Revisiting the interlanguage speech intelligibility benefit

Chang Shu, Ian Wilson, Jeremy Perkins

University of Aizu, Japan

m5192107@u-aizu.ac.jp, wilson@u-aizu.ac.jp, jperkins@u-aizu.ac.jp

Abstract

Bent and Bradlow (2003) first discovered evidence for an interlanguage speech intelligibility benefit, essentially non-native listeners finding similar-L1 non-native speech equally or more intelligible than native speech. We have refined their method by using 14 speakers from 7 languages (English, Chinese, Hindi, Japanese, Korean, Russian, and Vietnamese) and using reaction time (RT) to accented speech as a more sensitive measure of intelligibility than transcription tasks. Non-native participants (15 Japanese, 9 Chinese, and 6 Vietnamese) had significantly faster RTs to same-accent speakers than to other L2 speakers. L1 English participants had faster RTs to L1 English speakers than to L2 speakers.

Index Terms: reaction time, interlanguage speech intelligibility benefit, English, Japanese, Chinese, Vietnamese, L2 speech

1. Introduction

Bent and Bradlow [1] did an intelligibility study using noise-embedded spoken data from 1 native English speaker, 2 Chinese speakers of English, and 2 Korean speakers of English. The listeners who transcribed the spoken data were from a variety of first-language backgrounds: monolingual English (n=21), Chinese (n=21), Korean (n=10), and other (n=12) – where "other" meant one of Bulgarian, Dutch, French/Douala, German, Greek, Hindi, Japanese, Serbian, Spanish, or Tamil. The following two findings were especially interesting: (i) non-native listeners found high-proficiency non-native speakers of the same L1 equally as intelligible as they did native speakers, something they called the "matched interlanguage speech intelligibility benefit", and (ii) even when listeners were not of the same L1 as the speakers, they found high-proficiency non-native speakers equally as intelligible as they did native speakers, something they called the "mismatched interlanguage speech intelligibility benefit."

Stibbard and Lee's [2] replication of [1] did not show results that supported [1]; one of their findings, for example, was that native speakers were not more intelligible than non-native speakers even to their fellow native listeners. Some procedures in [2] were different from that of [1], though, making it more difficult to directly compare the results. For example, presentation of stimuli was randomised in [2], eliminating a possible familiarity effect in [1], and the sentences in [2] were not embedded in noise (although no ceiling effect occurred).

Xie and Fowler [3] investigated the intelligibility of native and Mandarin-accented English speech for native English and native Mandarin listeners, specifically at the acoustics of stop voicing. They also made a distinction between interlanguage speech intelligibility benefit for listeners and that for speakers.

One drawback, however, of both [1] and [2] is that they each used only five speakers: one native speaker of English and two non-native speakers from each of the respective languages studied (Chinese and Korean in [1] and Saudi and Korean in [2]). And in [3], only one native English speaker and one Mandarin speaker were used. Thus, as pointed out in [2], results could have been greatly affected by idiosyncrasies of the speakers' speech rather than a given foreign accent in general. The same limitation is true of a study by Chen [4], who investigated 29 native and non-native listeners' perceptual judgements of the intelligibility of Chinese-accented speech. She had one Cantonese speaker from HK and one Mandarin speaker from Taiwan, although the listeners were from various language backgrounds. On a dictation task measuring intelligibility, all groups scored higher for the Mandarin accent than for the Cantonese accent, except for Cantonese listeners. This suggests a benefit for listeners listening to speech produced with their own accent.

All four of the studies discussed above used a dictation task to measure intelligibility. Using a more sensitive measurement such as reaction time (RT) can improve the sensitivity of an intelligibility test [5]. Measurements of RT have been used in speech intelligibility tests for over 50 years [6], and speech spoken with a foreign accent is indeed less intelligible (i. e., people take longer to react to it – meaning RTs are longer) than native speaker speech, at least to a native listener [7].

Studies have found that listeners could transcribe utterances perfectly, even though they rated the speakers as heavily accented, indicating that accent does not necessarily result in reduced intelligibility [8], even though it could still increase RT. Although RTs to foreign-accented speech are initially slower than to native speech, it has been shown that listeners can very quickly adapt (in as few as 2–4 utterances) [7]. In that study, native (Tucson, AZ) English speakers' RTs were measured as they listened to the speech of (i) a native speaker of English, (ii) a non-native Spanish-accented speaker of English, and (iii) a non-native Chinese-accented speaker of English. Results showed that RT to non-native speech was slower than RT to native speech, but that the difference diminishes within 1 minute of exposure.

RT is also slower when listening to a dialect of one's native language (L1) that is different from one's own dialect [9]. Native French listeners' RTs to various dialects of French were measured, and they found a significant cost to listening to a different dialect of one's L1 – a 30 ms delay in word identification. Unlike [7], they used multiple speakers for each of the dialects of the L1. In a follow-up study [10], it was found that accent changes cause a temporary perturbation in RTs and that this delay in word identification does not disappear after repeated exposure to the same accent.

In this research, we tested for the interlanguage speech intelligibility benefit reported in past research [1], but we used a total of 14 speakers of 7 languages, 30 listeners, and used arguably a more sensitive measurement of intelligibility: the RT of participants identifying one of two images on the screen – the image corresponding to the audio prompt.

2. Method

2.1. Participants

Thirty-five participants took part in the RT experiment (26 male, 9 female), including 5 native English speakers, 15 native Japanese speakers, 9 native Mandarin Chinese speakers and 6 native Vietnamese speakers. In the English L1 group, there were 3 Canadians, 1 American, and 1 New Zealander. In the Japanese L1 group, there were 2 graduate students and 13 undergraduate students. In the Chinese and Vietnamese groups, all participants were either graduate students or working full time. All L2 participants had studied English for at least 7 years, although to different proficiency levels. Most of them (27 out of 30) had taken the TOEIC English proficiency test in the last 2 years, with scores ranging widely from 265 to 960. All participants were right-handed, except one Mandarin Chinese speaker.

2.2. Stimuli

Eight pairs of words were selected as stimuli, and each word in a pair shared similarities. Table 1 shows the list of stimuli, all of them nouns. The first number is the frequency ranking in the Corpus of Contemporary American English (CCAE) [11], where 1 = the most frequent word in American English, and the column with a two-digit code shows the approximate grade (Gr) that the word is learned in public Japanese schools (J = Junior High; S = Senior High). Note that both words in a given pair were approximately balanced in frequency, and all were common English words.

Table 1: *Stimuli list with frequency and grade level*

Pair	Stimulus 1			Stimulus 2		
No.	Word	CCAE	Gr	Word	CCAE	Gr
1	food	367	J2	foot	381	J2
2	glass	823	J1	gas	1026	J3
3	nose	1748	J2	snow	1795	J2
4	shape	1273	S1	shoe	1430	J2
5	cat	1788	J1	hat	2033	J2
6	flight	1302	J3	fight	1573	J3
7	lake	2204	J2	cake	2563	J1
8	wall	572	J1	ball	915	J2

Combinations of simultaneous visual and audio prompts were used to present the stimuli. Eight image pairs were created, with the left words in Table 1 on the left and the right words on the right (e. g. in Image 1, a picture of food on the left and foot on the right). All 16 stimuli were inserted into the carrier sentence “The picture you should choose is ______.” and they were read and recorded by 14 speakers in a light-type soundproof booth [12]. The 14 speakers were all University of Aizu professors; 2 speakers from each of 7 different countries (Canada, China, India, Japan, Korea, Russia, Vietnam).

A total of 224 audio-visual stimuli (8 image pairs × 2 words per pair × 14 speakers) were created, and then divided into 2 blocks. In Block 1, 1 randomly-chosen word from each pair was read by 1 randomly-chosen speaker from each of the 7 countries. In Block 2, the other word from the same pair was read by the other speaker from each of the 7 countries.

So, each block contained 112 stimuli and each participant (listener) was asked to listen to one of the blocks. For every participant from a given L1, we alternated the choice of block, so the first Japanese participant did block 1, the second block 2, the third block 1, etc.

2.3. Data collection

Before doing the RT experiment, participants filled out a questionnaire asking about handedness, English learning experience, standardized test scores, etc. All participants were offered money to participate, but a few of them refused to be paid. The experiment was conducted in the same soundproof booth as the stimuli were recorded in, ensuring a quiet environment.

Before starting the actual experiment, each participant had brief training. In the training session, participants were requested to respond as quickly as possible to 3 pairs of stimuli, which had the same kind of combination of picture and sound (but were different words from the ones used in the actual experiment). The training stimuli were all spoken by a Colombian speaker of Spanish, and his voice was not used outside of training. Participants adjusted the volume of the sound in their headphones to a comfortable level.

The input device used was an Xbox 360° controller joystick, and participants had to press the left or right button with their left or right index finger, as soon as they determined which picture matched the word they heard. The order of stimuli presentation was randomised by E-Prime 2 software running on an HP EliteBook 8570w laptop computer. As RT was measured from the beginning of the target word (the last word of the sentence) and the longest sound file was about 4.4 seconds, we allowed the images to be displayed after the audio prompt had stopped until a response was given, or until a response deadline of 8 seconds was reached.

2.4. Data analysis

We first analyzed the RT data generated by E-Prime 2 and found that participants had answered incorrectly in 7.2% of trials. All such trials were eliminated from further analysis. No single participant had an overall error rate greater than 20%. Responses from five individual stimuli were excluded from analysis due to error rates greater than 40% across all participants. Four of these stimuli were of non-native English speakers saying “food” (two Chinese speakers: 84% and 69% error rates; one Japanese speaker: 50%; and one Vietnamese speaker: 58%). The remaining stimulus was of a non-native speaker of Russian saying “foot” (63%). “Food” and “foot” have vowel length and quality differences that are difficult for non-native speakers, both in production and perception, possibly causing high error rates.

Also, responses were excluded that were more than 3 times the median average deviation (MAD) from the median RT, calculated separately for each listener. MAD was used rather than SD since it is less influenced by outliers, following [13]. Following [14], we calculated MAD per listener, rather than over the entire data set, since RTs had a large inter-speaker variation. A total of 502 out of 3920 responses were excluded (12.8%), leaving 3418 responses for analysis. Of these 502 responses, 313 (8.0%) were excluded because of high error rates, and 189 (4.8%) were excluded because they exceeded the threshold of 3 times the MAD from the participant median.

The lme4 package [15] and the lmerTest package [16] were used in R [17] to perform a linear mixed effects analysis of the relationship between RT and the factors Participant Language (PL), Speaker Language (SL), and Language Relation (LR). Incorrect responses were omitted from the analysis. The factor LR was coded according to whether the speaker and listener spoke the same native language or not (LR has two levels: ”same” or ”different”). The final model had three fixed effects (PL, SL, and LR) with no interaction terms.

As random effects, intercepts were included for Word and

Participant, but we did not include random slopes. The linear mixed model fit was performed via the REML method, with rival models assessed by likelihood ratio tests, incrementally removing fixed effects. Alternate models with interaction terms added were assessed as well, using the maximum-likelihood method. To assess whether a fixed effect parameter estimate was significant, t-tests were performed using Satterthwaite approximations to degrees of freedom. Type-III F-tests were used to assess the significance of a given fixed effect.

Since speaking rate differed across speakers, and since we did not want to artificially alter any stimuli, we checked Speaker as a random effect and found that it was insignificant in an analysis of random effects (p = 1). As a result, Speaker was excluded as a random effect in the final model.

3. Results

Mean RTs categorized by native English status of speaker and listener and by language relation (LR) are summarized in Table 2. The fastest RTs were by native English participants listening to native English speakers (651 ms), and the slowest RTs were by non-native participants listening to other non-native participants who did not share the same L1 (851 ms). Linear mixed effects analysis details are given in the following paragraphs.

Table 2: *Mean RTs (ms) categorized by native English status of participant & speaker, and by language relation*

Part. L1	Spkr L1	RT (ms)	S.E.	n
Engl.	Engl.	651	28	77
	Non-Engl.	681	8.4	2056
Non-Engl.	Engl.	807	9.5	430
	Non-Engl. (same)	811	17	430
	Non-Engl. (diff.)	851	16	425

Significant effects on RT for speaker language, $F(6, 3381) = 5.6$, $p < 0.01$, participant language, $F(3, 31) = 8.3$, $p < 0.01$, and LR, $F(1, 3381) = 5.3$, p = 0.02, were discovered. The model intercept estimate, using the condition with native English listeners and native English speakers (LR = "same") yielded an RT of 615 ms (SE = 86, $p < 0.01$). A significant positive effect was found on RT when listener and speaker had different native languages ($\beta = 31.748$, SE = 13.8, p = 0.02).

In addition, all participants responded more quickly to stimuli spoken by a native English speaker than to non-native speakers of Chinese ($\beta = 74.8$, SE = 16.7, $p < 0.01$), Hindi, ($\beta = 49.1$, SE = 16.6, $p < 0.01$), Japanese ($\beta = 43.4$, SE = 16.9, p = 0.01), Korean ($\beta = 70.6$, SE = 16.6, $p < 0.01$), and Russian ($\beta = 46.3$, SE = 16.4, $p < 0.01$); however the RTs were not significantly slower in responses for the Vietnamese speakers ($\beta = 13.5$, SE = 16.3, p = 0.41).

Mean RTs are listed by native language of speaker (columns) and participant (rows) in Table 3. In general, native English participants responded faster to all stimuli regardless of the speaker's native language (relative to Chinese participants: $\beta = 286$, SE = 102, $p < 0.01$; and Vietnamese participants: $\beta = 362$, SE = 110, $p < 0.01$); however, there was no significant difference between the RTs of native English and native Japanese participants ($\beta = 13.9$, SE = 94.1, p = 0.88).

There was no evidence of significant interactions between LR and SL, or between LR and PL (model AIC with interaction = 47854; model AIC without interaction = 47850; $\chi^2(3) = 2.56$, p = 0.46 for both likelihood ratio tests). This indicates that RTs were faster if the listener and speaker spoke the same native language, independent of the native language.

Table 3: *Mean RT (ms) by speaker L1 (columns) and participant L1 (rows); letters are the languages English, Chinese, Hindi, Japanese, Korean, Russian, and Vietnamese*

Part.	Speaker L1							
L1	E	C	H	J	K	R	V	All
E	651	704	659	680	701	690	656	677
C	933	974	957	1013	997	974	910	965
J	666	707	723	657	710	690	688	691
V	972	1077	1066	1001	1092	1048	969	1032

Of the 16 stimuli used, all had mean accuracy rates higher than 90%, except for "food" and "foot", which had rates of 71.8% and 75.1% respectively. Among correct responses, these two words also had significantly longer RTs than other words, and were the only words to have mean RTs higher than 1000 ms across all speakers. Notably, "shape" and "shoe" had relatively high RTs (958 ms and 968 ms respectively), perhaps because they were the only other pair with identical syllable onsets.

English proficiency (TOEIC test scores) were collected from non-native English speaking participants. Proficiency did not significantly lower RT; in fact, a small but significant effect was seen where proficiency correlated positively with RT ($\rho = 0.058$, t = 3.034, $p < 0.01$). For the purposes of the linear mixed model analysis, it was decided to exclude the effect of proficiency, and instead include participant as a random intercept. The correlation between RT and proficiency is illustrated in Figure 1.

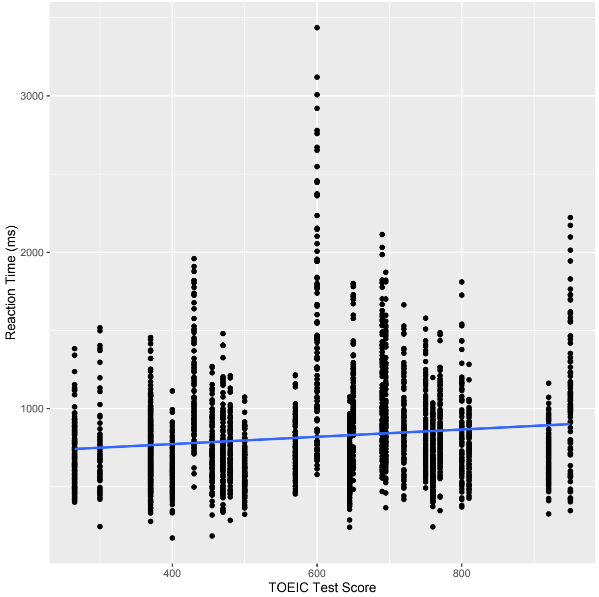

Figure 1: *RT (ms) plotted against English proficiency (TOEIC) test scores of participants*

4. Discussion

Using a more sensitive measure of intelligibility than transcriptions of speech in noise, we found that non-native listeners find speech significantly more intelligible if it is produced by speakers from the same accent group rather than a different accent group (e. g., Japanese speakers find Japanese-accented English more intelligible than other-language-accented English).

Non-native listeners even find their own accent group's L2 English more intelligible than native English, something that agrees with results found in [1]. Note that in [1], this was only found when the speaker was high proficiency, though, but the speakers we used had a great variety of degree of accent. For example, the Japanese speakers in our study are both low proficiency speakers of English, and yet Japanese listeners had faster RTs to their accented English than to native English.

All 14 speakers were professors at the same university, and all L2 participants except one were students at that university. It is certainly possible that some students recognised some voices and this may have affected the RTs. On the other hand, most of the graduate students who have a research supervisor of the same L1 speak to that professor in the L1 instead of in English. One of the Vietnamese participants was surprised to learn (after the experiment) that his research supervisor with whom he meets weekly was one of the 14 speakers. It is not so surprising, though, given the fact that that student and his supervisor converse in Vietnamese, not English, during their meetings.

Instead of simply using the target word in our audio prompts, we decided to embed those words in a carrier sentence. The reason for this is because it gives participants about 2 to 3 seconds to adapt to the speaker's accent and it has been shown that adaptation can be done in a short time [7]. The carrier sentence may also help to minimise RT perturbations that were found in [10] when the speaker changes from trial to trial. When a student is listening to a lecture in accented English, they have ample time to adapt to the speaker's accent, so using a carrier sentence helped to make the situation more realistic. It should be pointed out, though, that the participants' RTs may have been affected by expectations resulting from the pronunciation of the carrier sentence. One participant, a native speaker of English, told us that he was more likely to press the joystick button sooner when the carrier sentence was spoken by another native speaker because he had more confidence that a pronunciation error would not be made.

It is somewhat surprising that Japanese participants had RTs that were not significantly different from native English participants. One possibility for this is age; the native English participants were in their 30s and 40s, except one in his mid-20s, while the Japanese participants were all undergraduates about 21 years old.

It should be unsurprising that no strong correlation was found between RTs and the English proficiency scores of the participants, because the words were specifically chosen to be very common words learned before the end of the first year of senior high school. The weak positive correlation may be due to the fact that the graduate student participants, who were older (thus subsequently slower?) than the undergraduates, had higher TOEIC scores.

5. Conclusions and future work

In conclusion, L2-English participants (Japanese, Chinese, and Vietnamese) had significantly faster RTs to same-L1 speakers' English than to different-L1 speakers' non-native English, a type of matched interlanguage speech intelligibility benefit [1]. Native English listeners had faster RTs when listening to native English speakers than when listening to L2 speakers of English.

In the future, we would like to see how effective it would be to train students to perceive accented speech (e. g. crosslinguistic phonetic and phonological differences). We can then measure effectiveness by comparing pre- and post-training RTs.

6. Acknowledgements

This research was supported by 2016 University of Aizu Competitive Research Funding awarded to Ian Wilson. The authors are grateful to all 15 University of Aizu professors whose voices were recorded for stimuli, and to all the participants.

7. References

[1] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1600–1610, 2003.

[2] R. M. Stibbard and J.-I. Lee, "Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 433–442, 2006.

[3] X. Xie and C. A. Fowler, "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in mandarin speakers of english," *Journal of Phonetics*, vol. 41, pp. 369–378, 2013.

[4] H. C. Chen, "Judgments of intelligibility and foreign accent by listeners of different language backgrounds," *The Journal of Asia TEFL*, vol. 8, no. 4, pp. 61–83, 2011.

[5] C. Gooskens, "Experimental methods for measuring intelligibility of closely related language varieties," in *The Oxford Handbook of Sociolinguistics*, R. Bayley, R. Cameron, and C. Lucas, Eds. Oxford, UK: Oxford University Press, 2013, pp. 195–213.

[6] M. H. L. Hecker, K. N. Stevens, and C. E. Williams, "Measurements of reaction time in intelligibility tests," *Journal of the Acoustical Society of America*, vol. 39, no. 6, pp. 1188–1189, 1966.

[7] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented English," *Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[8] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[9] C. Floccia, J. Goslin, F. Girard, and G. Konopczynski, "Does a regional accent perturb speech processing?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 5, pp. 1276–1293, 2006.

[10] C. Floccia, J. Butler, J. Goslin, and L. Ellis, "Regional and foreign accent processing in English: Can listeners adapt?" *Journal of Psycholinguistic Research*, vol. 38, pp. 379–412, 2009.

[11] M. Davies. (2008) The corpus of contemporary american english: 520 million words, 1990-present. [Online]. Available: http://corpus.byu.edu/coca/

[12] Kawai. (2015) Kawai musical instruments manufacturing co. catalog. [Online]. Available: http://www.kawai-os.co.jp

[13] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[14] R. Ratcliff, "Methods for dealing with reaction time outliers," *Psychological Bulletin*, vol. 114, no. 3, pp. 510–532, 1993.

[15] Douglas M. Bates and Martin Maechler and Ben Bolker, *lme4: Linear mixed-effects models using S4 classes*, R package version 1.1-12, 2016.

[16] A. Kuznetsova and P. B. Brockhoff and R. H. B. Christensen, *lmerTest: Tests in Linear Effects Models*, R package version 2.0-32, 2016.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/