Cross-lingual talker discrimination

Mirjam Wester[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK

mwester@inf.ed.ac.uk

Abstract

This paper describes a talker discrimination experiment in which native English listeners were presented with two sentences spoken by bilingual talkers (English/German and English/Finnish) and were asked to judge whether they thought the sentences were spoken by the same person or not. Equal amounts of cross-lingual and matched-language trials were presented. The experiments showed that listeners are able to complete this task well, they can discriminate between talkers significantly better than chance. However, listeners are significantly less accurate on cross-lingual talker trials than on matched-language pairs. No significant differences were found on this task between German and Finnish. Bias (B'') and Sensitivity (A') values are presented to analyse the listeners' behaviour in more detail. The results are promising for the evaluation of EMIME, a project covering speech-to-speech translation with speaker adaptation.

Index Terms: evaluation, talker discrimination, cross-lingual

1. Introduction

The motivation for this study 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? 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.

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 talker similarity across language boundaries when the stimuli consist of natural speech.

Winters et al. (2008) [1] carried out a study in which they investigated the extent to which language familiarity affects a listener's perception of the speaker-specific properties of speech by testing listeners' identification and discrimination of bilingual talkers across German and English. They showed that listeners can generalize knowledge of talkers' voices across these two phonologically similar languages. However, it is unknown whether this is also the case for languages that are less closely related.

To investigate cross-lingual talker discrimination we recorded a database of bilingual speech. The language pairs we chose to record are English/German and English/Finnish. English/German was selected to be able to compare our results to [1] and we included Finnish/English talkers as Finnish is one of the EMIME languages, and so is English. The other languages under consideration in EMIME are Japanese and Mandarin.

The questions we want to answer are: how well can listeners discriminate between bilingual talkers across languages? Also, how much more accurate are listener's judgements of speaker similarity in the same language versus across two languages? Finally, do native English listeners perform better on German (which is more closely related to English) than on Finnish (which is less closely related to English)?

2. Discrimination experiment design

An important factor in talker identification or discrimination is talker familiarity. Whether or not a listener is familiar with a talker will influence how well they can recognise or identify them, as well as how well they can discriminate between them and other talkers [2, 3]. Of course unfamiliar voices can become familiar voices with training. In [4] talker-specific learning in speech perception was investigated. They found that listeners' familiarity with talkers facilitated the speech intelligibility.

Despite these findings, we decided to use untrained listeners, but to present them with sentence length stimuli rather that single word stimuli. Nygard and Pisoni (1998) [4] showed that learning is faster when using sentences rather than words and that it is much easier to identify talkers' voices from sentences than from isolated words. It can be expected that a sentence is long enough for some talker learning to occur. Therefore, using sentence length stimuli should provide the listeners with sufficient speaker-specific information about a talker to make an informed decision. Furthermore, in EMIME, the more likely scenario is that interlocutors are not familiar with each other.

2.1. Materials

German and Finnish were selected as the talkers' native languages (L1) for these experiments. German is an Indo-European language and Finnish is a part of the Finno-Ugric group of languages. English, also an Indo-European language, is the talkers' second language (L2).

A database of 14 German/English and 14 Finnish/English talkers (seven male / seven female per language) was collected (talkers were 20-30 years of age). Each talker read a set of 125 news sentences in both their native language and English. Of these 28 talkers, 20 were chosen to be present in the discrimination experiment presented in this paper. They were selected on the basis of an accent rating experiment in which native English listeners were asked to rate the degree of foreign accent for each talker on a scale from 0 ("no foreign accent") to 6 ("strong foreign accent"). For each language/gender category the five talkers with the least degree of foreign accent were selected. The reason for this is that we expect that the more native-like the bilingual talkers are in English the more difficult listeners

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

will find it to distinguish between them across languages. Per language, forty news sentences ranging in length from 7 to 10 words were selected for the talker discrimination experiment.

2.2. Design

There are four test conditions: German female, German male, Finnish female and Finnish male, each consisting of 160 trials (i.e. 320 sentences in total). 80 news sentences were used per test condition, 40 English and 40 German (or Finnish). Each sentence occurred four times – twice in a same-talker trial, twice in a different-talker trial. Each talker was presented in combination with every other talker twice and counterbalanced for order. We also ensured there were equal amounts of mixed-language and matched-language trials. Table 1 shows the number of trials for each language pair.

Table 1: *Number of trials per language pair.*

	Language pair			
test condition	matched		mixed	
German (F/M)	Eng-Eng	Ger-Ger	Eng-Ger	Ger-Eng
Finnish (F/M)	Eng-Eng	Fin-Fin	Eng-Fin	Fin-Eng
same	20	20	20	20
different	20	20	20	20

2.3. Listener Task

Forty native English listeners with no known hearing, speech and language problems, 20-30 years of age, were recruited at the University of Edinburgh. Each listener was given one of the test conditions (160 trials) to complete. This took between 35 and 45 minutes. Listeners were asked to judge whether the two sentences in each pair were spoken by the same talker or by two different talkers. In addition to giving same/different judgements they were asked to indicate on a 3-point scale how sure they were of their judgement. Subjects were paid for their participation.

3. Results

Each test condition was judged by 10 listeners. Per listener data were pooled for each test condition. Figure 1 shows a boxplot of percent correct for the four test conditions. In all boxplots in this paper, the median is indicated by a solid bar across a box which shows the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented by circles.

An analysis of variance (ANOVA) was conducted with test condition (German female, German male, Finnish female, Finnish male) as the between-test factor. The ANOVA shows there is no significant main effect of test condition $[F(3,36) = 0.53, p = 0.664]$. The listeners behave in a similar fashion for the different languages and genders.

Figure 2 is a boxplot showing percent correct for the various language conditions, the four test conditions have been combined here. A further ANOVA was conducted on the percent correct results with language pair condition as the within-test factor. The ANOVA shows there is a significant main effect of language pair $[F(7,192) = 8.04, p < 0.0001]$. Tukey HSD (Honestly Significant Difference) multiple comparisons of means with 95% family-wise confidence level were conducted to analyze the effect of language pair in more detail. The Tukey HSD test revealed that talker pairs are incorrectly classified significantly more often in mixed-language conditions than they are in matched-language conditions. Table 2 shows the mean percent correct for each of the language pairs, per test condition.

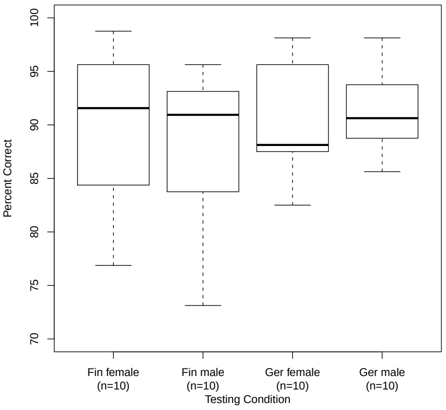

Figure 1: Percent correct judgements for the four test conditions.

Table 2: *Mean percent correct for each language pair, per test condition.*

	Language pair			
test condition	Eng-Eng	Eng-L1	L1-Eng	L1-L1
German female	97.5	85.8	85.0	93.8
German male	94.8	87.5	88.8	94.0
Finnish female	95.2	89.5	85.5	92.0
Finnish male	93.3	85.0	82.3	91.8
overall	95.2	87.0	85.4	92.9

The same/different responses were converted into nonparametric measures of sensitivity (A') and Griers' bias (B'') [5]. Both these measures are based on the proportion of "hits" and "false alarms". Hits in this context are when a listener judges a same-talker trial as same, and a false alarm is a same response to a different-talker trial. Sensitivity (A') is a measure of how sensitive a listener is to the same/different talker distinction. A' typically ranges from 0.5 which indicates that the trials cannot be distinguished from each other to 1.0 which corresponds to perfect performance. Griers Bias (B'') is a measure of the listeners' bias toward one response or the other. B'' ranges from -1.0 (extreme bias in favor of same) to 1.0 (extreme bias in favor of different). A B'' value of 0 indicates no bias in either direction.

Bias and sensitivity have been calculated per test condition. ANOVAs with language pair condition as the within-test factor and test condition as the between-test factor were conducted for the sensitivity (A') and bias (B'') measures.

Table 3 shows mean A' values, and Figure 3 shows a A' boxplot of listeners responses with test conditions pooled. Sensitivity measures in all language pair conditions are signifi-

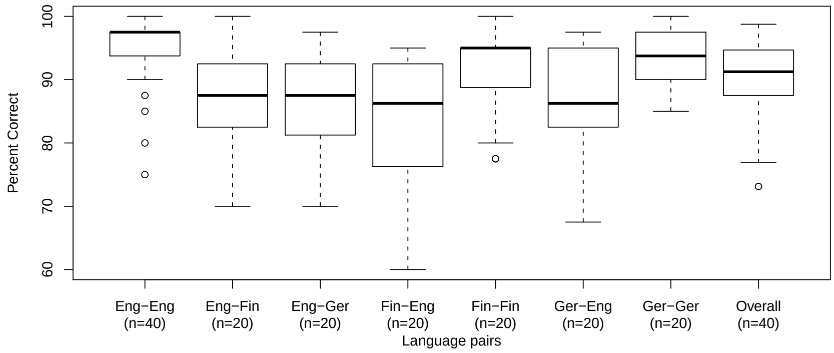

Figure 2: Percent correct judgements for each language pair condition, test conditions pooled.

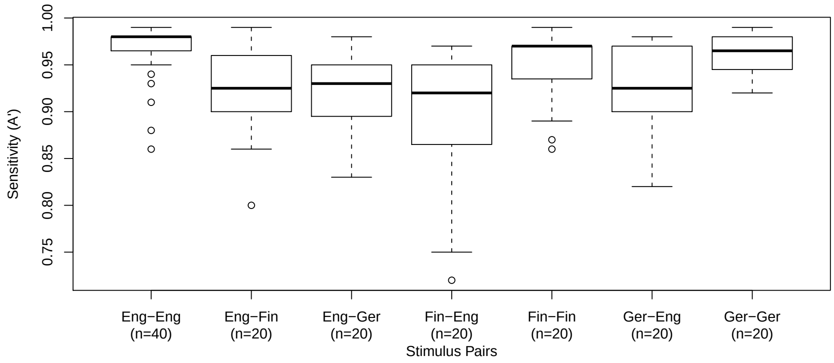

Figure 3: Sensitivity (A') for each language pair condition, test conditions pooled.

cantly above 0.5, chance performance. The sensitivity ANOVA showed a significant main effect of language pair $[F(6, 144) = 7.94, p < 0.001]$ but no significant main effect of test condition, nor was there a significant interaction between test condition and language pair. A Tukey HSD test revealed that listeners are significantly more sensitive to recognizing a talker as him/herself in the English-English condition than any of the mixed-language conditions ($p < 0.0001$). A significant difference was also found between German-German and Finnish-English ($p = 0.004$) and between Finish-Finnish and Finnish-English ($p = 0.005$). Basically, the listeners are more sensitive to matched-language trials than to mixed-language trials.

Table 4 shows mean B'' values, and Figure 4 shows B'' boxplots of listeners responses per test condition. The bias ANOVA yielded a significant main effect of language pair $[F(6, 144) = 2.76, p < 0.05]$ and of test condition $[F(3, 148) = 8.57, p < 0.0001]$ but no significant interaction between the two factors. Listeners behave significantly differently in the Finnish female test condition than in the other three test conditions. In the Finnish female test, the listeners have a negative bias which means they are more likely to judge a trial as same, whereas in the other three test conditions listeners are more likely to chose different in mixed-language trials and are closer to 0 – no bias – for matched-language pairs. A pairwise comparison of language pairs using the Tukey HSD test showed a significant difference between the German-English and German-German conditions. Listeners are more likely to judge trials as different in the German-English condition. All other pairwise comparisons were not significantly different.

Table 3: *Mean sensitivity (A') for each language pair, per test condition.*

	Language pair			
test condition	Eng-Eng	Eng-L1	L1-Eng	L1-L1
German female	0.980	0.916	0.912	0.960
German male	0.966	0.925	0.933	0.964
Finnish female	0.968	0.938	0.910	0.951
Finnish male	0.956	0.909	0.887	0.950

4. Discussion & Conclusions

This study shows that listeners perform well on a talker discrimination task. Discrimination accuracy is significantly higher

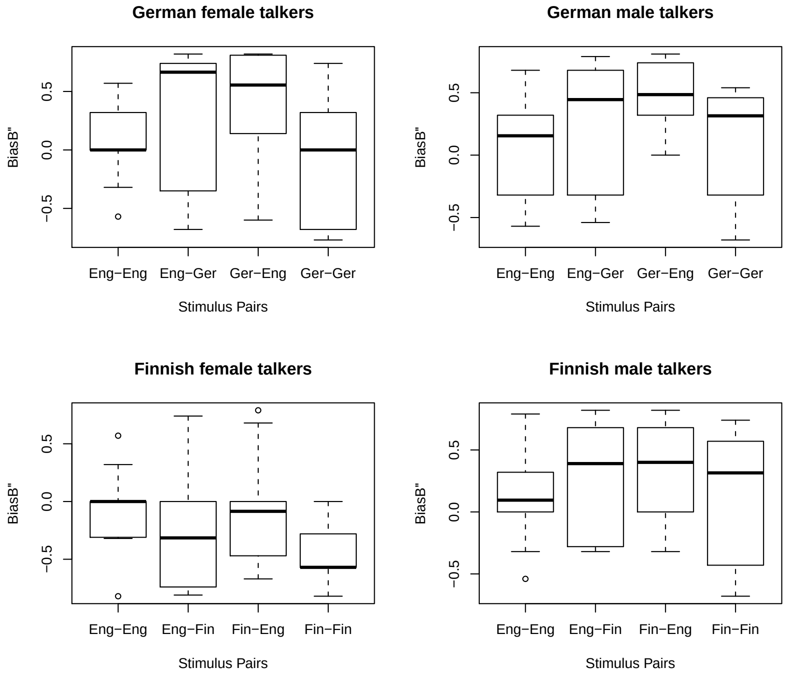

Figure 4: Bias (B'') for each language pair condition, per test condition.

Table 4: *Mean bias (B'') for each language pair, per test condition.*

	Language pair			
test condition	Eng-Eng	Eng-L1	L1-Eng	L1-L1
German female	0.064	0.333	0.415	-0.088
German male	0.110	0.283	0.488	0.084
Finnish female	-0.056	-0.234	-0.069	-0.452
Finnish male	0.144	0.275	0.306	0.137

than chance. However, listeners perform significantly better in matched-language conditions than in mixed-language conditions, percent correct is significantly higher as are the sensitivity values. We expected the cross-lingual condition to be the most difficult for listeners and this is corroborated by the results. On average the cross-lingual trials are scored incorrectly 8-10% (absolute) more often than the matched-language trials. Of the trials that are incorrectly judged 70% are cross-lingual (443 of 634), and 60% are same trials which are judged as different.

There is no clear indication that Finnish talker discrimination is more difficult for English native listeners than German talker discrimination. In the matched-language condition results are 2% lower on Finnish than on English, however in the cross-lingual condition results are comparable. Our further work will investigate whether this is also the case for Mandarin and Japanese.

The bias values we found show similar trends to [1], except for the Finnish female test condition. Further investigation is needed to determine why listeners show a different bias in this case. It may be due to the particular selection of talkers. Multidimensional scaling in further work will hopefully shed light on this. The confidence levels given by the listeners will then also be used.

The fact that listeners perform significantly better than chance in a talker discrimination task is a positive outcome for EMIME as it means the task we are ultimately looking at – whether or not a talker in L2 sounds similar to the original talker in L1 – is an achievable one. Further research will focus on investigating whether or not these findings remain the same for synthetic speech.

5. 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 for running the perception experiment.

6. 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] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.

[3] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[4] L. Nygaard and D. Pisoni, "Talker-specific learning in speech perception," *Perception & Psychophysics*, vol. 60, no. 3, pp. 355–376, 1998.

[5] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behaviour Research Methods, Instruments & Computers*, vol. 31, no. 1, pp. 137–149, 1999.