult speech, such as speech-in-noise, vocoded speech, and time-compressed speech. Accented speech can also be considered to be a type of difficult speech, and one might therefore assume that visual cues should also help with accented speech. On the other hand, accented speech is difficult for different reasons: Accented speech is acoustically intact but has certain phonetic variations that are not present in native speech. Thus, the benefit of visual cues found for degraded speech may or may not be found for accented speech. Therefore, in the current study, we examine the role of visual cues on perceiving accented speech. Rather than testing this by completely eliminating these cues (the AO versus AV manipulation used in prior studies), we provide observers with different levels of visual cue availability and measure whether this affects how well they can recognize words.

We manipulate the quality of visual speech cues by varying the apparent distance between the speaker and the listener. The results of this quantitative variation in visual cue availability can be compared to the effect of the all-or-none type of manipulation (AV versus AO)

used in prior work. Our test was embedded in a real-world question: Does it make sense to advise American undergraduate students to sit closer to their non-native instructors in order to facilitate their understanding of their instructors' accented speech? We created a lab situation to simulate classroom conditions in which the speaker is seen from close up, where the mouth is clearly visible, versus conditions in which the speaker is further away so that the mouth information is less clear.

In constructing our study, we made two fundamental design decisions. First, we chose to focus on Chinese-accented English because a large number of international TAs at our university, as at many others, are from China (Davis, 1988; Rubin 1992). Second, we chose to test how well listeners could understand Chinese-accented words under conditions that did not allow listeners to use sentence level context to guess the words. This choice was grounded in our desire to know how much the visual information actually improves speech perception. In sentence-level tests it is difficult to separate how well listeners are actually decoding the words from how well they can use the sentence context to guess word identity. In fact, it has been known for over a half century (e.g., Miller & Isard, 1963) that accuracy of word report can be heavily affected by these additional cues. Therefore, we had our participants make lexical decision judgments about Chinese-accented English words and pseudowords. The required response on each trial is simply to indicate whether the item is a real English word or not. In the literature on word recognition, this is by far the most widely used task to measure word-level intelligibility. An additional virtue of this task, in the current study, is that the responses to the pseudowords can provide insights into how participants process unfamiliar items. Many courses in the STEM fields require students to deal with new terms, such as "arcsine", which are similar to the pseudowords included in the lexical decision test.

As noted above, the primary aim of the current work was to test the influence of distance. To operationalize the distance manipulation, we videotaped two native Chinese speakers producing accented English words and pseudowords, and presented two versions of the videos to our participants: "far" and "close." The far version items were recorded with the camera about four meters away from the speakers. For the close version, we took the original video-recordings and zoomed in on the speaker's head, cropping the rest of the original frame. This approach ensured that the far and close stimuli had exactly the same sound quality and lip movements (at the cost of some subtle cues that vary with actually approaching an object).

In sum, the current study aims to examine the role of visual quality in accented speech perception. In addition, to examine how the effect of visual quality might change over time, especially for listeners who are exposed to Chinese-accented instructors over the course of a semester, we asked participants who were initially tested at the beginning of a semester to return for re-testing at the end of the semester. All participants filled out a questionnaire that asked for information about their language background. The questionnaire included questions about the person's prior experience with non-native instructors.

Method

Participants

We recruited 152 Stony Brook undergraduate students within the first six weeks of the semester, all of whom had self-reported normal vision and hearing. Nine participants were excluded because they were not native English speakers (as reflected in the questionnaire), and one participant's data were not used due to headphone problems. Thus, usable data were obtained from 142 participants (114 females, 28 males). All were native English speakers and

were 18 years of age or older. The mean age was 20.4 (SD=2.76), with a range of 18 to 44 years old. None of the participants reported knowing either of the two speakers presented in the experiment. A subset of the participants (41 females, 16 males; mean age 21.3, SD=3.7, range of 19 to 44 years old) agreed to come back for a second session a few months later. The results for the second session will be discussed after those from the first session. Participants were compensated with $10 or partial course credit for each session of their participation. The study was approved by the Stony Brook University IRB.

Materials

60 English words were selected, ranging between one and four syllables in length. These words included common terms used in the STEM field (e.g. *axis*) and regular English words (e.g., *fight*). They were all relatively high-frequency words (frequency for each token is shown in Appendix A, retrieved from http://subtlexus.lexique.org/moteur2/index.php). We then made 60 non-words that matched the words in structure and number of syllables. Non-words were made by changing one consonant in a word, making it easy for our two speakers to know the desired pronunciation from the orthography.

Two native Chinese speakers recorded the stimuli during their first semester in the US. Both were first year Ph.D. students at Stony Brook University. One speaker had a relatively strong Chinese accent (female, 22 years old, chemistry student) and the other speaker had a weaker Chinese accent (male, 26 years old, economics student), reflected in their pronunciation scores on the Versant Test (https://www.versanttests.com/) of 49 and 59 (out of 80), respectively. The Versant test is an automated speaking test; its reliability and validity have been verified in the literature (Chun, 2008; Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008). The pronunciation section of the test focuses on speakers' ability to produce vowels, consonants,

and stress in a native-like manner. In the following text, we will refer to the first speaker as "female" and the second as "male", without any implication that these individuals provide any information about female versus male speakers in general.

The speakers were instructed to stand in front of a blackboard, and to read words and non-words from a laptop that was placed next to a VIXIA HFG20 Canon HD camcorder at a distance of about 4 meters. Thus, in the video it looked as if the speaker was looking directly into the camera. Speakers were asked to read the stimuli in a natural and clear way, with a neutral facial expression. The speakers were asked to re-record an item if it was not good in any way (e.g., disfluency, cough, frowns, smiling, obvious body movements, background noise, or not looking into the camera). Each word (e.g., *advertise*) was followed by a non-word made from that word (e.g., *adverbise*). This procedure made it easy for the speakers to produce both words and non-words. They were asked to produce all items with the same confidence level. During the videotaping process, the speakers wore a CVL lavalier microphone using a Shure BLX 14/CVL-H10 wireless system to ensure the quality of the audios. The microphone was placed close to the neck of each speaker.

We used VSDC video editing software to make two versions of each video. First, we split the video and audio and saved the audios as separate .wav files. Second, we applied a noise-reduction function in Goldwave editing software to minimize any background noise. Third, we normalized the amplitude of the audios, using Goldwave's half dynamic range option, to match the overall volume across the two speakers. Forth, we inserted the audio stream back into the videos. Finally, we cropped the borders of the original videos to make two versions: In one version, the distance from the camera to the speaker remained relatively far, whereas in the second version we zoomed in on the speaker's head, making the distance from the camera to the

speaker seem relatively close. As noted above, while this method does not capture every cue to distance, it provides a good approximation to the desired change in apparent distance while assuring that the exact same lip movements were present in the two versions of each video, with the same high-quality audio stream. The final versions were all 720 × 480 pixels per frame, with 44100 Hz frequency and 29.970 fps. Each stimulus was saved as an individual video, starting right before the speaker opened his/her mouth and ending when the mouth was closed. In total, there were 480 videos: 60 words and 60 non-words, each presented at two distances by two speakers. See Appendix B for examples of visual images of the two speakers at the two distances.

Procedure

We tested up to three participants at the same time. Participants first completed a questionnaire regarding their present and past experiences with Chinese speakers (e.g., professors, instructors, TAs, etc.) in classroom settings (see the questionnaire in Appendix C). For simplicity, we refer to these speakers as "Chinese TAs" in the following text. After completing the questionnaire, participants were tested in a sound-attenuated booth. They watched videos of speakers producing speech on a standard 17-inch Dell computer monitor (60 Hertz refresh rate, 32-bit color quality, 1280 by 1024 pixels resolution) about 60 cm from the participants. The audio was presented through high-quality SONY MDR-V900 headphones. The participants' task was to determine whether a given utterance was a word or a non-word. They were asked to do the task as accurately as they could without taking too much time, and to keep their eyes on the screen in front of them. Participants were monitored through a window, ensuring that they looked at the monitor throughout the whole experiment. For each video, participants had up to 3s to respond; there was a half second of silence between trials. They

registered each word versus non-word response by pressing one of two labeled buttons on a button board. Response accuracy was used as the measurement of speech intelligibility.

Each participant watched a complete set of 60 word videos and 60 non-word videos. The experiment was run using a custom-designed C++ program. Within each set of 60 stimuli, an equal number of stimuli (15) were presented for the four cases created by crossing the two distances with the two speakers. Four test versions were created by rotating distances and speakers across items, so that across participants each item was presented in all four combinations. The order of stimuli was pseudo-randomized for each set of 1-3 participants tested together. The whole study took around 20 minutes.

Results

We calculated the average accuracy for each level of Lexicality (word vs. non-word), Distance (close vs. far), and Speaker (male vs. female) for each participant. A four-way repeated measures ANOVA was conducted with three within-subject factors: Lexicality, Distance, and Speaker, and one between-subject factor: Experience. Assignment to an Experience level was based on a participant's responses on the questionnaire. Participants were divided into four groups based on their linguistic experience with Chinese TAs, either in the past, or currently (see Table 1). Although there was a very small (2%) trend toward better accuracy for those with more experience, there was no significant effect of Experience, F (1, 138) = 0.90, p = .444, $\eta 2$ = .02. Experience did not interact significantly with any of the other factors.

Table 1

Accuracy of Four Participant Groups at Time 1.

Experience Group	Current Chinese TA	Previous Chinese TA	% Correct
Low (n = 45)	No	No	*M* = 78
Mid 1(n = 45)	Yes	No	*M* = 78
Mid 2(n = 38)	No	Yes	*M* = 79
High (n = 14)	Yes	Yes	*M* = 80

Figure 1 summarizes the results of Lexicality, Distance, and Speaker for the 142 participants tested at the beginning of the semester. As is usually the case, participants' performance was significantly better on words than on non-words, F (1, 138) = 84.21, $p < .001$, $\eta 2 = .38$. The main effects of Speaker and Distance were also both significant, F (1, 138) = 14.33, $p < .001$, $\eta 2 = .09$; F (1, 138) = 10.27, $p = .002$, $\eta 2 = .07$, respectively. The main effect of Speaker is consistent with the higher Versant test score for the male speaker than for the female speaker.

The interaction between Speaker and Lexicality was significant, F (1, 138) = 7.08, $p = .009$, $\eta 2 = .05$. Pairwise comparisons showed that the male speaker (who had a relatively weaker accent) was significantly more intelligible than the female speaker for non-words (mean difference = 5.2%, $p < .001$) but not for words (mean difference = 0.6%, $p = .591$). The Distance × Lexicality interaction was not significant, F (1, 138) = 1.38, $p = .243$, $\eta 2 = .01$. The interaction of Distance and Speaker also was not significant, F (1, 138) = .49, $p = .485$, $\eta 2 = .004$. In both cases, there was a trend for Distance to have a slightly stronger effect when conditions were

more difficult, on non-words (2.6% difference) rather than words (1.1%), and on the more-accented female's speech (2.3%) than on the less-accented male stimuli (1.3%).

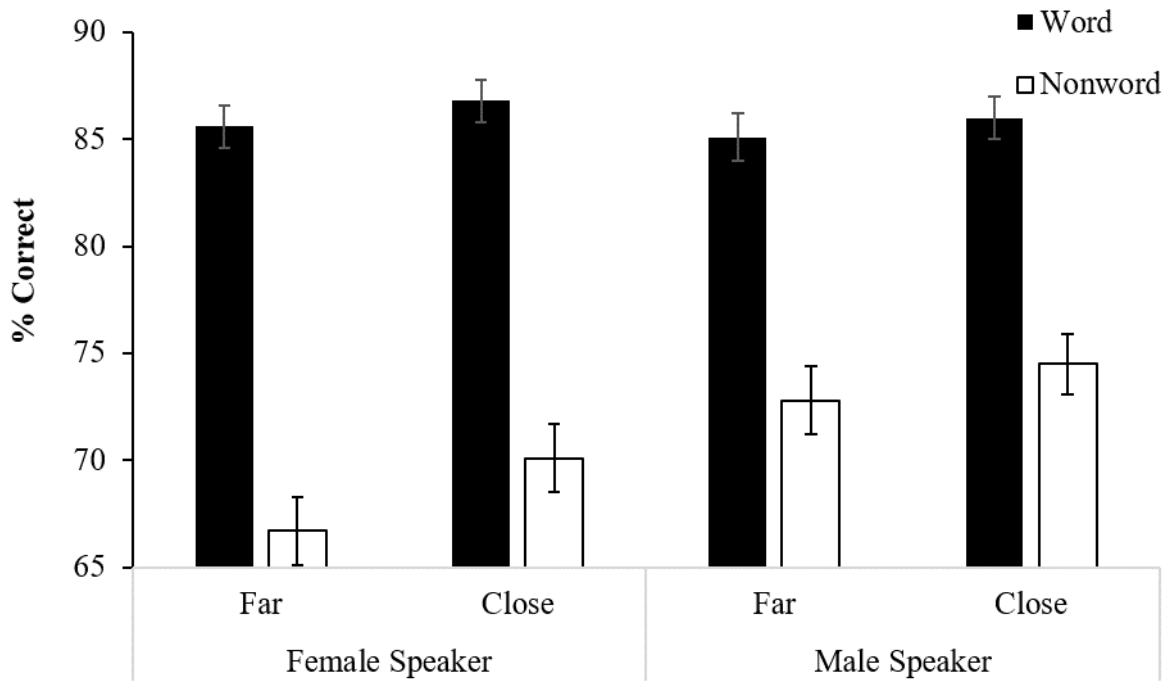


***Figure 1*. Accuracy as a function of Distance, Lexicality, and Speaker. Error bars represent the standard error of the mean.**

As Figure 1 shows, and as noted above, the effects of Distance and Speaker were primarily seen on the stimuli that were more difficult. This pattern can be summarized using Cohen's measure of effect size (see Table 2). Participants did significantly better at the close distance than at the far distance only when the stimuli were non-words and with the female speaker (who had relatively heavier accent). When words were pronounced by the female speaker, or when non-words were pronounced by the male speaker, Distance had a smaller and non-significant effect. The effect size of Distance was smallest when words were spoken by the

male speaker. Similarly, participants did significantly better at understanding the male speaker than the female speaker when the stimuli were non-words but not when the stimuli were words.

Table 2

Effect Sizes of Distance (Far vs. Close) as a function of Lexicality and Speaker.

Lexicality	Speaker	Mean Difference (Close –Far)	P value	Effect Size (Cohen's d)
Non-word	Female	3.4%	.038	.20
Non-word	Male	1.7%	.189	.12
Word	Female	1.2%	.289	.13
Word	Male	0.9%	.413	.05

Retest after Two Months: Of the 142 participants who were included in the first part of the study, 57 (50 females, 7 males) agreed to return to participate in the second part of the study. We compared the accuracy data (on the first session) for the 57 participants who returned and the 86 participants who did not return, and their performance did not differ, F (1, 140) = 1.15, p = .286, $\eta 2$ = .008. This suggests that the participants who agreed to return did not constitute a biased sample. The average time between the first part and the second part of the study was 60 (SD = 9.5) days. The participants were presented with the exact same set of videos at the two time points. We included a very long delay between the two tests (two months) to minimize any effect of experience with the stimuli, but it is of course possible that there could be some benefit from the first experience.

When participants returned for the second session, we had them complete the questionnaires again. Because a categorization based only on classroom experience may not fully reflect people's experiences with Chinese accented English (e.g., they may have Chinese friends or family), we added one question to the questionnaire that asked participants to rate their general familiarity with Chinese accented English on a scale of 1-10 (see Appendix C).

As before, we derived four levels of experience (Low, Mid1, Mid2, High), using the information provided in the questionnaires during the second session (13 of the 57 participants modified their answers about their TA experience on the second questionnaire). Table 3 shows that the four levels of experience (shown in the second and third columns) pattern in the same way as the familiarity ratings shown in the last column. Participants who had neither a previous nor a current Chinese TA reported low familiarity with the accent (a mean of 2.0 on a 10-point scale), while those with both previous and current Chinese TAs reported higher familiarity (5.5 out of 10). This convergence suggests that our categorization is in fact capturing the linguistic experience of the participants reasonably well.

Table 3 also shows the 57 participants' overall lexical decision accuracy at the two time points. We had expected that hearing a Chinese TA for two months in a classroom setting would help listeners to understand Chinese-accented English in the lab better. However, the results provide no support for this expectation. In fact, participants who did not have semester-long Chinese TAs (i.e., the Low and Mid2 groups, change of +3.5%) improved slightly more over time than those who had Chinese TAs (the Mid1 and Low groups, change of +1.4%).

Table 3

Accuracy and Familiarity Rating of Four Participant Groups at Time 1 and Time 2.

Experience Group	Current Chinese TA	Previous Chinese TA	% Correct (Time 1)	% Correct (Time 2)	Familiarity rating
Low (n = 8)	No	No	*M* = 76	*M* = 82	*M* = 2.0 (3.5)
Mid 1 (n = 22)	Yes	No	*M* = 79	*M* = 80	*M* = 4.5 (3.1)
Mid 2 (n = 14)	No	Yes	*M* = 80	*M* = 82	*M* = 5.2 (2.9)
High (n = 13)	Yes	Yes	*M* = 80	*M* = 82	*M* = 5.5. (2.7)

A five-way repeated measures ANOVA was conducted including four within-subject factors (Time, Speaker, Distance, Lexicality) and one between-subject factor (Experience). As in the larger sample of performance on the initial test, the main effect of Experience was not significant, F (3, 53) = .49, p = .692, $\eta 2$ = .027. Despite the absence of an effect of Chinese TA Experience, participants' overall performance significantly improved from Time 1 to Time 2, F (1, 53) = 23.13, $p < .001$, $\eta 2$ = .30, perhaps due to having taken part in the original test before.

Given this overall improvement, we can examine whether the differential effects of Distance across Speaker that we saw initially hold over time. Recall that for the full sample, performance was better for close videos than for far ones, and for the male speaker than for the female speaker, with these differences only being reliable for the more difficult non-word stimuli. The left panel of Figure 2 shows performance by the 57 participants as a function of Distance, Lexicality, and Speaker during the initial test session, while the right panel of Figure 2 provides the results when they returned two months later.

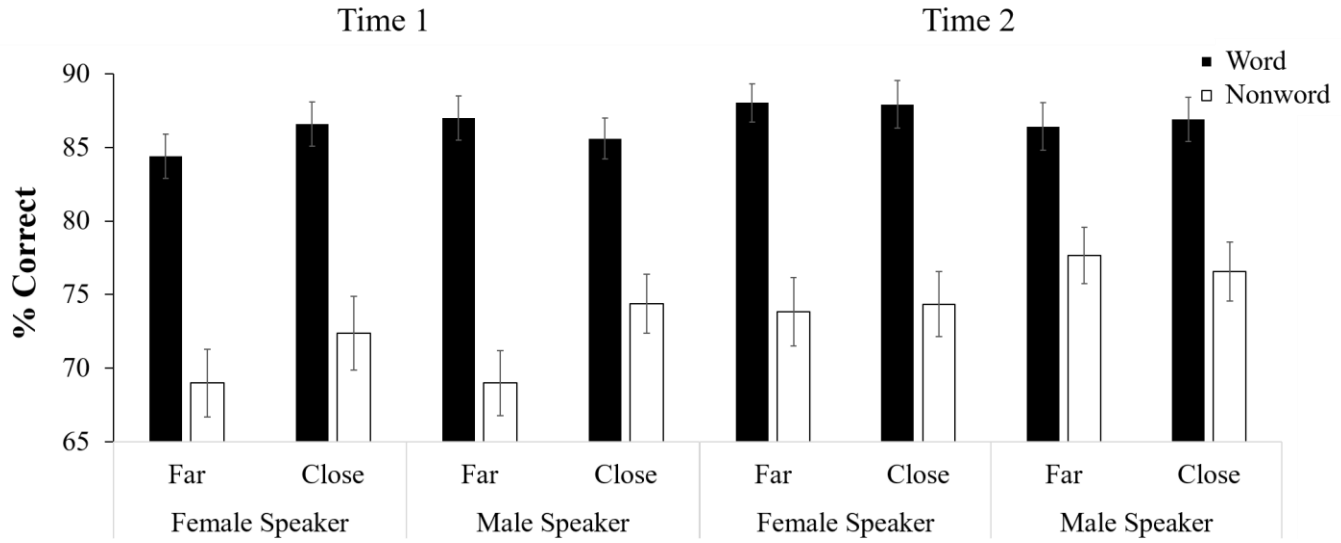


***Figure 2*. Accuracy as a function of Distance, Lexicality, and Speaker at Time 1 and Time 2. Error bars represent the standard error of the mean.**

Comparing the left panel of Figure 2 to Figure 1, we see that the 57 participants produced a pattern similar to that of the original larger sample (of which they were a subset). The effect of Distance was quite similar to that the full sample, with significantly better performance for the close than for the far distance (mean difference = 2.1%, p = .009). There was one difference: in the subset, the close distance led to significantly better accuracy for non-words with the male speaker (mean difference = 5.4%, p = .002, **d = .31**), whereas in the full sample the significant difference was for the female speaker. Overall, the patterns for the subset (Figure 2, left panel) and for the full sample (Figure 1) were quite similar.

Comparing the left and right panels of Figure 2 shows how the results for the 57 participants changed over time. The significant overall improvement in accuracy seems to have primarily been due to better performance on the stimuli that were originally most challenging.

As a result, performance on far stimuli no longer was significantly worse than on close stimuli (mean difference < 0.1%, p = .947), nor were items produced by the female speaker more difficult than those by the male speaker (mean difference = 0.9 %, p = .376).

Discussion

Prior research has shown that audiovisual information benefits the comprehension of difficult speech, such as compressed speech or speech in noise, but the role of visual information in understanding accented speech has not been studied extensively. In this study, we investigated the effect of visual quality, here operationalized as the apparent distance from the speakers, on accented speech recognition. We had participants make lexical decisions while seeing and hearing Chinese accented speakers producing words or non-words at two distances. Our results confirm that having more access to visual lip movements facilitates accented speech recognition, especially under non-optimal listening conditions. We examined whether the visual enhancement is modulated by factors such as the familiarity of the language stimuli (words vs. non-words), speaker differences, or a listener's experience with a particular accent. We found that the effect of distance was more reliable when the stimuli were non-words (compared to words) and when the speaker had a relatively stronger accent. Although a larger set of speakers would be needed to confidently attribute the difference specifically to accentedness, our findings are in line with research showing that accent strength significantly affects the comprehension of accented speech (e.g., Schmid & Yeni-Komshian, 1999). In a study of speech that was challenging for reasons other than accent, Sumby and Pollack (1954) showed that visual information helped more at low speech-to-noise ratios than when the speech was clear. Our results are consistent with this pattern: the impact of lip reading on speech perception is correlated with the difficulty of the listening conditions.

Previous studies suggested that listeners' language experience can potentially affect their perception of dialect variants (Larraza, Samuel, & Onederra, 2016; Sumner & Samuel, 2009; Witteman, Weber, & McQueen, 2013) and non-native phonemes (Wang, Behne, & Jiang, 2008; 2009). Witteman et al. (2013) suggested that listeners' familiarity with an accent affected the speed of perceptual adaptation to accented speech. They found that extensive long-term experience with German-accented Dutch facilitated learning of strongly German-accented Dutch, but limited experience with that accent did not. Sumner and Samuel (2009) also found an effect of long-term experience on the perception and representation of dialect variants. In the current study, we did not observe any effect of real-world linguistic experience on accented speech recognition (either overall, or more specifically by virtue of having a Chinese TA between the two testing sessions). In contrast, the overall improvement from the first session to the second session suggests that participants may have benefited from being exposed to the same Chinese speaker tested at the beginning of the semester. Taken together, we thus see some improvement through exposure to the same speakers, but no improvement due to different speakers. These results are generally consistent with talker-specific learning effects that have been reported in the literature (e.g., Bradlow & Bent, 2003; Gass & Varonis, 1984; Jongman, Wade, & Sereno, 2003): Much perceptual adjustment seems to be based on tuning perception to a particular speaker. Of course, it is possible to generate cross-talker or even cross-accent generalization with extensive training that contains enough variability along the dimension of desired generalization (Baese-Berk, Bradlow, & Wright, 2013; Bradlow & Bent, 2008).

We noted in the introduction that this project was partially motivated by an existing and growing problem: In college classrooms in the US, many teaching assistants speak with foreign accents, and American undergraduates often complain that it is difficult to understand their TA's

speech. In the 2013-2014 academic year, over half of the new graduate students at our university were international students, and over half of these were from China. This motivated the current study to focus on Chinese-accented English, and to examine whether visual cues can help listeners in perceiving accented speech.

A number of our results bear on this real-world issue. Our findings do suggest that having more access to the visual cues (loosely analogous to sitting up front in a classroom) can be helpful when listening to accented speech. The utility of doing so seems to be greatest for non-words (analogous to speech materials that are not familiar), when the speaker's accent is relatively strong. From the perspective of offering real-world solutions, these constraints are encouraging because these are the conditions that are most likely to be causing problems in the first place (e.g., a highly accented ITA using unfamiliar technical terms in the STEM field, such as "arcsine"). Somewhat less positively, our results are consistent with the view that cross-talker generalization is limited (i.e., speaker-specific), so that giving students more general training with accented speech may not help very much. Further research is needed to establish the generalizability of the current findings by including more accent types, more speaker variability, and distance manipulations that capture all of the cues that vary with distance.

The current findings provide insights into accented speech perception. Because we used words and non-words instead of sentences, we were able to measure perception of accented speech without the possible influence of contextually-driven guessing. Our results thus extend previous findings by showing adjustment to accented speech at the word level. Overall, the current research has clarified the contexts in which quantitative variation in visual quality (i.e., variation in apparent visual distance) matters in accented speech recognition. The results broaden our understanding of how listeners use visual information during speech perception

when the input is difficult, and offer a practical idea for improving how American undergraduates can better understand their international instructors.

Acknowledgements

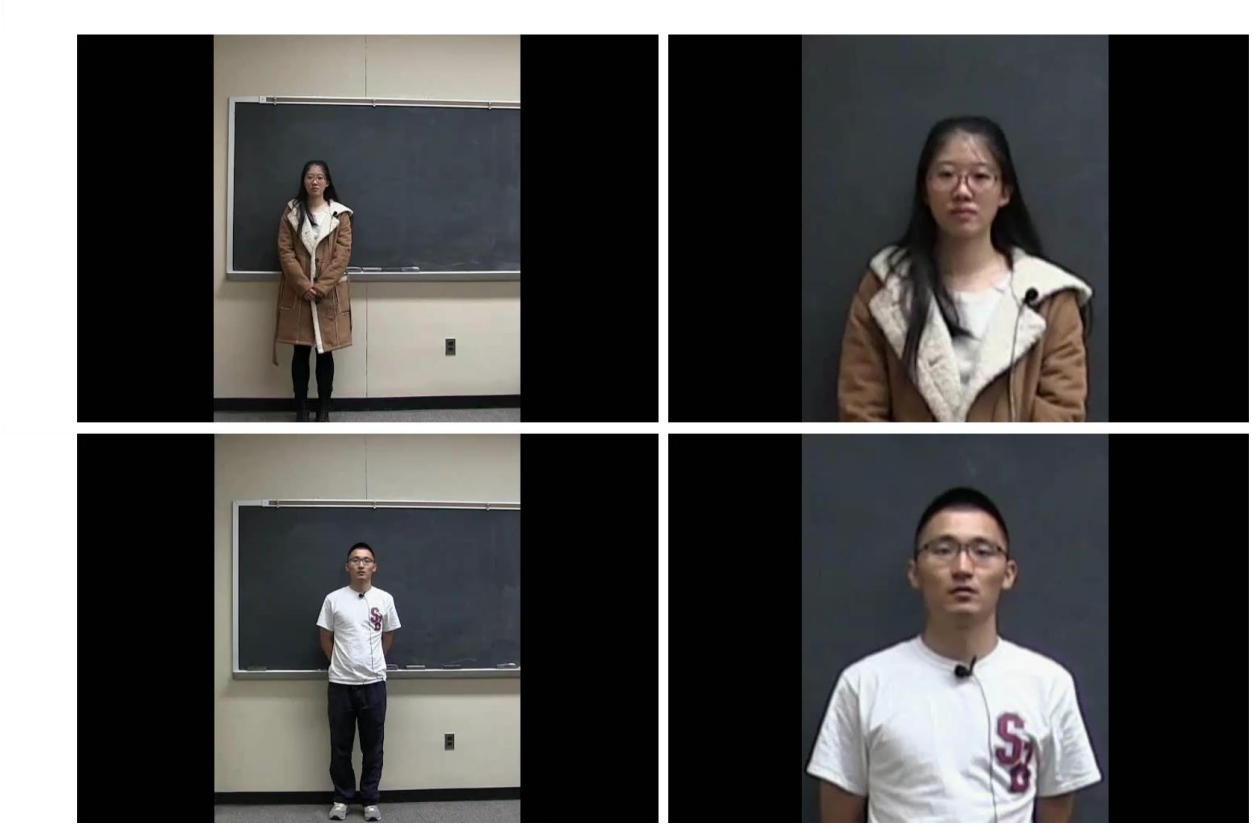
Support was provided by Ministerio de Ciencia E Innovacion, Grant PSI2014-53277, Centro de Excelencia Severo Ochoa, Grant SEV-2015-0490, and by the National Science Foundation under Grant IBSS-1519908. We also appreciate the constructive suggestions of Dr. Yue Wang and two anonymous reviewers.

Appendix A: Stimuli.

Word frequency (per million) is indicated in the parenthesis

Word (frequecy)	**Non-word**	**NSYL**	**Word (frequecy)**	**Non-word**	**NSYL**
scoff (0.42)	snoff	1	osmosis (0.3)	osnosis	3
graph (0.76)	glag	1	median (0.4)	fedian	3
slope (3)	sloge	1	isotope (0.5)	itrophope	3
flash(15.66)	prash	1	electron (0.74)	elestron	3
mass (17.6)	fazz	1	piracy (0.82)	caracy	3
solve (19.88)	bolve	1	density (1.46)	densimy	3
rope (23.16)	wope	1	advertise (1.68)	adverbise	3
square (32.4)	smare	1	transparent (1.84)	cransparent	3
strength (37.66)	prength	1	vertical (2.82)	verchical	3
gas (69.14)	nas	1	institute (3.32)	instipute	3
fight (205.1)	zight	1	triangle (4.36)	triamble	3
right (4088.56)	pight	1	classical (4.48)	brassical	3
vertex (0.04)	verpex	2	creation (6.2)	breation	3
tantrum (0.9)	tancrum	2	develop (9.8)	demelop	3
axis (1.62)	axip	2	pacific (9.86)	pashific	3
index (2.12)	inrex	2	absolute (11.54)	absotute	3
domain (2.64)	donain	2	critical (12.5)	crimical	3
atom (2.8)	abom	2	capital (12.96)	zapital	3
bagel (3.12)	vagel	2	oxygen (14.16)	oxyten	3
carbon (5.34)	darbon	2	reaction (16.6)	reaption	3
reserve (7.32)	resherve	2	occasion (16.88)	offasion	3
auction (9.08)	augren	2	exercise (18.08)	exergise	3
acid (10.16)	anid	2	collection (18.52)	rollection	3
function (11.3)	sunction	2	interrupt (19.1)	inferrupt	3
circle (21.94)	fircle	2	victory (21.88)	nictory	3
rescue (25.92)	reslue	2	vacation (33.46)	macation	3
garden (27.08)	garpen	2	energy (33.56)	enerfy	3
remain (33.88)	relain	2	concentration (5.64)	concenpration	4
master (89)	masber	2	universal (5.98)	cuniversal	4
power (152)	chower	2	original (28.8)	osiginal	4

Appendix B: The close and far distances of the two speakers used in the experiments



Appendix C

ITA Experience Questionnaire

(1) Your native language/first language is _____

(2) Year (circle one): Freshman, Sophomore, Junior, Senior

(3) Have you had any international TAs/instructors/professors (i.e., people who speak with foreign accents)? Yes or No

(4) If your answer to (3) is yes, please indicate below which countries they are from (e.g. China, India, Russia, etc.):_____

(5) Have you had any Chinese TAs/instructors/professors? Yes or No

(6) If your answer to (5) is yes, please list below in chronological order (starting from the most recent one) **when**, in **what** classes, and **how often/much** you heard them speak.

When (e.g. Fall 2016)	What Class	How often: **0-5** (approximately how many days per week?)	How much: (how many hours per week?)

PS: In the second questionnaire, (1) - (4) was replaced by the following two questions:

(1) Do you have any exposure to Chinese accent (family members, friends, teachers, etc.)? Yes or No

(2) If your answer to (1) is yes, please indicate below how familiar you are with this accent (circle a number on the scale below).

1 (not familiar at all) —2—3—4—5 –6—7—8—9—10 (very familiar)

References

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. Neuroimage, 49(1), 1124-1132.

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and aging*, *25*(3), 736-740.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. The Journal of the Acoustical Society of America, 133(3), EL174-EL180.

Banai, K., & Lavner, Y. (2012). Perceptual learning of time-compressed speech: More than rapid adaptation. PloS one, 7(10), e47099.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015a). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. *Frontiers in human neuroscience*, *9* (422), 1-13.

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015b). Cognitive predictors of perceptual adaptation to accented speech. *The Journal of the Acoustical Society of America*, *137*(4), 2015-2024.

Barros, P. C. M. D. (2010). *"It's easier to understand": the effect of a speaker's accent, visual cues, and background knowledge on listening comprehension* (Doctoral dissertation, Kansas State University).

Bernstein, L. E., Auer Jr, E. T., Jiang, J., & Eberhardt, S. P. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in neuroscience*, *7*, 34.

Borjas, G. J. (2000). Foreign-born teaching assistants and the academic performance of undergraduates. The American Economic Review, 90, 355-359.

Bradlow, A. R., & Bent, T. (2003). Listener adaptation to foreign-accented English. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2881-2884). Universitat Autònoma de Barcelona Barcelona.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707-729.

Chun, C. W. (2008). Comments on "Evaluation of the usefulness of the Versant for English test: A response": The author responds. *Language Assessment Quarterly*, *5*(2), 168-172.

Davis, T. M. (2000). Open Doors: Report on International Educational Exchange.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, *5*(2), 160-167.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech, Language, and Hearing Research*, *12*(2), 423-425.

Erber, N. P. (1971). Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *Journal of Speech, Language, and Hearing Research*, *14*(3), 496-512.

Finder, A. (2005). Unclear on American campus: What the foreign teachers said. New York Times, pp. A1, A18.

Fitch, F. & Morgan, S. E. (2003). "Not a lick of English": Constructing the ITA identity through student narratives. Communication Education, 52(3-4), 297-310.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, *34*(1), 65-87.

Grossman, L. A. (2011). The effects of mere exposure on responses to foreign-accented speech. San Jose State University.

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*(11), 996-1009.

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts a. *The Journal of the Acoustical Society of America*, *119*(3), 1740-1751.

Hosoda, M., Stone-Romero, E. F., & Walter, J. N. (2007). Listeners' cognitive and affective reactions to English speakers with standard American English and Asian accents. Perceptual and Motor Skills, 104(1), 307-326.

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *The Quarterly Journal of Experimental Psychology*, *65*(8), 1563-1585.

Jongman, A., Wade, T., & Sereno, J. (2003). On improving the perception of foreign-accented speech. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 1561-1564).

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. The Journal of the Acoustical Society of America, 136(3), 1352-1362.

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio–visual training enhances auditory adaptation process. *Neuroreport*, *20*(14), 1231-1234.Larraza, S., Samuel, A. G., & Oñederra, M. L. (2016). Listening to accented speech in a second language: First language and age of acquisition effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1774-1797.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, *21*(2), 131-141.

Miller, G., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*, 217-228.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological research*, *71*(1), 4-12. Pilling, M., & Thomas, S. (2011). Audiovisual cues and perceptual learning of spectrally distorted speech. *Language and speech*, 0023830911404958.

Pilling, M., & Thomas, S. (2011). Audiovisual cues and perceptual learning of spectrally distorted speech. Language and speech, 54(4), 487-497.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147-1153.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. Research in Higher Education, 33(4), 511–531.

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research*, *42*(1), 56-64.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212-215.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4), 487-501.

Wang, Y., Behne, D. M., & Jiang, H. (2008). Linguistic experience and audio-visual perception of non-native fricatives. *The Journal of the Acoustical Society of America*, *124*(3), 1716-1726.

Wang, Y., Behne, D. M., & Jiang, H. (2009). Influence of native language phonetic system on audio-visual speech perception. *Journal of Phonetics*, *37*(3), 344-356.

Wayne, R. V., & Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology: Applied*, *18*(4), 419.

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, *75*(3), 537-556.

Yi, H. G., Phelps, J. E., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*(5), EL387-EL393.

Zheng Y. & Samuel, A. G (2017). Does Seeing an Asian Face Make Speech Sound More Accented? *Attention, Perception, & Psychophysics. 79* (6). 1841-1859

Zhou, J. (2014). Managing anxiety: A case study of an international teaching assistant's interaction with American students. Journal of International Students, 4, 177-190.