Talker discrimination across languages

Mirjam Wester*

The Centre for Speech Technology Research, The University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
Tel:+44 (0) 131 650 4434, Fax: +44 (0) 131 650 6626

Abstract

This study investigated the extent to which listeners are able to discriminate between bilingual talkers in three language pairs - English-German, English-Finnish and English-Mandarin. Native English listeners were presented with two sentences spoken by bilingual talkers and were asked to judge whether they thought the sentences were spoken by the same person. Equal amounts of cross-language and matched-language trials were presented. The results show that native English listeners are able to carry out this task well; achieving percent correct levels at well above chance for all three language pairs. Previous research has shown this for English-German, this research shows listeners also extend this to Finnish and Mandarin, languages that are quite distinct from English from a genetic and phonetic similarity perspective. However, listeners are significantly less accurate on cross-language talker trials (English-foreign) than on matched-language trials (English-English and foreign-foreign). Understanding listeners' behaviour in cross-language talker discrimination using natural speech is the first step in developing principled evaluation techniques for synthesis systems in which the goal is for the synthesised voice to sound like the original speaker, for instance, in speech-to-speech translation systems, voice conversion and reconstruction.

Keywords: human speech perception, talker discrimination, cross-language

1. Introduction

The ability to recognise a person as an individual based on their voice is something most of us probably take for granted. However, if that same individual was speaking a different language – one we maybe didn't understand – would we still be able to recognise them as the same person? In most everyday situations this is not a naturally occurring scenario. But, with the advancement of speech technology, scenarios like this are becoming a reality and the question does arise. In the EMIME project[1], the goal was personalised speech-to-speech translation (S2ST) 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 (Wester et al., 2010). This objective raises a number of questions: What is the effect of the modelling techniques used? How to measure whether a person sounds like the same person across language boundaries? How does comparing natural and synthetic speech impact on this?

A survey of the literature in voice conversion –which can be seen as related to speaker adaptive cross-lingual speech synthesis– gives an impression of the types of evaluation that are most commonly used in the field. Abe et al. (1991) used bilingual data (Japanese/English) and measured similarity by calculating mutual information between speaker pairs. Mashimo et al. (2001) also used bilingual data (Japanese/English) and used the objective measure Mel Cepstral Distortion (MCD) to evaluate speaker individuality. In the S2ST project TC-STAR (Sündermann et al., 2006) data from monolingual speakers was used in a unit selection system. Evaluation was carried out using mean opinion scores (MOS) for similarity and quality. The work of Latorre and colleagues (Latorre et al., 2006) has a slightly different focus: multilingual synthesis, which is the ability to generate utterances in more than one language, or utterances of mixed language, from a single system. They also use MOS, for intelligibility, similarity and native accent.

A common technique, used in several of these studies, is to compare cross-lingual voice conversion to intra-lingual voice conversion. However, this does not directly measure how similar the speech sounds to that of the original speaker. Using mean opinion scores to evaluate similarity, although a widely-used technique, is not without problems: judging how similar utterances are on a scale from 1 to 5 may be too difficult for listeners, especially if the utterances are in different languages. The results in Liang et al. (2010) support this. Judgements of speaker similarity are also strongly correlated with the overall quality or

[☆] Preliminary reports of this work were presented at Interspeech 2010 (Wester, 2010a) and are due to be presented at Interspeech 2011 (Wester and Liang, 2011a).

[*] Corresponding author.
Email address: mwester@inf.ed.ac.uk (Mirjam Wester)

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

naturalness of the synthetic speech: listeners are probably unlikely to rate an utterance as sounding like the target speaker if the quality is poor.

In summary, the methods commonly employed to evaluate talker similarity for voice conversion are no more sophisticated than those already used to evaluate text-to-speech (TTS). Whilst listening tests based on pairwise comparisons or MOS ratings are simple to administer and analyse statistically, they offer no guarantee that what is being evaluated really is talker similarity, independent of other factors such as quality or naturalness.

There are issues associated with comparing synthetic to natural speech, for instance, synthetic speech is less intelligible than natural speech, it requires more cognitive resources, and it is more difficult to comprehend (Winters and Pisoni, 2005). However, there is a more fundamental question which has not been addressed in the voice conversion and speaker adaptation TTS fields which is: to what extent are listeners able to judge talker similarity across language boundaries? This study focuses on this more fundamental question and investigates how well listeners judge talker similarity across language boundaries using stimuli that consist of natural speech.

What are listeners doing when they recognise a talker? To start with there is a distinction that can be made between *what* somebody says and *how* they say it. The *what* is covered by the linguistic properties of speech, that is the message that the speaker is trying to convey, and the *how* is covered by the characteristics of the talker (age, gender, emotional state, health etc), i.e. the non-linguistic information or the indexical properties. One of the main questions that has been addressed in previous studies is whether or not there is perceptual integration of these indexical and linguistic properties or if they are independently processed (see e.g., Nygaard, 2005; Winters et al., 2008, for reviews).

Nygaard (2005) gives a comprehensive overview of the relationship between linguistic and non-linguistic information in spoken language processing. She argues, based on the available evidence, that linguistic and non-linguistic information are integrally related components of the same acoustic speech signal and consequently the speech perceptual process.

Neuroscientific evidence supporting the integration of the linguistic and non-linguistic information is given in Perrachione et al. (2009). Listeners without any familiarity of a particular foreign language appear significantly impaired in achieving native-like accuracy at identifying voices speaking that language, even after substantial training (Perrachione and Wong, 2007). Furthermore, Perrachione and Wong (2007) found that although English subjects improved at a task of Mandarin speaker identification they never came to perform as well as native speakers.

Winters et al. (2008) 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 generalise 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. Winters et al. (2008) concluded that listeners apparently process indexical information in a language-dependent fashion when they hear a language that they know; otherwise, they perform indexical tasks by more heavily relying on language-independent information in the signal.

An important factor in speaker 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 (Kreiman and Papcun, 1991; Van Lancker and Kreiman, 1987). Of course unfamiliar voices can become familiar voices with training. In Nygaard and Pisoni (1998) talker-specific learning in speech perception was investigated. They found that listeners' familiarity with talkers facilitated speech intelligibility and that listeners learned talker identity faster from sentences than from single words.

In addition to talker familiarity a listeners' familiarity with the languages under consideration is also of interest. Goggin et al. (1991) investigated talker identification performance in a foreign versus native language. Native English listeners identified bilingual talkers speaking either English or German. Goggin et al. (1991) found that listeners are better at this task when the talkers are using the listeners' native language than when speaking a foreign language. Similar findings have been reported in Philippon et al. (2007). There it was shown that ear-witnesses are more accurate at recognising voices speaking their native language than an unfamiliar language.

Stockmal et al. (2000) investigated whether listeners are able to separate talker voice from language characteristics, and found that they are able to make same-language/different-language discrimination judgements at better than chance levels. However, in Stockmal et al. (2004) when asked to focus on voice quality to judge voice similarity in a foreign language, monolingual listeners were not able to ignore language characteristics.

In this study, native English listeners carried out talker discrimination experiments which measure how well they are able to discriminate between bilingual talkers[2] speaking English and German, English and Finnish or English and Mandarin. The English-German language pair was selected to enable comparisons between our results and those reported in Winters et al. (2008). Finnish and Mandarin were selected in addition to German because they are more distant to English than German from a classical language classification point of view (Lewis, 2009), but also from a phonetic similarity point of view (Bradlow et al., 2010).

[2] A bilingual talker –in this context– simply refers to a person who has the ability to speak and read the two languages under investigation.

The questions we want to answer with this study are: (i) How well do listeners discriminate between bilingual talkers across languages? (ii) How much more accurate are listeners in discriminating between talkers within a language vs across two languages? and (iii) Are listeners able to carry out voice discrimination across different language pairs equally well? That is, do native English listeners perform equally well on an English-German talker discrimination task as they do on an English-Finnish or English-Mandarin talker discrimination task?

On the basis of the above discussed previous research, we expect that our native English listeners will not be as accurate at discriminating between voices speaking in one of the foreign languages as they are in English. Furthermore, we expect that mixed language trials pose a larger challenge to listeners than matched language trials. This is because without language familiarity the relevant phonetic information contributing to accurate talker identification is unavailable, and the listeners have to make do with indexical information, i.e., the listeners have less information at their disposal in a foreign language. Finally, we expect Mandarin and Finnish to be more difficult for listeners than German because these languages have less phonetic and phonological detail in common with English than German does.

A further important contribution of this work is the development of a framework which can be used to evaluate S2ST systems more accurately and concisely. Before we can interpret listeners' behaviour when processing synthetic speech, a baseline of listeners' behaviour on natural speech is needed to be able to compare to.

2. Method

To investigate listeners' ability to recognize talkers across languages we designed a talker discrimination task. The listeners were all monolingual native English listeners, the talkers all native talkers of a language other than English who were recorded producing sentences in their native language, and in English (their non-native language).

The EMIME S2ST scenario served as the guide for the choices made in the task design. In an S2ST system, the interlocutors most likely will be unfamiliar with each other, therefore we used untrained listeners. Furthermore, the task needed to be able to measure whether a synthetic speech sample in language X sounds like it could have been produced by the same speaker who produced the input natural speech in language Y. A discrimination task seemed most appropriate to fullfil this requirement.

2.1. Talkers and Materials

A database of bilingual speech was recorded to investigate cross-lingual talker discrimination. The language pairs we chose to record are English-German, English-Finnish and English-Mandarin. The talkers' native languages (L1) were either German, Finnish or Mandarin. German is an Indo-European language, Finnish is a member of the Uralic group of languages and Mandarin Chinese is part of the Sino-Tibetan language family group (Lewis, 2009). English, also an Indo-European language, is the talkers' second language (L2).

Databases of 14 English-German, 14 English-Finnish and 14 English-Mandarin talkers (seven male/seven female per language) were 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 42 talkers, 30 were selected for the discrimination experiments presented in this paper. The selection was made 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 motivation for selecting talkers with the least degree of foreign accent was that we expected that the more native-like the bilingual talkers were in English the more of a challenge it would be to listeners to distinguish between them across languages. More details on the data collected and the accent ratings can be found in Wester (2010b); Wester and Liang (2011b).

Per language, forty news sentences ranging in length from 7 to 10 words were selected for the talker discrimination experiment. As mentioned above, Nygaard and Pisoni (1998) 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. The use of sentence-length stimuli provides not only a richer phonetic environment from which to compute talker identity compared to isolated words, but also sentence-level linguistic information absent from isolated words, such as patterns of intonation, stress and coarticulation (Goggin et al., 1991; Thompson, 1987; Perrachione and Wong, 2007). Note that although the task in the present experiment is discrimination rather than identification it is to be expected that the findings for discriminating between talkers will translate to identifying talkers. Discrimination can be seen as a precursor to identification. Using sentence length stimuli should provide the listeners with sufficient speaker-specific information about a talker to make an informed decision.

2.2. Listeners

Sixty native English listeners with no known hearing, speech or language problems, 20-30 years of age, were recruited at the University of Edinburgh. All listeners were monolingual native English speakers - they had no experience with Finnish, German or Mandarin aside from the inevitable accidental exposure to foreign languages. Also none of the subjects were familiar with any of the talkers.

2.3. Talker discrimination experiment design

In the talker discrimination test design, there were three different language pairs: English-German, English-Finnish and English-Mandarin. For every language pair, there were two sets of talkers, who differed in terms of gender: a set of five males and a set of five females. Thus, in total there were six tests (3 language pairs x 2 genders).

Each test consisted of 160 trials (i.e. 320 sentences in total). The trials were made up of 80 news sentences (40 English and 40 German, Finnish or Mandarin). Each sentence occurred four times – twice in same-talker trials, twice in different-talker trials. The two sentences within a trial were always different. Each of the five talkers 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.

In other words, listeners encountered the following types of trials in each test. Taking English-Mandarin as an example: in matched-language trials, sentences 1 and 2 were either both in English "Eng/Eng" or both in Mandarin "Man/Man". In mixed-language trials, when sentence 1 was in English then sentence 2 was in Mandarin, and vice versa: so "Eng/Man" and "Man/Eng". In same-talker trials, both sentences were produced by the same talker and in different-talker trials, sentence 1 was spoken by a different talker than sentence 2.

2.4. Task

Each listener was given one of the six tests to complete, i.e., one listener heard, for example, only the German males, while another listener only heard Mandarin females. Listeners were asked to decide whether the two sentences in each pair were spoken by the same talker or by two different talkers. The task took between 35 and 45 minutes to complete. Subjects were paid for their participation.

3. Analysis

Each of the six test conditions was judged by 10 listeners. The same/different responses were converted into nonparametric measures of sensitivity (A') and Grier's bias (B'') (Stanislaw and Todorov, 1999). 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. Grier's Bias (B'') is a measure of the listeners' bias toward one response or the other. B'' ranges from -1.0 (extreme bias in favour of same) to 1.0 (extreme bias in favour of different). A B'' value of 0 indicates no bias in either direction. Bias and sensitivity have been calculated per listener.

ANOVAs with stimulus pair type as the within-subject factor and test condition as the between-subject factor were conducted for the sensitivity (A') and bias (B'') measures. *Post-hoc* Tukey HSD (Honestly Significant Difference) multiple comparisons of means with 95% family-wise confidence levels were conducted to determine which factors were significantly different from each other.

In addition to the sensitivity and bias results, non-metric multi-dimensional scaling (MDS) was used to visualise the same/different responses given by the listeners. Sammon's non-linear mapping, a form of non-metric multidimensional scaling was used (Sammon, 1969). All of the MDS plots are 2-dimensional solutions computed using Sammon in R (R Development Core Team, 2010). This implementation chooses a two-dimensional configuration to minimise the stress, the sum of squared differences between the input distances and those of the configuration, weighted by the distances, the whole sum being divided by the sum of input distances to make the stress scale-free.

4. Results

4.1. Overall percent correct discrimination

To get a general overview of the results, the range of percent correct scores achieved by the subjects in the various test conditions is shown in Figure 1 and Table 1 shows the mean percent correct values for each of the test conditions.

Table 1: *Mean percent correct for each language pair, per test condition.*

	Language pair		
Test condition	Eng-Eng	Eng-L1	L1-L1
Finnish female	98.2	90.4	97
Finnish male	93	85.4	85.4
German female	99	88.6	95
German male	93.5	85.6	93
Mandarin female	92.8	72.6	85.5
Mandarin male	94	84	94

4.2. Sensitivity

Figure 2 shows an A' boxplot of listeners' responses per test condition. An ANOVA with stimulus pair type (Eng/Eng, L1/L1, Eng/L1) as the within-subject factor and test condition as the between-subject factor was conducted. The sensitivity ANOVA showed a significant main effect of stimulus pair type [$F(6, 162) = 11.36, p < 0.0001$] and of test condition [$F(5, 162) = 9.53, p < 0.0001$]. There was, however, no significant interaction between the two [$F(6, 162) = 1.48, p = 0.189$].

Tukey HSD tests were conducted to look at the differences between the stimulus pair types and between the various test conditions. It revealed that listeners are significantly more sensitive to matched-language trials than to mixed-language trials. However, no significant differences

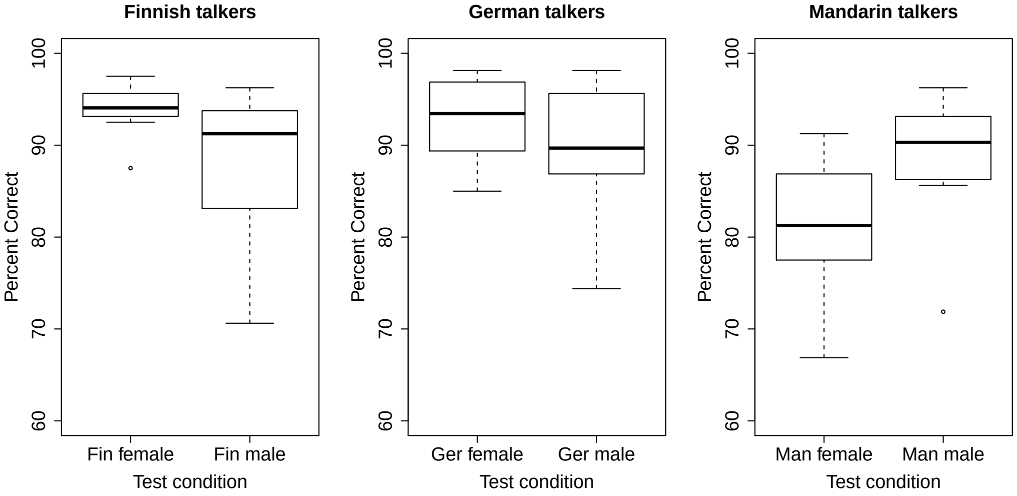


Figure 1: Percent correct discrimination for Finnish, German and Mandarin female and male test conditions.

in sensitivity were found between the various matched-language trials, which suggests that listeners discriminate as well between talkers speaking English as between talkers speaking German, Finnish or Mandarin. The Tukey HSD test also showed that listeners are significantly less sensitive to Mandarin female trials than to German and Finnish trials.

4.3. Bias

Figure 3 shows B'' boxplots of listeners' responses for each of the six test conditions. Again, an ANOVA with stimulus pair type (Eng/Eng, L1/L1, Eng/L1)) as the within-subject factor and test condition as the between-subject factor was conducted. The ANOVA for bias showed no significant effect of test condition $[F(5, 174) = 1.47, p = 0.202]$ nor of language pair $[F(6, 173) = 2.01, p = 0.066]$.

4.4. MDS

Figure 4 shows MDS plots of the listeners' same/different responses. The proximity between a talker's L1 and English data points indicates how well listeners classified talkers as themselves across the two languages. If the distance between a talkers' L1 and English points in the MDS space is small and doesn't overlap with another talkers' space, it indicates that the talker was recognised as the same person across languages by the listeners. One could also say that their indexical properties are perceivable across language boundaries. A large distance between a talker's L1 and English data points indicates they are difficult to recognise as one person. The MDS plot also shows which talkers are most confusable, as their data points are close together. Ellipses have been added to the plots to show L1-English proximity for each of the talkers. Note, however, that it is not clear from this analysis what the acoustic correlates of the dimensions are. Sammon (R Development Core Team, 2010) was used to obtain the 2-dimensional MDS solutions. The final stress levels are also indicated in Figure 4.

The MDS-plot for Mandarin females shows the largest degree of overlap between the different talkers. The overlap indicates that talkers 1 and 4 are very difficult for listeners to tell apart. This illustrates why the mean percent correct for the mixed-language trials for Mandarin females is only 72.6% (see Table 1). Contrast this with, for example, the Finnish females where there is a clear separation between all of the different talkers and the mean percent correct for mixed-language trials is 90.4%. For German female talkers, listeners achieve 88.6% and although some of the ellipses are closer together than for the Finnish females there is not the overlap seen for the Mandarin females or some of the male talker groups.

The MDS plots for the male talkers also give a good visual interpretation of the mean percent correct values achieved by the listeners (see Table 1). Mean percent correct for the Mandarin male talker group is 84.0%, for the Finnish male talker group it is 85.4% and for the German male talker group it is 85.6%. In the plots, the largest degree of overlap can be seen for Mandarin males, followed by Finnish males, with the least degree of overlap found for the German male talker group.

5. Discussion

In our experiments, native English listeners carried out talker discrimination tasks which measured how well they are able to discriminate between bilingual talkers. The results showed that listeners are able to perform this talker discrimination task well. Furthermore, listeners are significantly more sensitive to matched language trials than to mixed language trials. This means that listeners are better at distinguishing between talkers speaking the same language than when they are speaking different languages.

Looking only at the matched language trials we find that there are no significant differences between listeners'

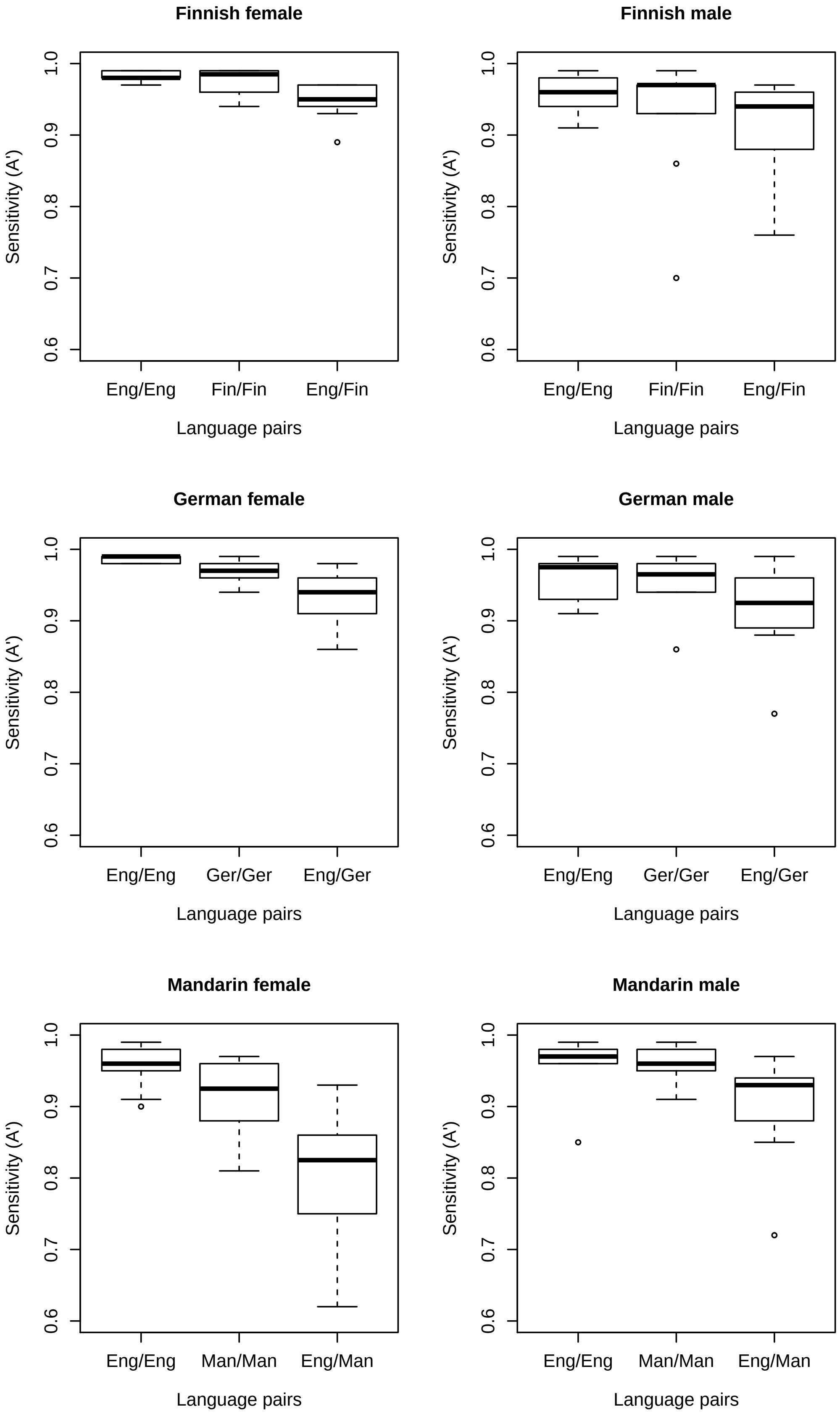


Figure 2: Sensitivity (A') values per test condition for each language pair.

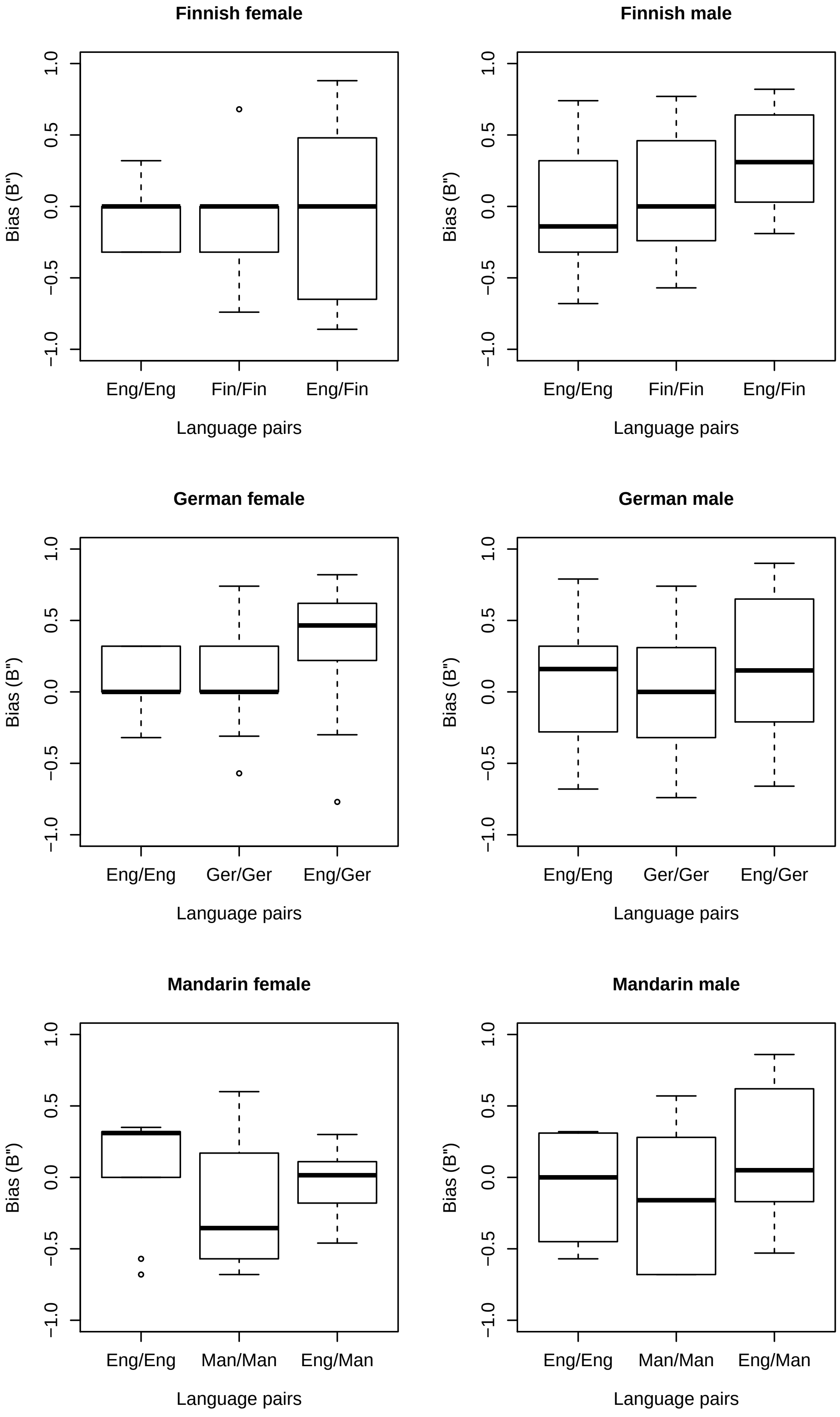
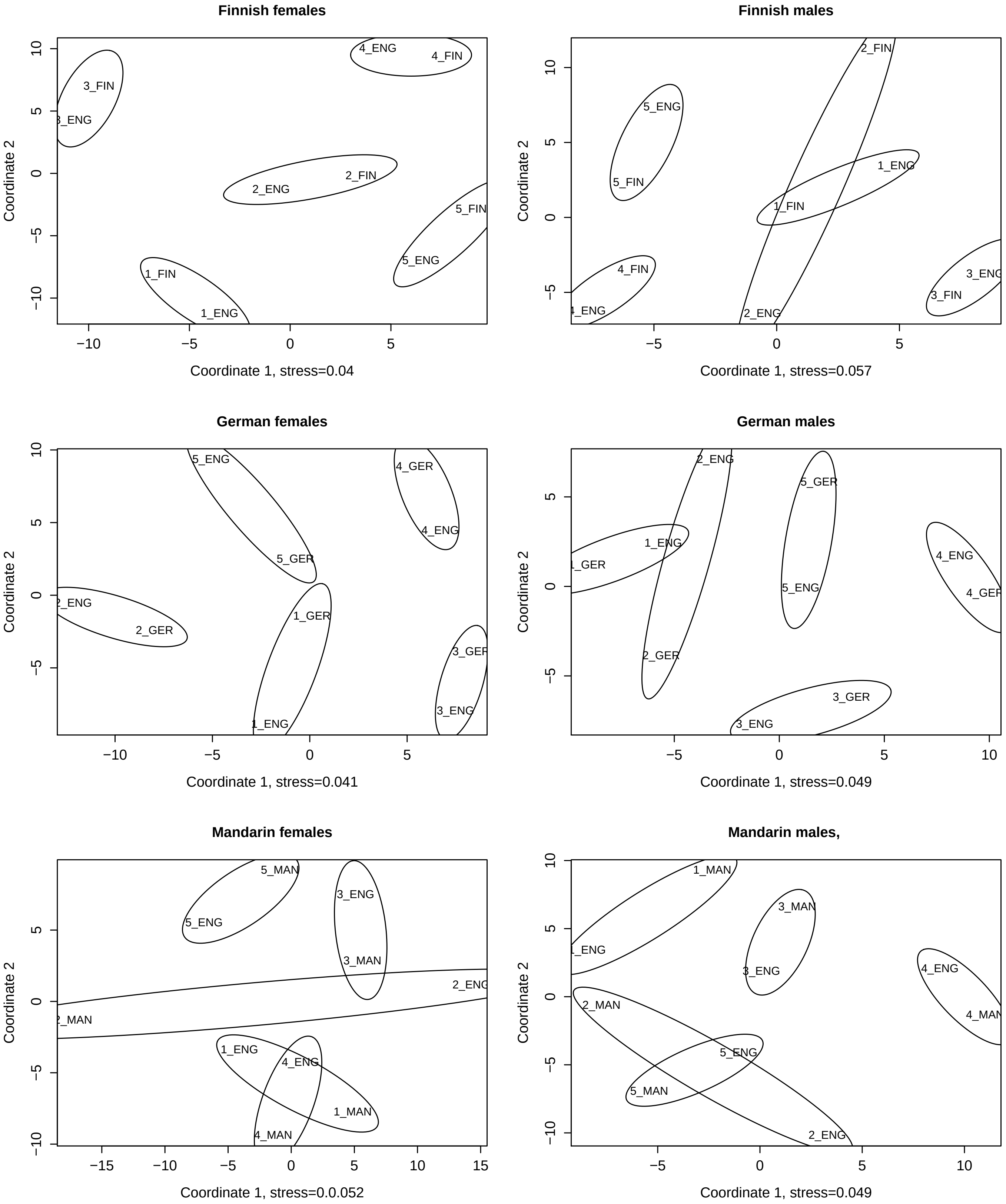


Figure 3: Bias (B'') values per test condition for each language pair.

]



Finnish females

Finnish males

German females

German males

Mandarin females

Mandarin males,

Figure 4: MDS per test condition.

performance on English-English trials and on matched foreign language trials. Although in all cases we observe that native English listeners achieve higher accuracy values for talker discrimination in English than in the foreign languages, the differences are relatively small. This finding is similar to Winters et al. (2008) in which they report that subjects perform better on English-English than on German-German trials. In their study, however, the differences are significant.

In mixed language trials, i.e., talker discrimination across language boundaries, listeners show a significantly degraded performance for all the female talker groups and for the male Mandarin talker group. This is line with expectations. In the foreign matched language condition, listeners are able to only focus on information in the signal relevant to talker discrimination, they are not distracted by linguistic properties in the signal. In the native matched language condition, listeners make use of all the information available to them - indexical and linguistic. In the cross-language scenario, listeners have access to both indexical and linguistic information for the language they are familiar with but only indexical information for the foreign language. This mismatch in available information will add to the cognitive load and thus the difficulty of this condition.

Our bias measures did not show any significant effects of test condition nor of language condition. It looks like there is a slight positive bias for mixed language trials, which means listeners are more likely to say the trial contains different talkers, while for matched language trials overall the B'' value is near to 0 which indicates listeners do not have a bias in either direction. However, note that the variance in bias is rather large. This large spread in bias values is an indication that different listeners show different patterns of bias.

The final question to be addressed is whether listeners are able to carry out voice discrimination of talkers across different language pairs. At a glance, the Mandarin female talker group gives the impression that Mandarin is more difficult for listeners than for instance, Finnish or German. However, if we consider the information in the MDS visualisation, which shows that for the Mandarin female talker group two of the five talkers are highly confusable, we can hypothesise that it is the set of Mandarin female talkers rather than the language which is posing the challenge for listeners. Add to this that listeners' behaviour on the Mandarin male talker set does not differ from their behaviour on the German and Finnish talker sets then, taken together, there is not enough evidence to support the view that talker discrimination across languages A and B is more difficult than across languages A and C. Rather, we can conclude that our results support the findings in Winters et al. (2008). Listeners are able to discriminate between talkers' voices across English and German and, in addition to that, listeners are also able to extend this to Finnish and Mandarin, languages that are quite distinct from English from a genetic and a phonetic similarity perspective.

The MDS plots were included here simply to illustrate the discrimination results more clearly. In future research, a more in-depth acoustic analysis will be carried out as for example in Tsuzaki et al. (2011). In their study, an ABX discrimination task was carried out using the recordings of bilingual talkers. MDS was carried out on the discrimination results and a number of different auditory features were extracted for each talker. Using regression analysis, Tsuzaki et al. (2011) found that the spectral centroid and loudness were the auditory features which contributed most to the perceptual dimensions in the MDS. This suggests that these are features which listeners employ to discriminate between talkers. Also, further investigation into the processes underlying a listener's perception when performing talker discrimination may include more sensitive measures such as reaction times. Response times may give more insight into talker pairs that are more confusable with each other and could show more subtle differences between different languages.

Our experiments show that talker discrimination across languages is a viable task for listeners. The talker discrimination results in combination with the MDS visualisation give a good picture of how listeners behave in what could be seen as a S2ST evaluation task consisting only of natural speech. These findings give a good basis to further explore the behaviour of listeners in S2ST system evaluations. Preliminary experiments investigating various aspects of listeners' behaviour on synthetic speech in a S2ST context can be found in Wester and Karhila (2011); Karhila and Wester (2011); Wester and Liang (2011a).

Acknowledgements

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

References

Abe, M., Shikano, K., Kuwabara, H., July 1991. Statistical analysis of bilingual speaker's speech of cross-language voice conversion. J. Acoust. Soc. Am. 90 (1), 76–82.

Bradlow, A. R., Clopper, C., Smiljanic, R., Walter, M., 2010. A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. Speech Communication 52, 930–942.

Goggin, J., Thompson, C., Strube, G., Simental, L., 1991. The role of language familiarity in voice identification. Memory and Cognition 19 (5), 448–458.

Karhila, R., Wester, M., 2011. Rapid adaptation of foreign-accented HMM-based speech synthesis. In: Proceedings Interspeech '11. Florence, Italy.

Kreiman, J., Papcun, G., 1991. Comparing discrimination and recognition of unfamiliar voices. Speech Communication 10 (3), 265–275.

Latorre, J., Iwano, K., Furui, S., 2006. New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. Speech Communication 48, 1227–1242.

Lewis, M. P. (Ed.), 2009. Ethnologue: Languages of the World Sixteenth Edition. SIL International.

Liang, H., Dines, J., Saheer, L., 2010. A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In: Proceedings ICASSP '10.

Mashimo, M., Toda, T., Shikano, K., Campbell, N., 2001. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In: Proceedings Eurospeech '01.

Nygaard, L., 2005. Perceptual integration of linguistic and nonlinguistic properties of speech. In: The Handbook of Speech Perception. Blackwell Publishing, pp. 390–413.

Nygaard, L., Pisoni, D., 1998. Talker-specific learning in speech perception. Perception & Psychophysics 60 (3), 355–376.

Perrachione, T., Pierrehumbert, J., Wong, P., 2009. Differential neural contributions to native- and foreign-language talker identification. J. of Experimental Psychology: Human Perception and Performance 35 (6), 1950–1960.

Perrachione, T., Wong, P., 2007. Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. Neuropsychologia 45, 1899–1910.

Philippon, A. C., Cherryman, J., Bull, R., Vrij, A., 2007. Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. Appl. Cognit. Psychol. 21, 539–550.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Sammon, J., 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C-18 (5), 401–409.

Stanislaw, H., Todorov, N., 1999. Calculation of signal detection theory measures. Behaviour Research Methods, Instruments & Computers 31 (1), 137–149.

Stockmal, V., Bond, Z., Moates, D., 2004. Judging voice similarity in unknown languages. In: Proceedings of the 17th Congress of Linguists. Prague.

Stockmal, V., Moates, D., Bond, Z., 2000. Same talker, different language. Applied Psycholinguistics 21, 383–393.

Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J., 2006. Text-independent cross-language voice conversion. In: Proceedings Interspeech '06.

Thompson, C., 1987. A language effect in voice identification. Applied Cognitive Psychology 1, 121–131.

Tsuzaki, M., Tokuda, K., Kawai, H., Ni, J., 2011. Estimation of perceptual spaces for speaker identities based on the cross-lingual discrimination task. In: Proceedings Interspeech '11. Florence, Italy.

Van Lancker, D., Kreiman, J., 1987. Voice discrimination and recognition are separate abilities. Neuropsychologia 25 (5), 829–834.

Wester, M., 2010a. Cross-lingual talker discrimination. In: Proceedings Interspeech '10.

Wester, M., 2010b. The EMIME Bilingual Database. Tech. Rep. EDI-INF-RR-1388, The University of Edinburgh.

Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y.-J., Saheer, L., King, S., Oura, K., Garner, P., Byrne, W., Guan, Y., Hirsimäki, T., Karhila, R., Kurimo, M., Shannon, M., Shiota, S., Tian, J., Tokuda, K., Yamagishi, J., 2010. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In: Proceedings SSW7.

Wester, M., Karhila, R., 2011. Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In: Proceedings of ICASSP.

Wester, M., Liang, H., 2011a. Cross-lingual speaker discrimination using natural and synthetic speech. In: Proceedings Interspeech '11. Florence, Italy.

Wester, M., Liang, H., 2011b. The EMIME Mandarin Bilingual Database. Tech. Rep. EDI-INF-RR-1396, University of Edinburgh.

Winters, S., Levi, S., Pisoni, D., 2008. Identification and discrimination of bilingual talkers across languages. J. Acoust. Soc. Am. 123, 4524.

Winters, S., Pisoni, D., 2005. Speech synthesis, perception and comprehension of. In: Brown, K. (Ed.), Encyclopedia of Language and Linguistics. Elsevier, pp. 31–49.