ACCENT RATING BY NATIVE AND NON-NATIVE LISTENERS

Mirjam Wester and Cassie Mayo

Centre for Speech Technology Research,
University of Edinburgh, UK

ABSTRACT

This study investigates the influence of listener native language with respect to talker native language on perception of degree of foreign accent in English. Listeners from native English, Finnish, German and Mandarin backgrounds rated the accentedness of native English, Finnish, German and Mandarin talkers producing a controlled set of English sentences. Results indicate that non-native listeners, like native listeners, are able to classify non-native talkers as foreign-accented, and native talkers as unaccented. However, while non-native talkers received higher accentedness ratings than native talkers from all listener groups, non-native listeners judged talkers with non-native accents less harshly than did native English listeners. Similarly, non-native listeners assigned higher degrees of foreign accent to native English talkers than did native English listeners. It seems that non-native listeners give accentedness ratings that are less extreme, or closer to the centre of the rating scale in both directions, than those used by native listeners.

Index Terms— Perceptual evaluation, native vs non-native listeners

1. INTRODUCTION

Accent rating—the degree of foreign accent or type of accent of a talker—is a perceptual evaluation task that is relevant to a variety of different tasks within speech technology, e.g., in computer assisted language learning [1, 2], for accent conversion [3, 4], for accent identification [5, 6], to reduce the impact of non-native accents on word error rates in ASR [7, 8], and in the context of adverse listening conditions [9]. The study presented here was conducted in the context of an EU project which aimed for personalized speech-to-speech translation such that a user's spoken input in one language was used to produce spoken output in another language, while continuing to sound like the user's voice [10]. Accent rating experiments were conducted to select the talkers from the EMIME bilingual database [11, 12] with the least degree of perceived foreign accent for use in talker discrimination experiments [13]. This paper discusses how the accent rating experiments were carried out and looks specifically at the influence of listener native language with respect to talker native language on the perception of degree of foreign accent.

There are many factors that influence the degree of foreign accent that a non-native talker is perceived to have. For example, age of onset of second language learning, years of formal instruction, length of residence in the second language environment, gender, language learning aptitude, and language use all have an impact on how foreign-accented a talker is perceived to be (see e.g., [14] for a review). Research has shown there is also an influence of the listening *task* on perception of foreign accent [15]. Further studies [16, 17] have shown that there is also a role for *listener*-specific factors in perception of foreign-accentedness. Specifically, the role of native versus non-native status of the listener in the perception of talker accentedness has been the focus of a number of studies e.g, [18, 19, 20].

Flege [18] looked at "whether non-natives who themselves speak L2 with a foreign accent, can gauge foreign accent accurately" and found that native and non-native listeners both give native talkers lower accented ratings than they give to non-native talkers, but for the non-native listeners the difference between the perceived accentedness of native and non-native talkers is much smaller than that perceived by the native listeners. Flege focussed on the effect of non-native listeners' experience with a second language (L2), rather than on other listener-related factors such as listener native language (L1). We would like to know to what extent listener factors other than degree of experience with L2 impact on perceived foreign accent. In particular, what is the role of listener L1 with respect to talker L1 in talker accentedness ratings?

A study carried out by Bent and Bradlow [19] showed how native language background influenced the intelligibility of native and non-native English speech. Native English listeners found native English talkers most intelligible and for non-native listeners, non-native (highly proficient) talkers were as intelligible as native talkers. The results of Bent and Bradlow's study suggest that the language background of the listener with respect to a talker—encompassing native versus non-native, but also shared L1 versus different L1s—can affect the perceived intelligibility of a talker. However, while intelligibility may be part of and/or related to accentedness, there is evidence that the two are not equivalent [21, 22]. The question is, therefore, whether shared L1 between talker and listener can affect the perception of talker accentedness, as found for talker intelligibility. A study by Munro and colleagues [20] aimed to address this issue (see also [23, 24]) and found that there was actually very little difference between listener groups in assessment of the three different aspects of accentedness: intelligibility, comprehension and accentedness. However, whereas Flege and Bent and Bradlow included native talkers in their study, Munro and colleagues, on the other hand, did not elicit accent ratings for native talkers. Without such ratings it is impossible to determine whether the accent ratings seen by Munro et al. [20] pattern with those of Flege [18]—which showed an apparent quantitative difference between native and non-native listener behaviour—or with the intelligibility scores of Bent and Bradlow [19]—which showed a more qualitative difference between native and non-native listener behaviour.

The current study aimed to address this issue by (i) collecting accentedness ratings of both native and non-native talkers, (ii) from native and non-native listeners, (iii) with the non-native listeners either sharing or not sharing an L1 with the non-native talkers. The results of this study will allow us to determine the influence of lis-

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

tener native language with respect to talker native language on the perception of degree of foreign accent.

In the first part of this study (Experiment I), native English, Finnish and German listeners rated the accentedness of native English, Finnish and German talkers producing a controlled set of English sentences. In the second part of the study (Experiment II), native English and Mandarin listeners rated the accentedness of native English and Mandarin talkers producing the same sentences as in Experiment I; additionally, the English and Mandarin listeners also rated the accentedness of either the Finnish or the German talkers from Experiment I.

2. METHOD

2.1. Talkers

In total, 56 talkers were included in this study. There were three groups of non-native adult talkers, each consisting of seven males and seven females with German, Finnish or Mandarin native language backgrounds. The non-native talkers were all recruited via the Edinburgh University Careers Services to be included as subjects in the EMIME bilingual database [11, 12]. The non-native talkers learned English in a variety of places and/or from a variety of English-accented teachers (covering Scottish, American, Southern British English, Australian and Canadian English accents); additionally, some non-native talkers had been exposed to more than one variety of English.

Work by Flege [16] has shown that it is important to include native talkers in a non-native accent rating task. Therefore, for this study, a group of native English speakers (seven males and seven females) was also recorded for the EMIME database. The native English talkers were selected locally from staff and students in the School of Informatics at the University of Edinburgh. As the accents of the non-native talkers cover a range of English accents, a variety of native English accents were also included in the experiment. Both male and female talker sets included two Scottish, two Southern-English and two American talkers. In addition, there was also one New Zealand female and one Australian male talker.

2.2. Materials

The stimuli used in the current study were English sentences which were selected from the EMIME bilingual database. For all talkers the same sentences were used. The selected sentences were:

- Sometimes it helps to take a step back.
- A second meeting is reportedly scheduled for today.
- Microbiology is the study of organisms that cannot be seen by the naked eye.

2.3. Listeners

Accent ratings for Experiment I were collected from three groups of listeners: (i) 28 native, monolingual English listeners (18 female and 10 male), (ii) 24 native German listeners (14 female and 10 male), and (iii) 24 native Finnish listeners (11 female and 13 male). One of the male German listeners was omitted from further analyses as his responses were incomplete. For Experiment II, accent ratings from two groups of listeners were collected: (i) 24 native monolingual English speakers (12 female and 12 male), and (ii) 24 native Mandarin listeners (12 female and 12 male). The reason that two experiments were carried out, rather than one, was that Mandarin-English was not available until after Experiment I had been completed.

All German, Finnish and Mandarin listeners reported being fluent in English. None of the native English listeners reported being fluent in any other languages. German listeners started learning English between 0 and 13 years of age (mean age = 9 years) and were in an English speaking environment for a period of 0.5 to 19 years (mean time = 4 years). Finnish listeners started learning English between 3 and 11 years of age (mean age = 8) and were in an English speaking environment for a period of 0 to 42 years (mean time = 4 years). Mandarin listeners started learning English between 0 and 14 years of age (mean age = 9) and were in an English speaking environment for a period of 0.5 to 22 years (mean time = 3 years).

Most listeners were recruited at the University of Edinburgh. Twelve of the Finnish listeners were recruited at Aalto University, Helsinki. None of the listeners had any known hearing, speech or language problems. All listeners were paid for their participation.

2.4. Experimental design

In Experiment I there were four different test conditions: 1) German and English female talkers, 2) German and English male talkers, 3) Finnish and English female talkers and 4) Finnish and English male talkers. In Experiment II there were six different test conditions: the four above test conditions, plus 5) Mandarin and English female talkers, and 6) Mandarin and English male talkers.

Each test condition consisted of 84 trials: 14 talkers x 3 sentences x 2 repetitions. Within a test condition, trials were divided into six blocks of 14 utterances. Each block consisted of 14 presentations of the same sentence, produced once by each of the 14 talkers. The order of the talkers was different for every block to control for any possible effect of talker order. To control for any possible effect of order of presentation of the sentences, there were six different orders for the blocks (3 sentences x 2 repetitions), and each listener was assigned to one of these six orders. Within a test condition, presentations alternated between male and female sets.

In Experiment I, each listener heard all four Experiment I test conditions. In Experiment II, listeners heard the two new test conditions—Mandarin/English male and Mandarin/English female—in addition to either the two Finnish/English conditions (male and female) or the two German/English conditions (male and female).

2.5. Listening task

Both Experiment I and Experiment II were carried out using a web interface. The subjects' task was to click on an audio file, listen to the sentence stimulus and then score the degree of foreign accent for each utterance on a scale from 0 to 6, where 0 = "no foreign accent at all" and 6 = "strong foreign accent". Subjects were informed that they would be listening to three different sentences read by both native and non-native speakers of English. Additionally, they were told the native speakers were from various different English speaking backgrounds. The subjects were instructed not to consider any of these native English accents as foreign. Subjects were free to listen to the utterance as often as they needed to make a judgement. Responses were entered by clicking with a mouse on the relevant button on the web interface.

The listening experiments in Edinburgh were carried out in sound isolated booths. Audio was presented from a Mac mini computer using Beyerdynamic DT 770 PRO headphones. The 12 Finnish subjects, recruited at Aalto University, Helsinki, conducted

the experiment over the web in a quiet environment using high quality audio equipment.

3. RESULTS

In order to examine group trends across the data, listener ratings were first converted to normalised z-scores. As Levene's F test revealed that the homogeneity of variance assumption was not met for these data Welch's ANOVA was used. ANOVA's with listener native language as the between-subjects factor were conducted on the z-scores of the accent ratings for the six different test conditions. Within a test condition, responses to the two talker groups (Finnish/English, German/English and Mandarin/English) were each analysed separately. Post-hoc comparisons using the Games-Howell post-hoc procedure, were conducted to determine which pairs were significantly different from each other.

3.1. Listener Agreement

Before the main analyses were carried out, intra-class correlations [25] were computed on the raw accent judgement data to assess single raters compared with themselves (ICC3) and inter-listener agreement (ICC2) for the various groups of listeners. To summarise, the ICC3 results showed that native and non-native listeners did not differ from each other a great deal in terms of intra-listener reliability, although non-native listeners had somewhat lower ICC3 values than native listeners, and showed a larger degree of variance. The ICC2 results showed that native listeners predominantly had moderate to substantial inter-rater agreement for the various test conditions. The non-native listeners showed lower levels of inter-rater reliability than native listeners. Moderate levels of agreement were achieved by German and Finnish listeners on the Finnish male data set and Mandarin listeners showed much higher degrees of agreement on Mandarin data sets than on the other test conditions.

3.2. Effect of listener native language on the rating of talkers

Figure 1 shows the effect of listener native language in judging accentedness in female talkers and Figure 2 shows the results for male talkers. The results for both groups of talkers are similar.

ANOVAs (female talkers) showed a significant effect of listener native language in all cases: all listeners, English talkers (finset): $[Welch's\ F(3, 1618.2) = 45.2,\ p < 0.0001]$, Finnish talkers: $[Welch's\ F(3, 1734.5) = 153.1,\ p < 0.0001]$, English talkers (gerset): $[Welch's\ F(3, 1667.5) = 34.6,\ p < 0.0001]$ and German talkers: $[Welch's\ F(3, 1827.4) = 120.7,\ p < 0.0001]$. Post-hoc Games-Howell tests revealed that non-native listeners judge non-native talkers as *less* accented than do native listeners, and non-native listeners judge native talkers as *more* accented than do native listeners. A summary of the Games-Howell results for female talkers is given in Table 1.

ANOVAs (male talkers) showed a significant effect of listener native language in all cases: all listeners, English talkers (finset): $[Welch's\ F(3, 1547.5) = 45.3,\ p < 0.0001]$, Finnish talkers: $[Welch's\ F(3, 1705.9) = 74.7,\ p < 0.0001]$, English talkers (gerset): $[Welch's\ F(3, 1621.6) = 29.6,\ p < 0.0001]$ and German talkers: $[Welch's\ F(3, 1720) = 54.7,\ p < 0.0001]$. However, Games-Howell tests show that not all the significant differences between listener groups found for female talkers are also found for male talkers. The main finding that non-native listeners judge non-native talkers as *less* accented than do native listeners and non-native listeners judge native talkers as *more* accented than do native listeners still stands, but there are smaller differences between the different

Table 1. *Summary of Games-Howell results for the female talker sets. Order from left to right is from highest degree of accentedness to lowest degree of accentedness, according to the different sets of listener groups. '=' indicates no significant difference between listener groups and ' >' indicates the direction of a significant difference between listener groups.*

Female talkers	Listener group order
English (Finnish set)	$man > fin > ger > eng$
Finnish	$eng > ger = fin > man$
English (German set)	$fin = man > man = ger > eng$
German	$eng > ger > fin > man$
English (Mandarin set)	$man > eng$
Mandarin	$eng > man$

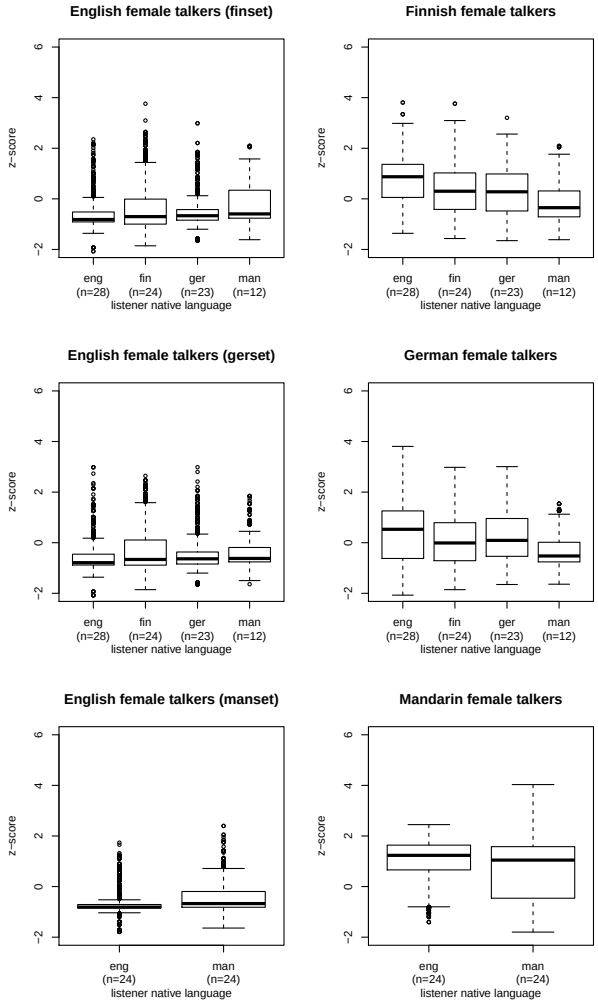

Fig. 1. Boxplot of listeners' z-scores per native language for female talkers. 'n' = number of listeners, 'eng' = English, 'fin' = Finnish, 'ger' = German, 'man' = Mandarin, 'finset' = Finnish/English set, 'gerset' = German/English set & 'manset' = Mandarin/English set (Experiments I and II).

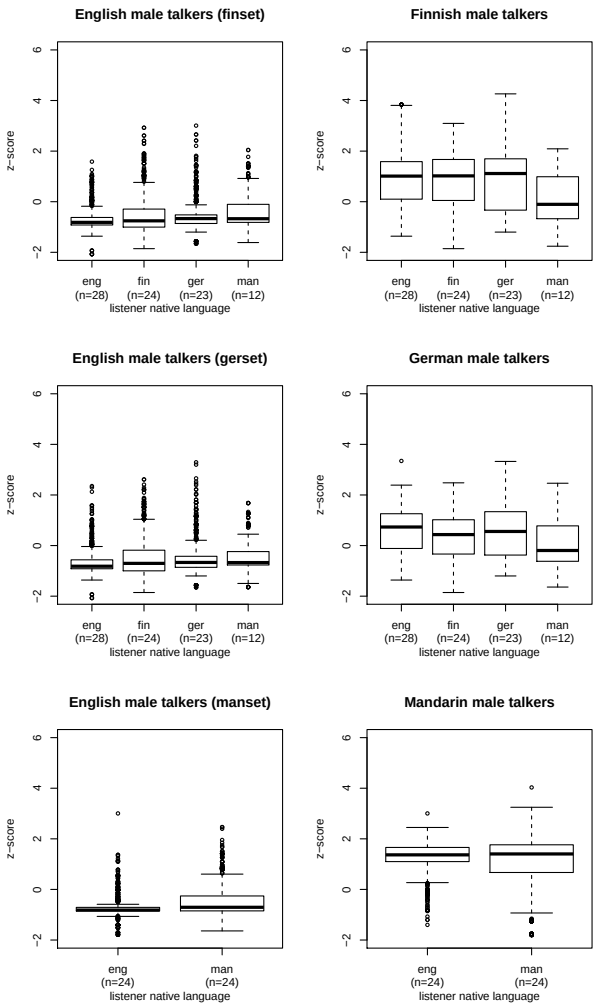

Fig. 2. Boxplot of listeners' z-scores per native language for male talkers. (See caption for Figure 1 for abbreviations.)

non-native listener groups. A summary of the Games-Howell results for male talkers is given in Table 2.

3.3. Summary of listener behaviour by native language group

Examining the pattern of responses grouped by listener native language shows a complex interaction between listener and talker native language. Native English listeners, as noted above, were always the listener group that gave the lowest accentedness rating to the native English talkers, and the highest accentedness ratings to each of the non-native talker groups. However, the same pattern was not observed for the non-native listener groups: that is, non-native listeners did not automatically give talkers from the same non-native language the lowest accentedness ratings. One non-native listener group—native Mandarin listeners—was consistently the group that gave the lowest accentedness rating to talkers from their own native language. However, as the only other listener group to rate the Mandarin talkers was the native English listener group who always rated non-native talkers as highly accented, this result should be interpreted somewhat cautiously. It should be noted, though, that the

Table 2. *Summary of Games-Howell results for the male talker sets. See caption for Table 1 for interpretation of symbols and ordering.*

Male talkers	Listener group order
English (Finnish set)	$man > fin = ger > eng$
Finnish	$eng > ger = fin > man$
English (German set)	$man = ger = fin > eng$
German	$eng = ger > fin > man$
English (Mandarin set)	$man > eng$
Mandarin	$eng > man$

Mandarin listener group was the group that gave the lowest accentedness ratings to all of the non-native talker groups and the highest accentedness rating to the native English talkers. At the other end of the scale, the German listener group was not the group that gave the least accentedness rating to native German talkers: Mandarin listeners and Finnish listeners gave lower accentedness ratings than did German listeners to both male and female German talkers. When listening to native Finnish and native English talkers, the Finnish listeners generally gave accentedness ratings that were the same as, or that sat in between, those given by the Mandarin and German listener groups.

4. DISCUSSION

While there are differences in accentedness rankings of different talker groups by different listener groups, overall all listeners behaved qualitatively the same. Non-native listeners, like native listeners, rated native talkers as relatively unaccented, and non-native talkers as relatively accented. Similarly, non-native listeners and native listeners tended to agree as to which non-native talkers were more heavily accented, and which were less heavily accented. Finally, and most importantly, there was no significant difference between the accent rating behaviour of non-native listeners who shared an L1 with a non-native talker, and those non-native listeners who did not share the talker's L1. Other than for the Mandarin listeners, who consistently rated Mandarin-accented English as less accented than did other listeners, there was no particular advantage of shared L1 between listener and talker. These results lend support to those of Munro et al. [20], and provide further evidence that the matched interlanguage intelligibility benefit demonstrated by Bent and Bradlow [19] does not extend to accentedness. This in turn underlines the proposal by Munro et al. [20] that while intelligibility is part of, or related to, accentedness, the two do not operate in the same way: the acoustic-phonetic features that might increase intelligibility between two non-natives who share an L1 (e.g., lack of reduction) might also increase a non-native talker's perceived degree of accent.

In general, non-native listeners gave accentedness ratings that were less extreme, or closer to the centre of the rating scale in both directions, than those used by native listeners. This trend was even seen for those few non-native talkers who were judged by most listeners to be relatively unaccented. Non-native listeners considered these talkers to be somewhat more accented than did the native listeners. This pattern of results matches that of Flege [18], and thus supports his proposal that experience with a target language gives listeners "more accurate information concerning how the phonetic segments in English 'ought' to sound" ([18], p. 77). Thus, our data and Flege's data suggest that the reduced scale of accentedness used by the non-native listeners reflects an inability to fully detect a native accent.

5. REFERENCES

[1] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[2] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 83–93, 2000.

[3] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009.

[4] D. Felps and R. Gutierrez-Osuna, "Developing objective measures of foreign-accent conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1030–1040, 2010.

[5] K. Amino and T. Osanai, "Native vs. non-native accent identification using Japanese spoken telephone numbers," *Speech Communication*, vol. 56, pp. 70–81, 2014.

[6] F. de Wet, P. Louw, and T. Niesler, "Human and automatic accent identification of Nguni and Sotho Black South African English," *South African Journal of Science*, vol. 103, no. 3/4, p. 159, 2007.

[7] D. van Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," *Speech Communication*, vol. 35, no. 1, pp. 71–79, 2001.

[8] A. Faria, "Accent classification for speech recognition," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 285–293.

[9] J. Volín and R. Skarnitzl, "The strength of foreign accent in Czech English under adverse listening conditions," *Speech Communication*, vol. 52, no. 11, pp. 1010–1021, 2010.

[10] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimäki, R. Karhila, and M. Kurimo, "Personalising speech-to-speech translation: Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," *Computer Speech & Language*, 2011.

[11] M. Wester, "The EMIME Bilingual Database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1388, 2010.

[12] M. Wester and H. Liang, "The EMIME Mandarin Bilingual Database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1396, 2011.

[13] M. Wester, "Talker discrimination across languages," *Speech Communication*, vol. 54, pp. 781–790, 2012.

[14] T. Piske, I. R. A. MacKay, and J. E. Flege, "Factors affecting degree of foreign accent in an L2: A review," *Journal of Phonetics*, vol. 29, pp. 191–215, 2001.

[15] S. V. Levi, S. J. Winters, and D. B. Pisoni, "Speaker-independent factors affecting the perception of foreign accent in a second language," *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2327–2338, 2007.

[16] J. E. Flege and K. L. Fletcher, "Talker and listener effects on degree of perceived foreign accent," *Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 370–389, 1992.

[17] A. R. Bradlow and D. B. Pisoni, "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," *Journal of the Acoustical Society of America*, vol. 106, p. 2074, 1999.

[18] J. E. Flege, "Factors affecting degree of perceived foreign accent in English sentences," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 70–79, 1988.

[19] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *Journal of the Acoustical Society of America*, vol. 114, p. 1600, 2003.

[20] M. J. Munro, T. M. Derwing, and S. L. Morton, "The mutual intelligibility of L2 speech," *Studies in Second Language Acquisition*, vol. 28, pp. 111–131, 2006.

[21] S. K. Sidaras, J. E. D. Alexander, and L. C. Nygaard, "Perceptual learning of systematic variation in spanish-accented speech," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, p. 3306, 2009.

[22] S. Winters and M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of f0 and duration," *Speech Communication*, 2012.

[23] I. R. A. MacKay, J. E. Flege, and S. Imai, "Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent," *Applied Psycholinguistics*, vol. 27, no. 157-183, 2006.

[24] R. C. Major, S. F. Fitzmaurice, F. Bunta, and C. Balasubramanian, "The effects of nonnative accents on listening comprehension: Implications for ESL assessment," *TESOL Quarterly*, vol. 36, pp. 173–190, 2002.

[25] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.