IDIAP RESEARCH REPORT

CROSS-LINGUAL SPEAKER DISCRIMINATION USING NATURAL AND SYNTHETIC SPEECH

Mirjam Wester Hui Liang

Idiap-RR-18-2011

JUNE 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Cross-Lingual Speaker Discrimination Using Natural and Synthetic Speech

Mirjam Wester[1], *Hui Liang*[2,3]

[1] Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2] Idiap Research Institute, Martigny, Switzerland
[3] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
mwester@inf.ed.ac.uk, hui.liang@idiap.ch

Contents

Abstract

This paper describes speaker discrimination experiments in which native English listeners were presented with natural speech stimuli in English and Mandarin, synthetic speech stimuli in English and Mandarin, or natural Mandarin speech and synthetic English speech stimuli. In each experiment, listeners were asked to judge whether the sentences in a pair were spoken by the same person or not. We found that the results of Mandarin/English speaker discrimination were very similar to those found in previous work on German/English and Finnish/English speaker discrimination. We conclude from this and previous work that listeners are able to discriminate between speakers across languages *or* across speech types, but the combination of these two factors leads to a speaker discrimination task that is too difficult for listeners to perform successfully, given the fact that the quality of across-language speaker adapted speech synthesis at present still needs to be improved.

Index Terms: speaker discrimination, speaker adaptation, HMM-based speech synthesis

1. Introduction

In the EMIME project, we are aiming for personalized 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 (http://www.emime.org). However, how do we measure whether our modeling attempts are successful or not? That is, how are we to measure whether or not a speaker sounds similar in two different languages? Does synthetic speech which has been adapted to sound like an original speaker actually sound like him/her?

In previous studies, we partially addressed these issues. [1] looked at across-language speaker discrimination (German/English and Finnish/English) using natural speech stimuli. The experiments in [1] showed that listeners were able to complete this task well, and could discriminate between speakers significantly better than chance. However, listeners performed significantly worse on across-language speaker trials than on matched-language trials.

Winters et al. [2] showed that listeners could generalize knowledge of speakers' voices across English and German, which are two phonologically similar languages. In [1] we looked at Finnish which is from the Uralic language family rather than Indo-European like English and German. The results in [1] showed there was no indication that Finnish speaker discrimination was more difficult for native English listeners than German speaker discrimination.

Listeners' ability to discriminate between speakers when comparing synthetic speech to natural speech within a single language (English) was investigated in [3]. It was found that listeners also completed this task well, with classification results significantly above chance. However, listeners performed significantly worse on mixed trials (synthetic vs natural) than on matched trials (synthetic-synthetic or natural-natural). Furthermore, the degradation of listeners' ability to discriminate between speakers was worse when comparing across different speech types (synthetic vs natural), than when comparing across different languages.

This paper investigated how well listeners were able to discriminate between speakers when they had to deal with stimulus pairs that crossed both language and speech type boundaries. We investigated whether previous findings for German and Finnish speaker discrimination also held true for a language from another language family: Mandarin Chinese from the Sino-Tibetan language family. Using speaker discrimination tests, we measured how well listeners were able to discriminate between speakers first in natural Mandarin and English, then in synthetic Mandarin and English, and finally in natural Mandarin and synthetic English.

2. Experimental Design

2.1. Speech Database

For our speaker discrimination experiments, we recorded a bilingual (Mandarin and English) speech database [4] at the University of Edinburgh [1]. It contains seven female and seven male speakers reading Mandarin and English prompts. For the experiments mentioned in this paper, five females and five males with the least degree of foreign accent in their English were selected from the 14 speakers. An accent rating task was used to decide the degree of foreign accent for each of the speakers [4].

2.2. Preparation of Stimuli

HMM-based speech synthesis enables the generation of unique synthetic voices by adapting an average voice model [5]. By using HMMs with explicit duration modelling and by adapting spectral, pitch and duration parameters using sentence-wide phonological and linguistic context information, it is possible to adapt speaking styles and phonetic features of synthetic speech [5, 6]. A foreign accent can be viewed as a certain type of speaking style and these techniques allow for adaptation of speaking rhythm, regular mispronunciation patterns and other types of features that are distinctive of foreign accents. The following subsections describe how we generated synthetic stimuli for our experiments. All the synthetic stimuli were speaker-adapted speech samples, in either Mandarin or English.

2.2.1. Average voice models to be adapted

We trained two average voice, single Gaussian-per-state synthesis model sets on the corpora Speecon (12.3 hours in Mandarin) and WSJ-SI84 (15.0 hours in English), respectively, in the HTS-2007 framework [7]. The HMM topology was five-state and left-to-right with no skip. Speech features were 39th-order STRAIGHT [8] mel-cepstra, $\log F_0$, 5-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz WAV files with a window shift of 5ms.

2.2.2. Within-language speaker adaptation

Speech data for within-language speaker adaptation was sourced from the bilingual (Mandarin and English) speech database [4]. The two average voices were adapted to each of the 10 selected speakers with 105 English and 60 Mandarin adaptation utterances (i.e. on average, 86060 English and 84715 Mandarin speech frames per speaker), respectively. The 45 utterance difference was due to the fact that Mandarin sentences were much longer than English ones. To ensure the amount of adaptation data for the two languages was comparable, we limited the number of Mandarin sentences used.

[1] Available for download at http://www.emime.org/participate/emime-bilingual-database

The adaptation procedure followed the supervised within-language case in [9], which used the CSMAPLR algorithm [6] for transform estimation. For stimulus synthesis, we used global variances calculated on the adaptation data, but duration models of the average voices in order to ensure the synthetic speech would have natural prosody and not be affected by foreign prosody present in the adaptation data.

2.2.3. Across-language speaker adaptation

In the context of across-language speaker adaptation, we adapted the English average voice to each of the 10 selected speakers using their 60 Mandarin adaptation utterances. The adaptation procedure followed the supervised across-language data-mapping case in [9] using the CSMAPLR algorithm [6]. We constructed a set of mapping rules between the two average voice model sets to ensure each Mandarin HMM state was linked to an English one, then associated Mandarin adaptation data with English HMM states via these mapping rules and finally performed "within-language" speaker adaptation on the English side by ignoring the language identity of the Mandarin adaptation data. As in Sec. 2.2.2, we used global variances calculated on the adaptation data and duration models of the English average voice for stimulus synthesis.

2.3. Evaluation – Listening Test Design

Four listening experiments (Exp. I-IV) were conducted. Each experiment consisted of two parts: a female and a male test conditions. There were five speakers in each test. We did not combine genders within any of the tests. 80 news sentences were used per test condition, 40 English and 40 Mandarin sentences which were selected from the bilingual database [4]. None of these sentence were used for speaker adaptation. Each test consisted of 160 trials (i.e., 320 utterances in total). Each sentence occurred four times – twice in same-speaker trials, twice in different-speaker trials. The two sentences within a trial were always different. Each of the five speakers was presented in combination with every other speaker 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. 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-speaker trials, both sentences were produced by the same speaker and in different-speaker trials, sentence 1 was spoken by a different speaker than sentence 2. The four listening tests included the following types of speech:

Exp. I – natural English and natural Mandarin

Exp. II – synthetic English and synthetic Mandarin (both *within-language* speaker adaptation)

Exp. III – synthetic English (*within-language* speaker adaptation) and natural Mandarin

Exp. IV – synthetic English (*across-language* speaker adaptation) and natural Mandarin

2.4. Listeners' Task

Eighty 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 to complete. This took between 35 and 45 minutes. The listeners were asked to judge if the two utterances in each pair were spoken by the same speaker or by two different speakers. In addition to giving same/different judgements, they were asked to indicate on a 3-point scale how sure they were of their judgements. Listeners 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 the results for the female and male test conditions. In all boxplots in this paper, a median is indicated by a solid bar across a box which shows quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented by circles.

An analysis of variance (ANOVA) with test condition (female, male) as the between-test factor showed there was a significant main effect of test condition $[F(1, 18) = 6.49, p = 0.02014]$. Therefore, female and male test conditions are presented separately in the following analyses.

Figure 2 shows boxplot results for all four experiments. The order of presentation of the mixed-language conditions – "Eng/Man" and "Man/Eng" – did not have a significant effect on percent correct, so they were combined. ANOVAs with

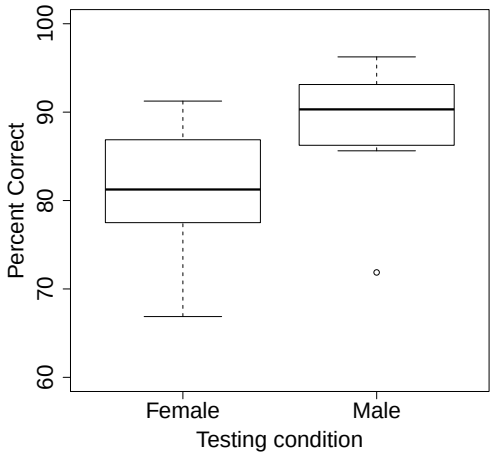

Figure 1: *Exp. I – Percent correct discrimination for the female and male test conditions, all natural speech.*

language pair (Eng/Eng, Man/Man and Eng/Man) as the within-test factor were conducted for all four experiments. In all cases, a significant main effect of language pair was found. Tukey HSD tests showed that listeners performed significantly worse on mixed-language trials than on matched-language trials. In Exp. IV, for both female and male test conditions there was also a significant difference between Man/Man and Eng/Eng. This was in contrast to the other experiments, in which no significant differences between matched-language trials had been found, irrespective of the speech being natural or synthetic.

Table 1 shows the results in terms of mean percent correct per language pair, for each of the four experiments. Differences in terms of percent correct between the various experiments are also given.

4. Discussion

It was shown in [1] that when comparing stimuli across languages (English/German and English/Finnish), listeners' performance dropped on average 10 percentage points, from 90-100% correct (matched-language) to 80-90% correct (mixed-language). Exp. I showed a similar picture. For the Mandarin male test set, listeners followed this pattern exactly. For the Mandarin female test set the results were about 10% lower.

Mandarin speaker discrimination did not seem to be more difficult for native English listeners than German or Finnish speaker discrimination when we looked at the male test condition. However, for the female Mandarin speakers we found significant differences between the results of listeners on female Mandarin speakers and the other female speaker sets, as well as between the female Mandarin speakers and the male German speakers. The most likely explanation would be that the set of five female Mandarin speakers is intrinsically more confusable than the other sets of speakers.

To illustrate this, Figure 3 shows non-metric multidimensional scaling (MDS) plots for the same/different scores given by the listeners for Mandarin male and female speakers. The plots are 2-dimensional projections of a 4-dimensional space. (stress = 0.02 for the male data, and 0.014 for the female data.)

The MDS plot can be interpreted as follows. The proximity between a speaker's English and Mandarin data points indicates how well listeners recognized speakers as themselves across the two languages. A large distance between a speaker's English and Mandarin data points indicates they are difficult to be recognized as one person. The MDS plot also shows which speakers are most confusable, as their data points are close together. Note, however, that it is not clear from this initial analysis what the acoustic correlates of the dimensions are.

In the female plot, the data points for speakers 1 and 4 totally overlap, meaning that listeners were not able to distinguish between these two speakers. Speaker 2's English and Mandarin data points are quite far removed from each other. Speaker 3's English and Mandarin data points merge but are quite close to speaker 5's data points. Three out of five speakers were clearly difficult for listeners. Compare this to the male plot in which speakers 2, 3, 4 and 5 all have Mandarin and English data points that are near each other, i.e., listeners were able to recognize these speakers well across the two languages. Only speaker 1 seems more difficult to identify across the languages and is more confusable with

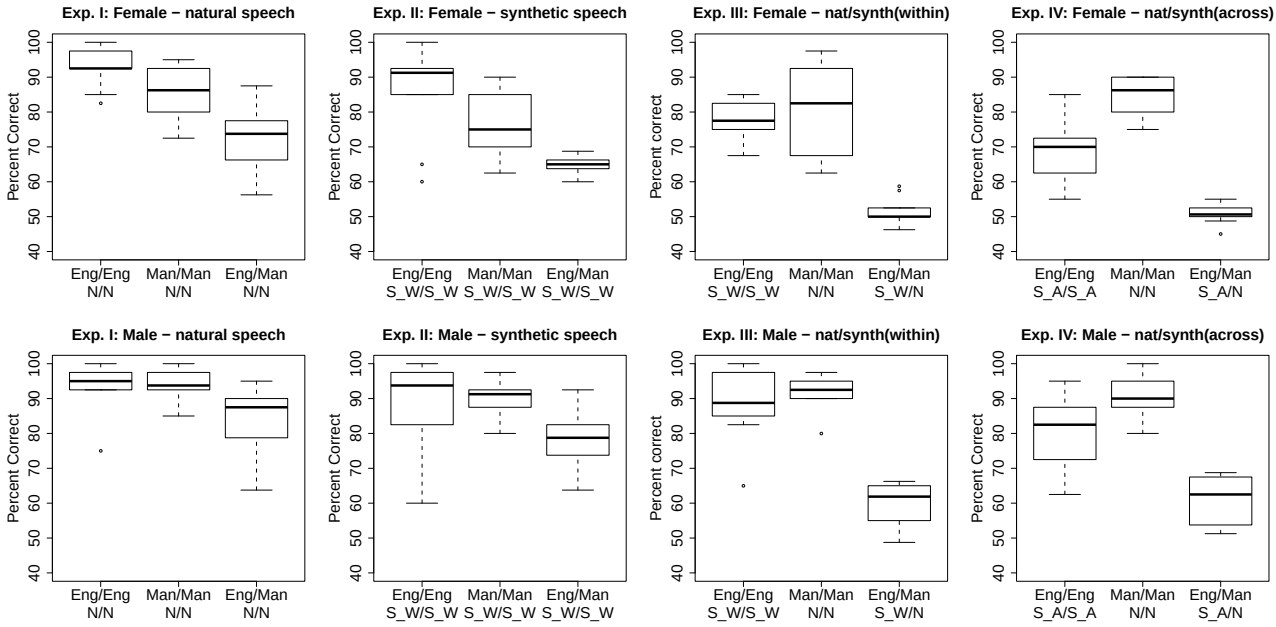

Figure 2: *Percent correct discrimination per language pair for male and female test conditions for the four different listening tests. N=Natural speech, S=Synthetic speech, _W=Within-language adaptation, _A=Across-language adaptation.*

speaker 3 in Mandarin and speaker 2 in English.

When going from Exp. I to Exp. II, i.e., from natural to synthetic speech, we observed small drops in listeners' performance of 7-9% in the female and 4-6% in the male test conditions. The synthetic speech created using within-language adaptation led to speaker identities that were recognized as individuals in the matched-language conditions. The results for synthetic speech are very similar to those found for natural speech.

In Exps. III and IV, the focus was on the mixed-language condition. Going from Exp. II to Exp. III, we saw a 13% degradation in listeners' performance for females and an 18% drop for males. When applying across-language speaker adaptation there was no further drop in performance in the mixed-language condition, but in this condition, for the female test set, listeners already performed at near chance levels. There was a drop in performance in the English matched-language condition of about 8% when going from within-language adaptation to cross-language adaptation.

5. Conclusions

Listeners are able to carry out speaker discrimination tasks well – deciding whether or not a speaker in one language sounds similar to the original speaker in another language is an achievable task. The current study has shown that native English listeners did not experience Mandarin as any more difficult than Finnish or German in such a speaker discrimination task.

[1] showed us listeners were well able to compare natural stimuli across languages (on average, 82-90% correct). The discrimination study in [3] showed that listeners were also reasonably able to discriminate speakers across speech types (synthetic vs natural) *within* a language (on average, 69-73% correct). The experiments in this paper show that when, in addition to comparing different speech types, listeners also had to contend with across-language trials, their ability to correctly discriminate between speakers suffered quite substantially (on average, 51-61% correct). To summarize, listeners are able to discriminate between speakers across languages *or* across speech types, but the combination of these two factors leads to a speaker discrimination task that is too difficult for listeners to perform successfully, given the fact that the quality of across-language speaker adapted speech synthesis at present still needs to be improved.

Our speaker discrimination set-up forms a good framework to measure to what extent listeners are able recognize a speaker as themselves across various conditions. It is more suited to measuring whether listeners perceive a speaker as himself/herself than a MOS-style rating task in which listeners are asked to judge speaker similarity [3]. Future research in personalized S2ST will need to concentrate on further improving a speaker's synthetic identity to achieve the goal of

Table 1: *Mean percent correct for each language pair, per test condition (Female or Male) and experiment.*

		Language pair		
M/F	Exp.	Eng/Eng	Man/Man	Eng/Man
F	I	92.8	85.5	72.6
	II	86.3	76.3	64.6
	III	77.3	81.0	51.5
	IV	69.3	84.5	50.6
(Diff)	*I – II*	*6.5*	*9.2*	*8.0*
	II – III	*9*	*-4.7*	*13.1*
	III – IV	*8.0*	*-3.5*	*0.9*
M	I	94.0	94.0	84.0
	II	89.3	89.8	78.1
	III	88.3	92.3	60.4
	IV	80.5	90.8	61.1
(Diff)	*I – II*	*4.7*	*4.2*	*5.9*
	II – III	*1.0*	*-2.5*	*17.7*
	III – IV	*7.8*	*1.5*	*-0.7*

sounding like the original speaker.

6. 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). Thanks to Vasilis Karaiskos for running the perception experiments.

7. References

[1] M. Wester, "Cross-lingual talker discrimination", in *Proc. of Interspeech*, Sep. 2010, pp. 1253–1256.

[2] S. Winters, S. Levi, and D. Pisoni, "Identification and discrimination of bilingual talkers across languages", *Journal of the Acoustical Society of America*, vol. 123, p. 4524, 2008.

[3] M. Wester and R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation", in *Proc. of ICASSP*, May 2011.

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

[5] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis", *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[8] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", *Speech Communication*, no. 27, pp. 187–207, 1999.

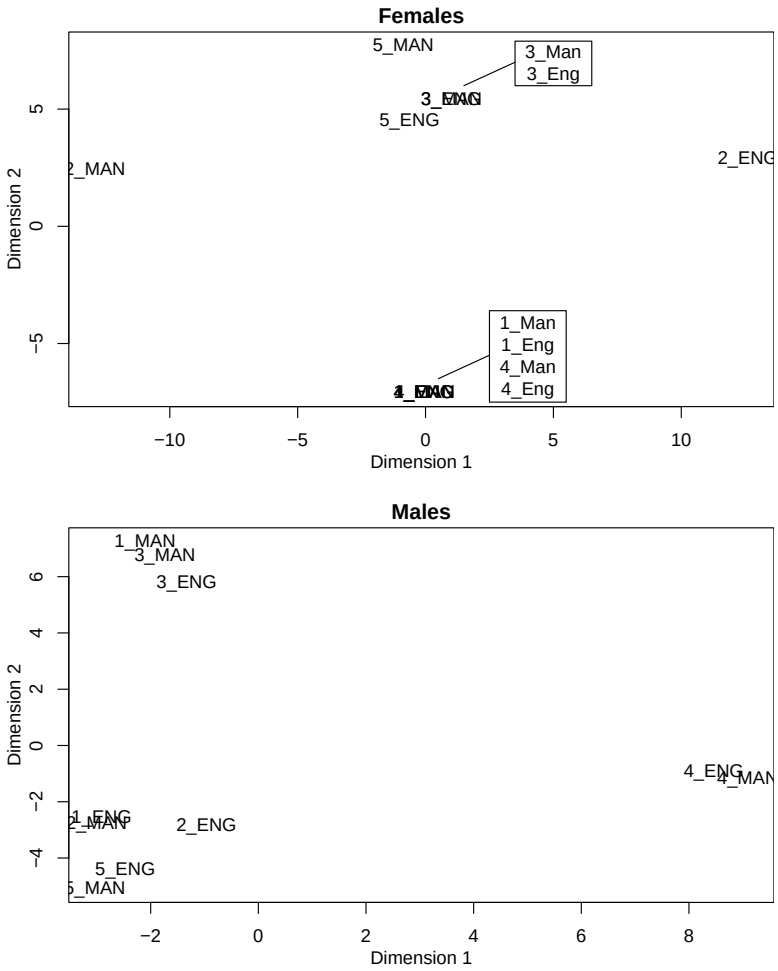

Figure 3: *MDS plots of female and male speakers' English and Mandarin data.*

[9] H. Liang, J. Dines, and L. Saheer, "A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis", in *Proc. of ICASSP*, Mar. 2010, pp. 4598–4601.