The EMIME Bilingual Database

Mirjam Wester

Centre for Speech Technology Research, University of Edinburgh, UK
mwester@inf.ed.ac.uk

Abstract

This paper describes the collection of a bilingual database of Finnish/English and German/English data. In addition, the accents of the talkers in the database have been rated. English, German and Finnish listeners assessed the English, German and Finnish talkers' degree of foreign accent in English. Native English listeners showed higher inter-listener agreement than non-native listeners. Further analyses showed that non-native listeners judged Finnish and German female talkers to be significantly less accented than do English listeners. German males are judged less accented by Finnish listeners than they are by English and German listeners and there is no difference between listeners as to how they judge the accent of Finnish males. Finally, all English talkers are judged more accented by non-native listeners than they are by native English listeners.

Index Terms: evaluation, accent rating, cross-lingual

1 Introduction

The motivation for recording a bilingual database arose from the EMIME speech-to-speech translation task. In this project, we are aiming for personalized speech-to-speech translation such that a user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice[1]. However, how do we measure whether our modeling attempts are successful or not - that is how are we to measure whether or not a user sounds similar in two different languages? Aside from the complications associated with asking listeners to compare natural speech to synthetic speech there is an even more fundamental question we would like to see answered first. How well do listeners judge speaker similarity across language boundaries when the stimuli consist of natural speech. To investigate this we needed a database of bilingual data. This paper describes the design and collection of this database.

In designing the bilingual database for our talker discrimination experiments we have two assumptions. First of all, we assume talker discrimination is easier when the different languages spoken by individual talkers are from the same language family. That is, listeners should be able to judge more accurately whether or not a talker is the same when the talker is speaking two closely related languages. Secondly, if bilingual talkers are highly fluent in their two languages, talker discrimination should be more difficult. Anecdotal evidence seems to suggest that proficient non-native talkers of English do not necessarily sound like the same person when speaking their native language.

The languages under consideration in EMIME are Japanese, Mandarin, Finnish and English. In order to link to previous research [1] and to be able to look at our first assumption we chose to record German/English (same language family) and Finnish/English (different language families) talkers. The aim of the experiment described in this paper is to select talkers with the least degree of perceived foreign accent[2] because, as stated above, we expect that the more native the bilingual talker sounds in English, the more difficult it will be for

[1] http://www.emime.org

[2] The definition we take of accent is a manner of pronunciation of a language.

listeners to recognize them as the same talker in both their native language (L1) and their second language (L2).

The main objective of this accent rating task is to obtain native listener's judgements, but we were also very interested to find out how non-native listeners perform on the same task. As [2] showed, degree of foreign accent is not only defined by the pronunciation of a specific talker, it also reflects the listener's perception of the L2 speech. Listeners with different language backgrounds may well judge accentedness differently. Especially if the non-native accent is their own type of accent. Therefore, we not only asked native English listeners to participate in the accent rating experiment but we also recruited Finnish and German listeners.

This paper aims to answer the following questions. Which talkers have the least degree of perceived foreign accent? And, do Finnish and German listeners perform differently than English listeners when rating the accentedness of the English speech of Finnish, German and English talkers?

2 Data collection

2.1 Stimulus materials

Three sets of prompts were created; one for each of the languages: English, Finnish and German. Each set contains 25 Europarl sentences, 100 news sentences and 20 semantically unpredictable sentences (SUS). The 25 Europarl sentences were selected from the ACL WMT 2008 test set of the Europarl (proceedings of the European Parliament) parallel corpus [3], i.e. the same 25 sentences were selected for each language. The news sentences for English were taken from the Wall Street Journal 1 corpus [4], comprising 40 enrolment sentences and 60 test set sentences. The Finnish sentences were selected from the Speecon corpus [5]. German news sentences were selected from the test set portion of German Globalphone [6]. In selecting the news sentences an effort was made to ensure they were easy to read by discarding sentences which contained names, long numbers, dates and anything else that looked like it might pose problems for reading aloud. The 20 SUS for English were taken from the Blizzard 2009 set [7], the Finnish set were supplied by TKK (Aalto University, Helsinki) and the German set was handcrafted by a native German speaker at the Centre for Speech Technology Research (CSTR).

2.2 Talkers

In total 42 talkers were recorded. Seven male and seven female talkers for each language: English, German and Finnish. The bilingual talkers were recruited via the Edinburgh University Careers Services, the native English talkers were selected locally from CSTR and Informatics. Bilingual in this context means a person who has the ability to speak and read two languages.

Work by Flege & Fletcher [8] has shown that it is important to include native talkers in a non-native accent rating task. It was found in [8] that the degree of foreign accent is influenced by the proportion of native (or near-native) speakers included in the test set. Increasing the proportion of native speakers in the stimulus set caused the non-native speakers to be rated as more accented. It also ensured that listeners used a wider range of the rating scale. Furthermore, native controls serve to confirm that listeners are correctly performing the task by testing that they can distinguish native from non-native speech.

The bilingual talkers learned English in a variety of places and/or from a variety of English-accented teachers (covering Scottish, American, British, Australian and Canadian English accents); some bilingual talkers had been exposed to more than one variety of English. As the accents of the non-native talkers cover a large part of the English speaking world 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.

2.3 Recording procedure

Recordings were carried out using ProTools HD hardware and software in a hemi-anechoic chamber. Two different microphones were used, a close-talking DPA 4035 mounted on the subjects headphones and a Sennheiser MKH 800 p48 microphone placed about 10cm from the subject using an omnidirectional pattern. The speech was sampled at 96kHz 24bit depth and stored directly to a computer. These recordings were subsequently downsampled, using Pro-Tools to 22 kHz 16bit and segmented into sentence level chunks.

The recordings took about an hour per person to complete. The bilingual talkers first read the English sentences and then their native language with a short break between the two sessions. When they made an error, the talkers were asked to re-read the sentence. A remuneration of £20 was given to the bilingual subjects for their time and effort. The native English talkers took about 20 minutes to record the English sentences and did not receive any remuneration.

3 Accent rating experiment

This section describes the accent rating experiment: the materials used, the design of the experiment and the participants.

3.1 Stimuli

The accent rating experiment was performed on English sentences. Three sentences (short, medium, long) were selected. For all talkers the same sentences were used. The selected sentences are:

- Sometimes it helps to take a step back. (9 syllables)
- A second meeting is reportedly scheduled for today. (15 syllables)
- Microbiology is the study of organisms that cannot be seen by the naked eye. (25 syllables)

3.2 Design

There are four different test conditions:1) German and English females, 2) German and English males, 3) Finnish and English females and 4) Finnish and English males. Each test condition consists of 84 trials: 14 talkers x 3 sentences x 2. Within a test condition there are six blocks of 14 talkers reading the same sentence. The order of the talkers is different for every block to control for any possible effect of talker order. There are six different orders for the three sentences. For each of the four testing conditions six versions were generated to control for sentence order. Each listener was assigned a different selection of the four test conditions, always alternating between male and female sets. Note that in this set-up the same English talkers are encountered in two test conditions, however alternating male and female tests ensured that this was not (too) noticeable for the listeners.

3.3 Listeners

Accent ratings were collected from three groups of listeners; 28 native monolingual English listeners, not fluent in any other languages, 22 native German listeners and 24 native Finnish listeners. Most listeners were recruited at the University of Edinburgh. A subset of 12 Finnish listeners were recruited at Aalto University, Helsinki. None of them had any known hearing, speech or language problems. Subjects were paid for their participation.

The subjects were asked to score the degree of foreign accent for each utterance on a scale from 0 to 6, with “0” = no foreign accent at all and “6”= strong foreign accent. They were told the native speakers were from various different English speaking backgrounds, and that none of these native English accents should be considered foreign.

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 experiment was conducted via a web interface. The 12 Finnish subjects conducted the experiment over the web in a quiet environment using high quality audio equipment. The subjects task was to click on an audio file, listen to the sentence stimulus and click with a mouse on one button in the range from 0 to 6 to indicate their judgement of degree of accent. Subjects were free to listen to the utterance as often as he/she needed to to make a judgement. The experiment can be found at `http://homepages.inf.ed.ac.uk/mwester/accent_rating/start_accent.html`

4 Results

First an analysis of which talkers have the least degree of foreign accent is given. This is followed by determining whether there are significant differences between listener groups.

4.1 Talkers with least degree of foreign accent

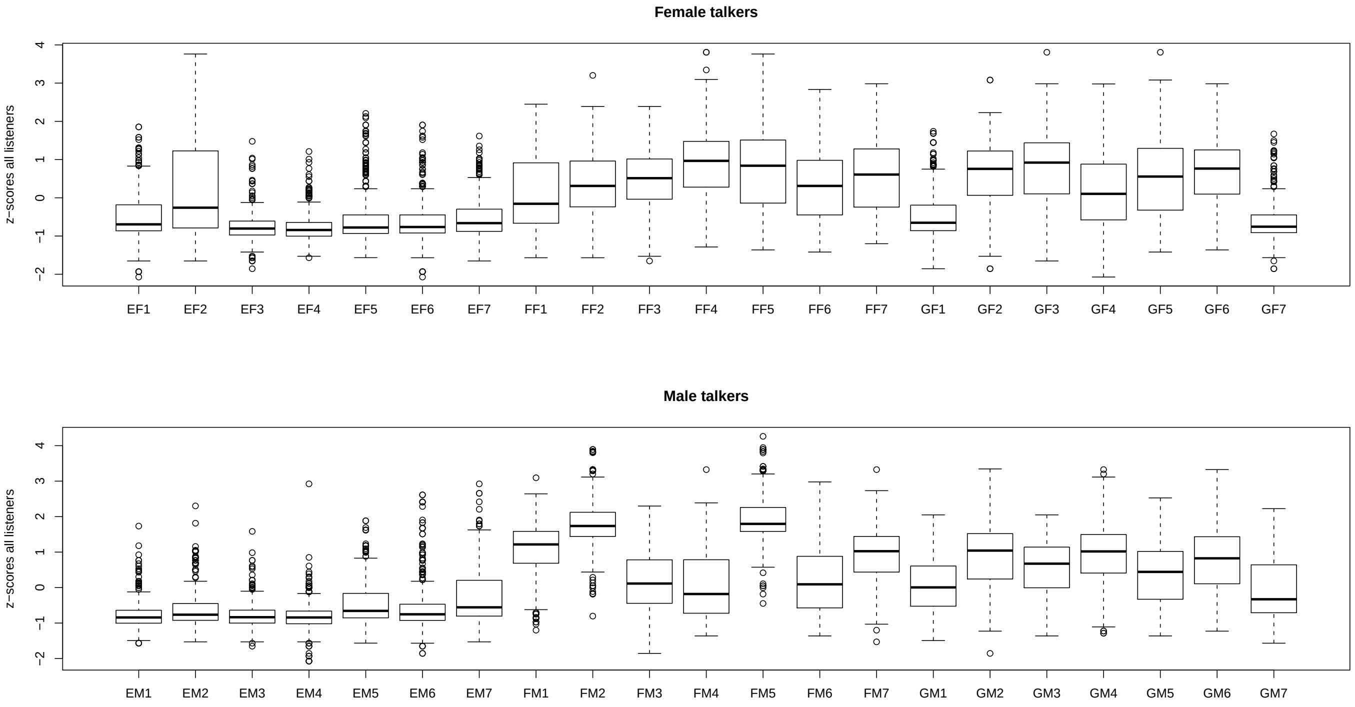

Figure 1: Female and male talkers' z-scores based on 74 listeners' judgements (n=444).

Listener ratings were converted to normalized z-scores. The boxplot in Figure 1 shows the overall z-score results for the female talkers in the top half and the male talkers in the bottom half, as judged by all listeners. Larger z-score ratings indicate a greater degree of foreign accent. Abbreviations in Figure 1 and the rest of this paper are: English female = EF, Finnish female = FF, German female = GF, English male = EM, Finnish male = FM, German male = GM. In order to analyze which talkers differ significantly from each other Tukey

HSD (Honestly Significant Difference) multiple comparisons of means with 95% family-wise confidence level were conducted.

On the basis of the data in Figure 1 the selected female talkers for the talker discrimination experiments are talker FF1, FF2, FF3, FF6 and FF7 and GF1, GF4, GF5, GF6 and GF7. Pairwise comparisons of Finnish female talkers showed that FF1-FF6, FF2-FF3, FF4-FF5, FF5-FF7 and FF6-FF7 do *not* differ significantly from each other, but all other pairs do. For German female talkers, the non-significantly different pairs are GF1-GF7, GF2-GF4, GF2-GF6, and GF3-GF5.

Regarding the English female talkers, EF2 is judged to be non-native rather than native. This is due to a clear hesitancy in her reading, which listeners may perceive as a cue for foreignness. GF1 and GF7 are pretty much regarded as native talkers.As far as comparisons with English talkers are concerned, GF7 is only significantly different from English talkers EF1 and EF2. GF1 is significantly different from EF2, EF3 and EF4. GF4 differs significantly from all English native talkers except for EF2. Finally, FF1, FF2 and FF6 are not significantly different from EF2, but all other comparisons between the ratings of English and Finnish talkers are significantly different.

For the male talkers we can see that there is a larger degree of variation in the degree of foreign accent for the Finnish male talkers compared to the other talker groups. The selected Finnish male talkers are FM1, FM3, FM4, FM6 and FM7. Pairwise comparisons of Finnish male talkers showed that FM1-FM7, FM2-FM5, FM3-FM4, FM3-FM6, FM4-FM6 are non-significantly different pairs. The selected German male talkers are GM1, GM3, GM5, GM6, GM7. All German male talkers differ significantly from each other except for GM1-GM7, GM2-GM4, GM2-GM6 and GM1-GM5. All Finnish and German male talkers are rated significantly different than the English male talkers.

4.2 Inter-listener agreement

Accentedness intra-class correlations on the raw data (ICC2 [9]) were computed to assess inter-listener agreement for the three groups of listeners, see Table 1. We can interpret the ICC values in a similar way to Cohen's Kappa so by convention $K = 0.40$ to 0.59 is moderate inter-rater reliability, 0.60 to 0.79 substantial, and 0.80 outstanding. On this basis, only native listeners show moderate to substantial agreement. The low values for non-native listeners could indicate this is a more difficult task for them than for native listeners.

Table 1: *ICC for inter-listener agreement*

judges	test conditions			
	German female	German male	Finnish female	Finnish male
English	0.51	0.47	0.46	0.65
German	0.26	0.36	0.22	0.59
Finnish	0.29	0.28	0.33	0.51

4.3 Differences between listener groups

An initial analysis of variance (ANOVA) was conducted with order of sentence as the within-subject factor to determine if differences existed between the first and second presentation of each sentence. The ANOVA showed there is no significant effect of the order of the sentences. Next, ANOVAs with sentence as the within-subjects factor and listener nationality as between-subjects factor were conducted on the z-scores of the accent ratings for the different test conditions, the English talkers were analyzed separately from the German and Finnish talkers.

The ANOVA for the Finnish female talkers showed a significant main effect of sentence [$F(2, 3099) = 205.46, p < 0.0001$], a significant main effect of listener nationality [$F(2, 3099) = 73.96, p < 0.0001$], and a significant interaction between sentence and nationality. A Tukey HSD test revealed significant pairwise differences between German & English and Finnish & English listeners ($p < 0.0001$) but no significant difference between Finnish & German listeners ($diff = 0.012$, $p = 0.945$). The direction of the effect showed that non-native listeners score Finnish female talkers lower (less degree of foreign accent) than do native listeners. Tukey's HSD also revealed pairwise differences between sentence 1 & 3 and 2 & 3. Talkers were assigned lower scores on sentence 3 than on sentences 1 and 2. A Tukey HSD test shows for the interaction between sentence and listener nationality that non-native listeners score all three sentences lower (i.e. less accented) than English listeners score sentences 1 and 2. English listeners score sentence 3 lower than the non-native listeners score sentences 1 and 2. And finally, non-native listeners score sentence 3 lower than English listeners score sentence 3.

The ANOVA for Finnish male talkers showed a significant main effect of sentence [$F(2, 3099) = 44.68, p < 0.0001$] but no significant main effect of listener nationality and also no significant interaction between the two factors. All pairs of sentences are scored significantly different ($p < 0.0001$), 2 lower than 1, and 3 lower than both 1 and 2.

The ANOVA for the German female talkers showed the same patterns as for the Finnish female talkers: a significant main effect of listener nationality [$F(2, 3099) = 33.7, p < 0.0001$], a significant main effect of sentence [$F(2, 3099) = 23.6, p < 0.0001$] and a significant interaction between the two [$F(4, 3099) = 3.39, p < 0.01$]. A Tukey HSD test showed that German and Finnish listeners score German female talkers significantly lower than do English listeners ($p < 0.0001$), and that Finnish listeners score German female talkers significantly lower than German listeners do ($p < 0.0001$). There are significant differences between all sentence pairs.

The ANOVA for German male talkers showed a significant main effect of sentence [$F(2, 3009) = 92.8, p < 0.0001$], a significant main effect of listener nationality [$F(2, 3099) = 18.0, p < 0.0001$] and a significant interaction between the two factors [$F(4, 3099) = 3.86p < 0.01$]. A Tukey HSD test shows that Finnish listeners judge German talkers significantly lower than German and English listeners do, but that there is no significant difference between German and English listeners. For German male talkers there is no significant difference between the judging of sentence 2 and 3, but they are both judged significantly lower than sentence 1.

ANOVAs for the English talkers were also conducted. Once again sentence and listener nationality as well as the interaction between the two were all found to be significant. Non-native listeners judge English talkers (male and female) significantly more accented than English listeners do. In this case, there is no significant difference between German and Finnish listeners.

5 Discussion and Conclusions

The accent rating experiment was successful in that it enabled the selection of talkers with the least degree of foreign accent. Tukey HSD tests showed that most of the pairs of talkers are significantly different from each other. The non-significant differences are either between pairs that were selected for the talker discrimination experiment, or between talkers that were omitted. The only exception to this is the non-significant difference between GF2-GF4, GF2-GF6 and GM2-GM6.

We also found that non-native listeners show lower intra-class correlations than native listeners. This could be an indication that the task is more difficult for them than for native listeners. However, the ICC values for native listeners are also not very high. This suggests that listeners' internal standards of accentedness differed.

Findings in this paper show that non-native listeners judge talkers with non-native accents less harshly than English native listeners do. On the other hand, non-native listeners assign higher degrees of foreign accent to native talkers. Maybe a process similar to the interlanguage speech intelligibility benefit is playing a role here [10]. In short [10] showed how native language background influenced the intelligibilty 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.

This study has shown that the language background of the listener has a significant impact on how they judge the degree of perceived accent of a talker. This research adds to the body of work which has investigated a number of different factors which affect the judgement of degree of foreign accent. Finally, a database — The EMIME Bilingual Finnish/English German/English Database — has been created and has been made available under the Open Datatbase License: http://opendatacommons.org/licenses/odbl/1.0/.

6 Acknowledgements

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). Thanks to Timothy Mills and Reima Karhila for running the perception experiment.

References

[1] S. Winters, S. Levi, and D. Pisoni, "Identification and discrimination of bilingual talkers across languages," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4524–4538, 2008.

[2] S. Levi, S. Winters, and D. Pisoni, "Speaker-independent factors affecting the perception of foreign accent in a second language," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2327–2338, 2007.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit 2005*, 2005.

[4] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. of the DARPA Speech and Natural Language Workshop*, 1992, pp. 357–362.

[5] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002, pp. 329–333.

[6] T. Schultz, M. Westphal, and A. Waibel, "The GlobalPhone project: Multilingual LVCSR with JANUS-3," in *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, Plzen, Czech Republic, 1997, pp. 20–27.

[7] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, Edinburgh, U.K, 2009.

[8] J. Flege and K. Fletcher, "Talker and listener effects on degree of perceived foreign accent," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 370–389, 1992.

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

[10] T. Bent and A. Bradlow, "The interlanguage speech intelligibility benefit," *J. Acoust. Soc. Am.*, vol. 114, p. 1600, 2003.