Work Package 5 : D5.5
Final report of formal internal evaluations and of real-time demonstrator evaluation

Effective Multilingual Interaction in Mobile Environments

Date of preparation	February 2011
Version number	1
Project full title	Effective Multilingual Interaction in Mobile Environments
Proposal acronym	EMIME
Funding scheme	STREP
Project co-ordinating person	Simon King Simon.King@ed.ac.uk
Deliverable co-ordinating person	Mirjam Wester mwester@inf.ed.ac.uk

Participant no.	Participant organisation name	Part. short name	Country
1 (Coordinator)	University of Edinburgh	UEDIN	UK
2	Institut Dalle Molle d'Intelligence Artificielle Perceptive	IDIAP	Switzerland
3	Aalto University	Aalto	Finland
4	Nagoya Institute of Technology	NIT	Japan
5	Nokia Corporation	Nokia	Finland
6	University of Cambridge	UCAM	UK

Contents

1 Introduction

The scope of this deliverable is to describe how we evaluated the EMIME research and real-time systems and the outcome of our evaluations.

In the EMIME 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. One of the main issues that follows from this statement and which needed to be addressed in our evaluation of the EMIME system is 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? Does the synthetic speech which has been adapted to sound like the original speaker actually sound like them? Furthermore, how should the synthetic voice of a person actually sound in a foreign language?

The content of this report includes a description of the main approach that we have developed for evaluating across-language speaker discrimination. In order to be able to describe the final evaluation of the research system a number of intermediate evaluation stages will first be discussed. The final section in this deliverable covers the qualitative evaluation of the real-time demonstrator.

1.1 List of papers in WP 5

This deliverable report is based on the following manuscripts submitted for publication in leading conferences in the field by the project partners. The content of the papers will be described in the following chapters and citations will be used to highlight appropriate details of the analysis and results.

1. Wester, "Cross-lingual talker discrimination" Interspeech 2010 [12]
2. Wester & Karhila "Speaker Similarity Evaluation of Foreign-accented speech Synthesis using HMM-based speaker adaptation" ICASSP 2011 [15]
3. Karhila & Wester "Rapid Adaptation of Foreign-accented HMM-based Speech Synthesis" submitted to Interspeech 2011 [4]
4. Wester & Liang "Cross-lingual Speaker Discrimination using Natural and Synthetic Speech" submitted to Interspeech 2011 [16]

2 Research System Evaluation

2.1 Introduction

In the research system evaluation, the main questions that we need to answer are: 1) How well can listeners discriminate between bilingual speakers across-languages? 2) How well can listeners judge speaker identity when comparing synthetic speech to natural speech? and finally 3) How well can listeners discriminate between bilingual speakers across-languages and across speech types (natural, synthetic)? In order to answer these questions, a number of different experiments were carried out. Five perceptual experiments were conducted, in which the following types of speech were compared:

- Exp. I: Finnish/German/Mandarin natural speech compared to English natural speech.
- Exp. II: Mandarin synthetic speech compared to English synthetic speech (both within-language adaptation).
- Exp. III: Mandarin natural speech compared to English synthetic speech (within-language adaptation).
- Exp. IV: Mandarin natural speech compared to English synthetic speech (across-language adaptation).

- Exp. V: English natural speech compared to English synthetic speech (within-language adaptation) – English and Finnish speakers.

In going from Exp. I to Exp. IV an increasing number of the elements that play a role in the EMIME scenario are included. Exp. IV should be viewed as the final EMIME evaluation. It most closely resembles how the EMIME system as a whole works and therefore how it should be evaluated. To clarify, in Exp. IV, listeners were asked to decide on a speaker's identity whilst comparing a user's Mandarin natural speech to their synthetic English speech – which had been adapted to the user's voice by using their Mandarin natural speech as adaptation data. This is a very challenging task for listeners as they have to deal with the combination of across-language (Mandarin versus English) and across speech type (natural versus synthetic) factors while trying to identify speakers. We carried out the intermediate experiments to answer questions about how listeners deal with these various factors. Exp. V differs from the other four in that it deals with within-language speaker discrimination rather than across-language speaker discrimination.

In Exp. I we investigate how well listeners discriminate between bilingual speakers across languages. Winters et al. [18] 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 speakers across German and English. They showed that listeners can generalize knowledge of speakers' voices across these two phonologically similar languages. However, it is unknown whether this is also the case for languages that are less closely related. This is a key reason that we were interested in investigating listeners' discrimination of speakers across English and Finnish, and across English and Mandarin, as both these languages are in different language families from English (Finno-Ugric and Sino-Tibetan respectively) In addition, they are both focus languages within EMIME.

Exp. II focuses on how well listeners are able to identify speakers when the trials consist of synthetic speech instead of natural speech. In Exp. III, we compare natural Mandarin to synthetic English – created using within-language adaptation. Within-language adaptation means that the sentences used for adaptation are from the same language as the synthetic speech that is being created. In Exp. IV, the comparison in still between natural Mandarin and synthetic English, but in this case, across-language adaptation has been applied. Across-language adaptation means that sentences from Mandarin are used to adapt English synthetic speech. These experiments touch on the question: Does the synthetic speech which has been adapted to sound like the original speaker actually sound like them? As Exp. II – IV all deal with across-language speaker discrimination which makes it impossible to say whether changes in the listeners' performance are due to across-language issues or across-speech type factors.

Exp. V fills this gap by only looking at speaker discrimination across-speech types. It has a slightly different focus as it deals with within-language speaker discrimination and additionally the effect of accent is investigated. First, in this experiment we aimed to determine how listeners perform in a discrimination task when asked to compare synthetic speech to natural speech. Difficulties associated with comparing synthetic to natural speech have been discussed in detail in [19]. It has been shown, for example, that synthetic speech is less intelligible than natural speech, it requires more cognitive resources, and it is more difficult to comprehend. All these factors will influence how listeners compare synthetic stimuli to natural stimuli. We want to find out to what extent this impacts the ability of listeners to identify a speaker in synthetic stimuli. To study only this across-speech type factor, we disregard the across-language element of our evaluation by restricting Exp. V to discrimination within one language – English.

Second, in Exp. V, it is to be expected that a person's accent in a foreign language will influence the perception of their identity. So, how should the synthetic voice of a person in a foreign language sound? There are as many ways of speaking a second language as there are speakers, but some regional characteristics can be observed, e.g., a type of foreign-accent [2]. We explored this by creating synthetic speech with different accents by using differently accented average voice models. Our aim is to find out whether using different average voice models affects listeners' ability to discriminate between speakers. Exp. V covers two sets of experiments: Exp. V-A deals with speaker adaptation using all the available adaptation material (105 sentences per speaker), and Exp. V-B looks at rapid adaptation, i.e. using only 15 or 5 adaptation sentences per speaker. Both experiments investigate the role

of accent in this context.

In the following sections, a short description of the data and the speech synthesis techniques that were employed are given. Where it is relevant, references to other EMIME deliverables are given rather than lengthy descriptions. Next, the set-up of the speaker discrimination experiments is described. In the results section, each of the five experiments (Exp. I – Exp. V) are discussed in turn. Finally, in the conclusions section, the findings of our speaker discrimination experiments will be summarized and conclusions drawn.

2.2 Data

To investigate across-language speaker discrimination we recorded a database of bilingual speech. The language pairs we chose to record are English/German, English/Finnish and English/Mandarin. English/German was selected to be able to compare our results to [18]. Finnish and Mandarin were included as they are focus languages of EMIME.

The English/German and English/Finnish speech databases were recorded at the University of Edinburgh in 2009 and have, for the most part, been described in Deliverable D1.2. Additional Mandarin/English data was collected at UEDIN in November 2010. More details on all of the data collected can be found in [13, 17].

Accent-rating experiments have been carried out on the English data of all the subjects to ensure speakers with the least degree of foreign accent were selected for inclusion in the perception experiments [13, 17]. We assume that if bilingual speakers are highly fluent in their two languages, speaker discrimination should be more difficult. Anecdotal evidence seems to suggest that proficient non-native speakers of English do not necessarily sound like the same person when speaking their native language, as when speaking English.

2.3 Speaker adaptation for speech synthesis

Speaker adaptation has been described in great detail in WP3 deliverables. Here it suffices to mention that for Exp. II, III and IV, the within-language and across-language adaptation was carried out as described in [7]. For Exp. IV, two English average voices were trained, one using a Finnish-accented English data set, another using an American-accented English data set [15].

2.4 Listening tests

To evaluate the success of our modelling attempts we use speaker discrimination tests. In this type of test, listeners are asked to listen to two sentences and decide whether or not they think the sentences could have been produced by the same speaker. In Exp. V-A, in order to compare our results to the mainstream, we also include a MOS-style task. Finally, in Exp. V-B, an accent classification task is included. It is used as a tool to further analyse listeners' behaviour.

2.4.1 Speaker discrimination test set-up

An important factor in speaker identification or discrimination is speaker familiarity. Whether or not a listener is familiar with a speaker will influence how well they can recognise or identify them, as well as how well they can discriminate between them and other speakers [5, 11]. Of course unfamiliar voices can become familiar voices with training. In [8], speaker-specific learning in speech perception was investigated. They found that listeners' familiarity with speakers 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 [8] showed that learning is faster when using sentences rather than words and that it is much easier to identify speakers' voices from sentences than from isolated words. It can be expected that a sentence is long enough for some speaker learning to occur. Therefore, using sentence length stimuli should provide the listeners with sufficient speaker-specific information about a speaker to make an

informed decision. Furthermore, in EMIME, the more likely scenario is that interlocutors are not familiar with each other.

Three different language pairs were investigated. Each language pair test consisted of two parts: a female set and a male set of data. We did not combine genders within any of the tests. Each part of the test consisted of 160 trials (i.e. 320 sentences in total). 80 news sentences, ranging in length from 7 to 10 words, were used per test condition, 40 English and 40 German, Finnish or Mandarin. Each sentence occurred four times – twice in a same-speaker trial, twice in a different-speaker trial. Each speaker 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. Table 2a shows the number of trials for each language pair.

Table 2a: *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
Mandarin (F/M)	Eng/Eng	Man/Man	Eng/Man	Man/Eng
same	20	20	20	20
different	20	20	20	20

2.4.2 MOS test

Almost all previous studies, for example the S2ST project TC-STAR [10], work by Latorre and colleagues [6] as well as the EMIME project [14] evaluate the success of across-language adaptation or multi-lingual synthesis by using mean opinion scores (MOS) for similarity and quality. Using MOS 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 and the speech types being compared are natural and synthetic speech [3, 1].

To find out how speaker discrimination tests compare to MOS style rating tasks and to illustrate how certain types of information are lost in MOS ratings we carried out a MOS style rating task using our data in Exp. V-A [15]. In this task, the listeners were asked to rate the similarity of a synthetic speech stimulus compared to the original target speaker on a 5-point scale ranging from 1 for "sounds like a totally different person" to 5 for "sounds like exactly the same person". Listeners rated all five male native Finnish speakers. All sentences were in English. In each trial, the natural reference stimulus was played first followed by the synthetic stimulus.

2.4.3 Accent classification test

In Exp. V-B [4], in addition to speaker discrimination tests, an accent classification task was carried out to ascertain whether listeners perceive different accents depending on which average voice model is used as a basis for the synthetic speech. In this task, for each speaker (five English and five Finnish) one sentence for each of the five speech types (natural speech and four synthetic speech types) was selected. The four synthetic types were defined by being based on either an American-accented average voice or a Finnish-accented average voice and whether five or fifteen adaptation sentences were used for adaptation. The abbreviations used are as follows:"S"= Synthetic, "A" = American,"F" = Finnish and "5"/"15" = number of adaptation sentences. This results in the synthetic types: SA5 SA15, SF5 and SF15. Listeners were asked to listen to the sentences and decide whether they thought the accent was mainly American, British or Scandinavian (Scandinavian was chosen as a label rather than Finnish as we felt the broader label would be easier for native English listeners to deal with).

2.5 Listeners

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 speaker discrimination test conditions (160 trials) to complete. This took between 35 and 45 minutes. Listeners were asked to decide whether the two sentences in each pair were spoken by the same speaker or by two different speakers. In addition to making same/different decisions they were asked to indicate on a 3-point scale how sure they were of their decision. In the case of Exp V-A and V-B, listeners also did a MOS test or an accent classification test. Subjects were paid for their participation.

2.6 Results

Each test condition was judged by 10 listeners. Per listener data were pooled for each test condition. In all boxplots, 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. Abbreviations that are used in the figures are "Eng" for English, "Fin" for Finnish, "Ger" for German and "Man" for Mandarin.

2.6.1 Exp. I: Natural Speech - across-language discrimination

In this section, the results for speaker discrimination of natural speech across-languages are presented. First, the results for Finnish/English and German/English are given (these are taken from [12]). Next, the results for Mandarin are presented (taken from [16]).

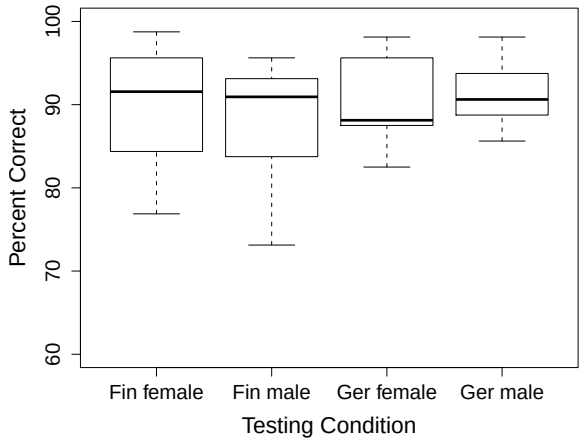

Figure 2a: Exp. I – Percent correct discrimination for German and Finnish male and female test conditions, all natural speech.

Figure 2a shows a boxplot of percent correct for the German and Finnish test conditions. 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. Therefore, in further analyses the results are combined.

Figure 2b is a boxplot showing percent correct for the Finnish and German test 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 $[F(7, 192) = 8.04, p < 0.001]$. Tukey HSD (Honestly Significant Difference) multiple comparisons of means

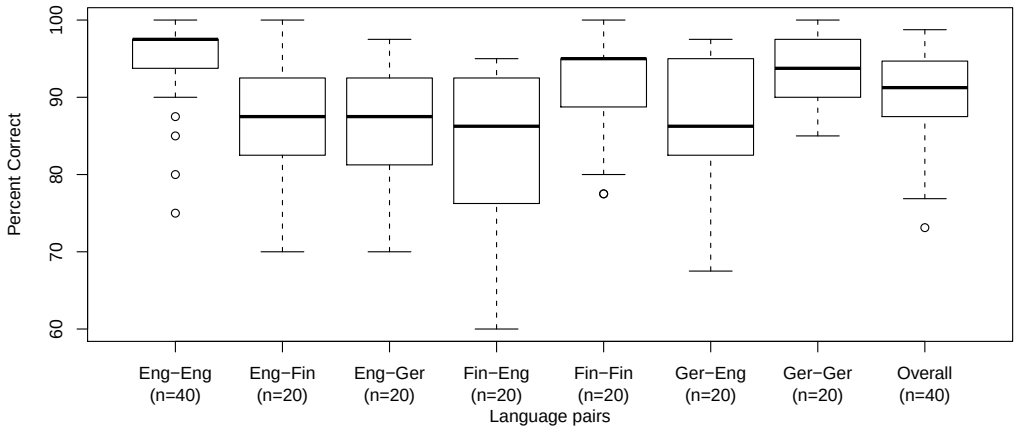

Figure 2b: Exp. I – Percent correct discrimination for German and Finnish language pairs, test conditions pooled, all natural speech.

with 95% family-wise confidence level were conducted to analyze the effect of language pair in more detail. The Tukey HSD test revealed that speaker pairs are incorrectly classified significantly more often in mixed-language conditions than they are in matched-language conditions.

Figure 2c shows the results for the Mandarin test conditions. An ANOVA with test condition (Mandarin female, Mandarin male) as the between-test factor shows there is a significant main effect of test condition $[F(1, 18) = 6.49, p = 0.02014]$. There is a significant difference between listener responses to Mandarin males and their responses to Mandarin females. Therefore, results for Mandarin females will be presented separately from the results for Mandarin males in the following analyses.

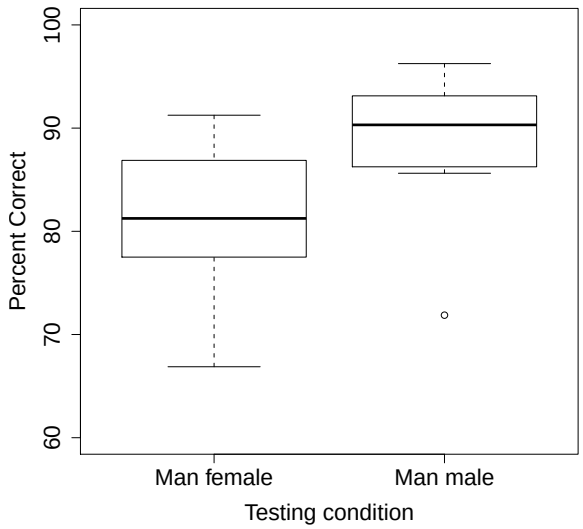

Figure 2c: Exp. I – Percent correct discrimination for Mandarin male and female test conditions, all natural speech.

Figure 2d is a boxplot showing percent correct for the Mandarin male and female test conditions. For all our Mandarin results, the mixed-language conditions "Eng/Man" and "Man/Eng" have been combined as no significant differences were found between them (this could also have been done for the earlier Finnish and German results, but those figures were taken directly from [12] in which the merging was not applied). An 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 $[F(3, 36) = 11.87, p < 0.001]$. The Tukey HSD test revealed that Mandarin speaker pairs are also incorrectly classified significantly more often in mixed-language conditions than they are in matched-language conditions.

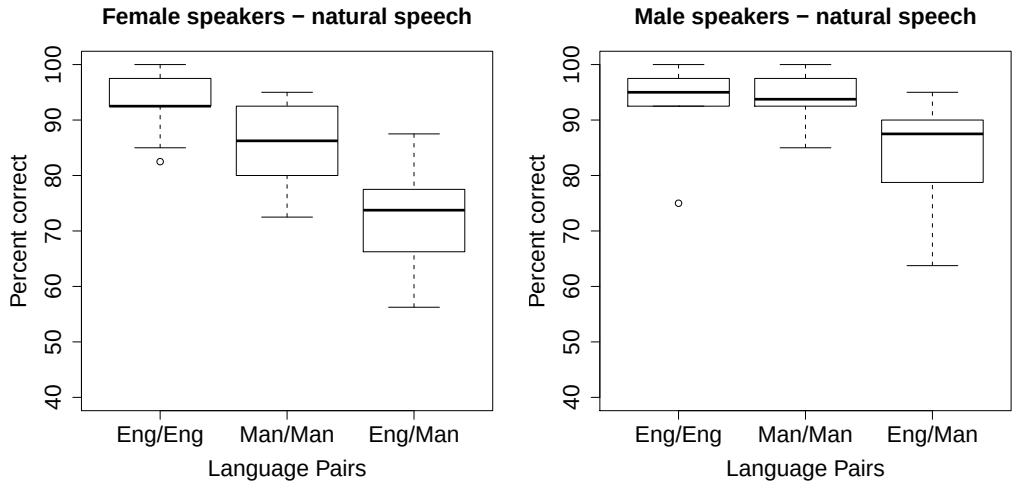

Figure 2d: Exp. I – Percent correct discrimination per language pair for Mandarin male and female test conditions, all natural speech.

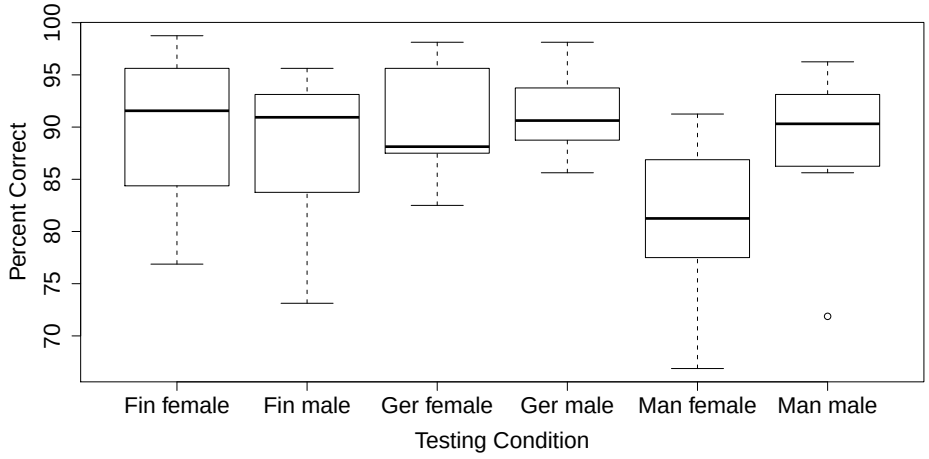

Figure 2e: Exp. I – Percent correct discrimination for Finnish, German and Mandarin male and female test conditions, all natural speech.

Figure 2e is a boxplot which shows results for all three languages combined. An ANOVA conducted on the

percent correct with test condition as the between-test factor shows there is a significant effect of test condition $[F(5, 54) = 3.59, p < 0.001]$. A Tukey HSD test showed that listeners incorrectly identify Mandarin female speakers significantly more often than they do Finnish females, German females and German males.

2.6.2 Exp. II: Synthetic Speech - across-language discrimination

In this section, the results for speaker discrimination of synthetic speech across-languages are presented. In this case, we have data comparing listener's responses for speaker pairs speaking Mandarin and English (taken from [16]). Figure 2f shows boxplots of percent correct for the comparison between Mandarin and English synthetic speech. The synthetic stimuli were all created using within-language adaptation. For each speaker, 60 Mandarin or 105 English sentences were used as the adaptation data for each respective language. Adaptation data sets are the same size, Mandarin sentences are simply longer so fewer are needed to achieve the same amount of speech data.

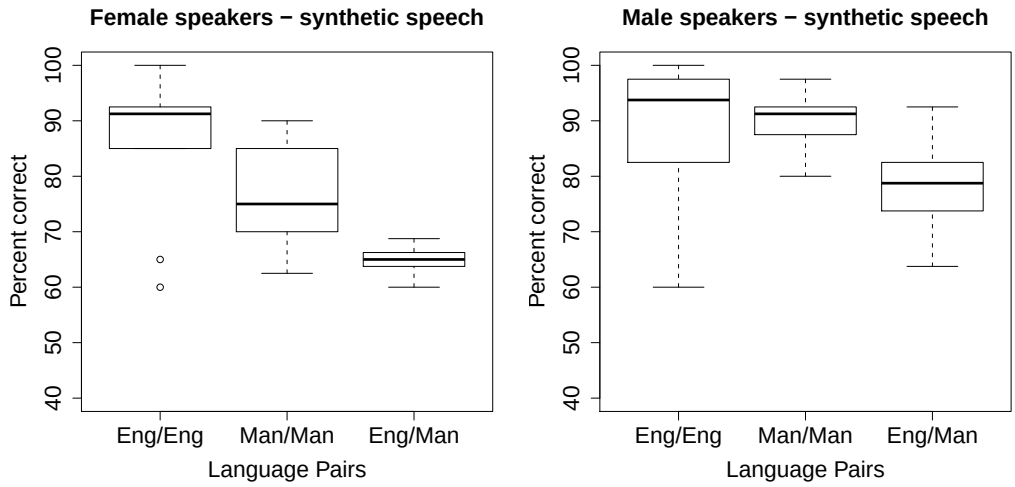

Figure 2f: Exp.II – Percent correct discrimination for synthetic Mandarin and English speech.

ANOVAs with language pair (Eng/Eng, Man/Man and Eng/Man) as the within-test factor were conducted. For both female and male test sets a significant main effect of language pair was found; for females: $[F(3, 36) = 11.49, p < 0.001]$ and for males: $[F(3, 36) = 3.89, p < 0.01655]$. The Tukey HSD tests showed that listeners act in the same way for synthetic speech as they do for natural speech in that they perform significantly worse on mixed-language trials than on matched-language trials. Table 2b shows the mean percent correct for each of the language pairs, per speech type and test condition. The drop in performance when going from natural speech to synthetic speech is about 7-9% in the Mandarin female test condition and 4-6% in the Mandarin male test condition.

2.6.3 Exp. III: Natural and Synthetic Speech (within-language adaptation) - across-language discrimination

Figure 2g shows boxplots of percent correct for the comparison between natural Mandarin and English synthetic speech. The synthetic stimuli were created using *within-language* adaptation. Again, ANOVAs with language pair (Eng/Eng, Man/Man and Eng/Man) as the within-test factor were conducted. For both female and male test sets a significant main effect of language pair was found; for females: $[F(3, 36) = 31.58, p < 0.001]$ and for males: $[F(3, 36) = 46.6, p < 0.001]$. Once again the Tukey HSD tests clearly showed that listeners perform significantly worse on mixed-language trials than on matched-language trials. No significant differences were found between Eng/Eng and Man/Man.

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

Test condition	Speech type	Language pair		
		Eng/Eng	Man/Man	Eng/Man
Mandarin female	Natural	92.8	85.5	72.6
	Synthetic	86.3	76.3	64.6
Nat - Syn Difference		6.5	9.2	8.0
Mandarin male	Natural	94.0	94.0	84.0
	Synthetic	89.3	89.8	78.1
Nat - Syn Difference		4.7	4.2	5.9

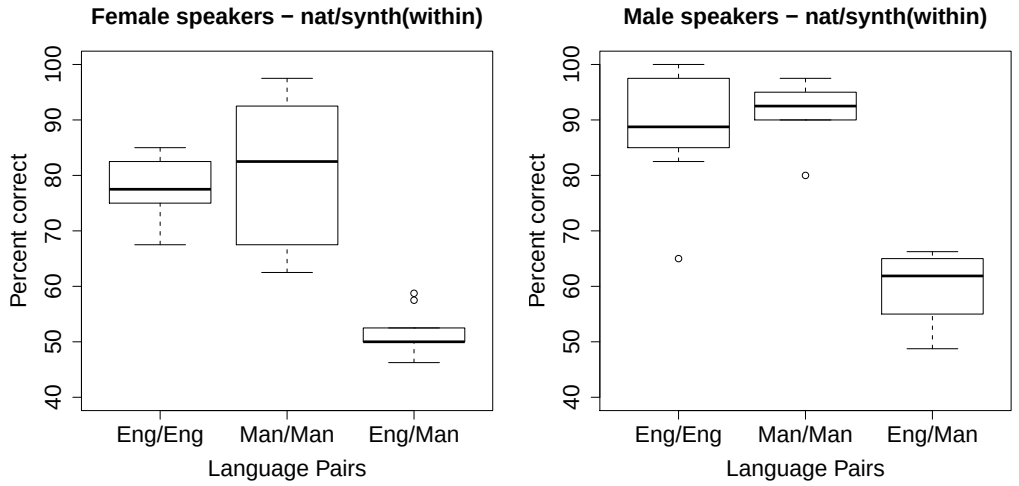

Figure 2g: Exp. III – Percent correct discrimination for natural Mandarin and synthetic English (within-language adaptation.)

Table 2c shows the mean percent correct for each of the language pairs, per speech type and test condition. There is a drop in listeners' performance when we compare Exp. II (synthetic speech) to Exp. III (synthetic and natural speech). For the Mandarin female test condition the drop is 13% and for Mandarin males it is 18%, in the mixed-language condition. The improvement found in the Mandarin matched language condition is due to going from a test containing Mandarin synthetic speech to Mandarin natural speech. Notice, however, that the results for Man/Man are somewhat lower here than those in Exp. I where all the conditions were natural speech. This is to be expected as results are always influenced to a certain extent by the other comparisons present in any one test set. A somewhat surprising result is the drop in performance (9%)for the Mandarin female Eng/Eng condition, because in both cases the trials consist of synthetic English (within-language adaptation), i.e., it is the exact same condition.

2.6.4 Exp. IV: Natural and Synthetic Speech (across-language adaptation) - across-language discrimination

Figure 2h shows boxplots of percent correct for the comparison between natural Mandarin and English synthetic speech. The synthetic stimuli were created using *across-language* adaptation. Again, ANOVAs with language pair (Eng/Eng, Man/Man and Eng/Man) as the within-test factor were conducted. For both female and male test sets a significant main effect of language pair was found; for females: $[F(3, 36) = 70.00, p < 0.001]$ and for males:

Table 2c: *Mean percent correct for each language pair, per test condition and speech type. Speech type comparisons are Synthetic Mandarin /Synthetic English or Natural Mandarin/Synthetic English. All synthesis based on within-language speaker adaptation.*

		Language pair		
Test condition	Speech type	Eng/Eng	Man/Man	Eng/Man
Mandarin female	Synthetic/Synthetic	86.3	76.3	64.6
	Natural/Synthetic	77.3	81.0	51.5
Syn/Syn - Nat/Syn Difference		9	-4.7	13.1
Mandarin male	Synthetic/Synthetic	89.3	89.8	78.1
	Natural/Synthetic	88.3	92.3	60.4
Syn/Syn - Nat/Syn Difference		1.0	-2.5	17.7

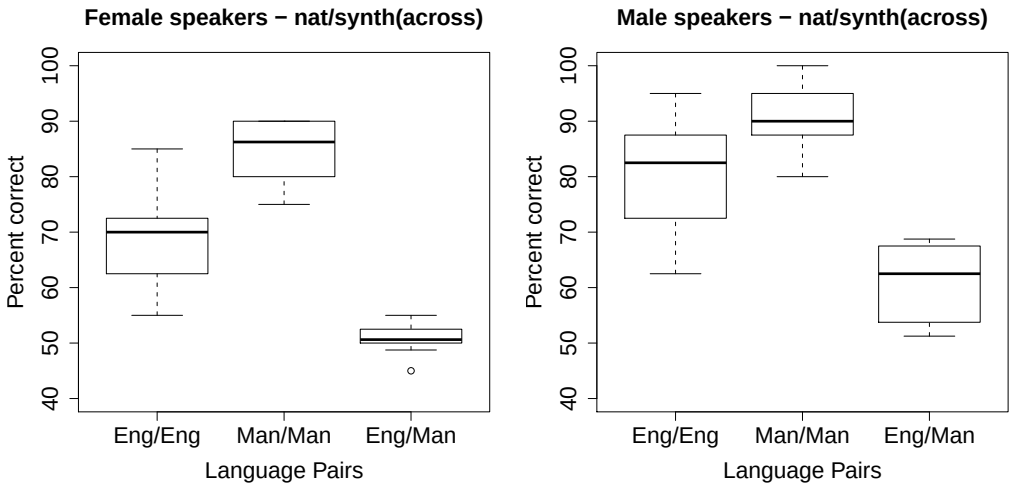

Figure 2h: Exp. IV – Percent correct discrimination for natural Mandarin and synthetic English (across-language adaptation.)

$[F(3, 36) = 29.92, p < 0.001]$. Once again the Tukey HSD tests clearly showed that listeners perform significantly worse on mixed-language trials than on matched-language trials. For both female and male test conditions there was also a significant difference between Man/Man and Eng/Eng. This is in contrast to the previous experiments (Exp. 1 – Exp. III) in which no significant differences between matched-language trials were found, irrespective of the speech being natural or synthetic.

Table 2d shows the mean percent correct for each of the language pairs, per speech type and test condition. The drop in performance when going from synthetic speech with within-language speaker adaptation to synthetic speech with across-language speaker adaptation is about 8% in both the Mandarin female and male test conditions for the English matched-language trials. The Mandarin matched-language trials are all natural speech so no degradation is expected to occur here. Not much of a difference is found between across-language adaptation and within-language adaptation as far as the mixed-language condition is concerned.

In order to be able to say a little more about listener's behaviour in Exp. IV, the same/different responses were converted into nonparametric measures of sensitivity (A') and Griers' bias (B'') [9]. Both these measures are based on the proportion of "hits" and "false alarms". Hits in this context are when a listener labels 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

Table 2d: *Mean percent correct for each language pair, per test condition and speech type. Speech type comparisons are natural Mandarin and synthetic English. Synthesis either based on within-language speaker adaptation or across-language speaker adaptation.*

		Language pair		
Test condition	Speech type	Eng/Eng	Man/Man	Eng/Man
Mandarin female	Natural/Synthetic(within-language)	77.3	81.0	51.5
	Natural/Synthetic(across-language)	69.3	84.5	50.6
Within - Across Difference		8.0	-3.5	0.9
Mandarin male	Natural/Synthetic(within-language)	88.3	92.3	60.4
	Natural/Synthetic(across-language)	80.5	90.8	61.1
Within - Across Difference		7.8	1.5	-0.7

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.

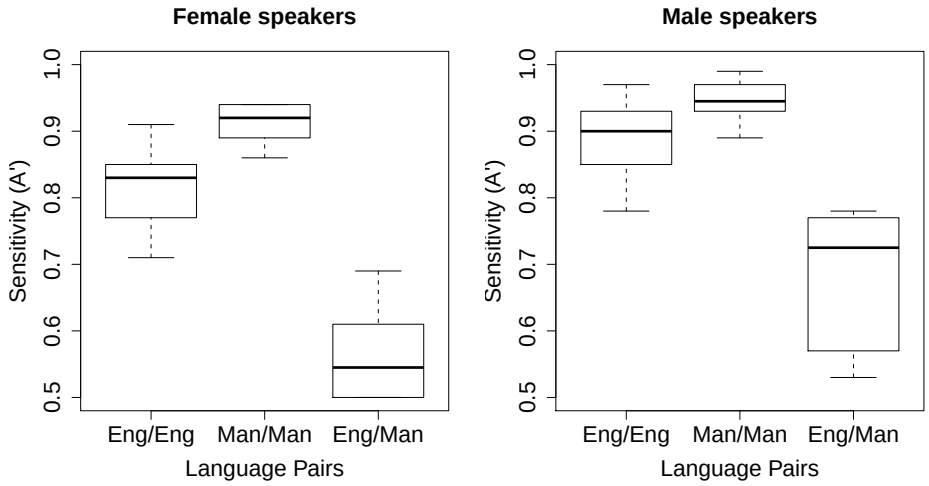

Figure 2i: Exp. IV – Sensitivity (A') for each language pair condition, per test condition.

Figure 2i shows a A' boxplot of listeners' responses for female and male Mandarin speakers and Figure 2j shows B'' boxplots of listeners' responses. The results for sensitivity show us that listeners are performing pretty much at chance level for female Mandarin speakers. Listeners are better at distinguishing the male Mandarin mixed-condition trials. The bias figure shows that there is a little bias towards saying mixed-language pairs are different but it is near to 0. For the English matched language condition, there is a clear bias towards judging trials as "same". For the Mandarin matched language trials there is larger variation in what listeners do but generally the bias is towards labelling these trials as "same" too.

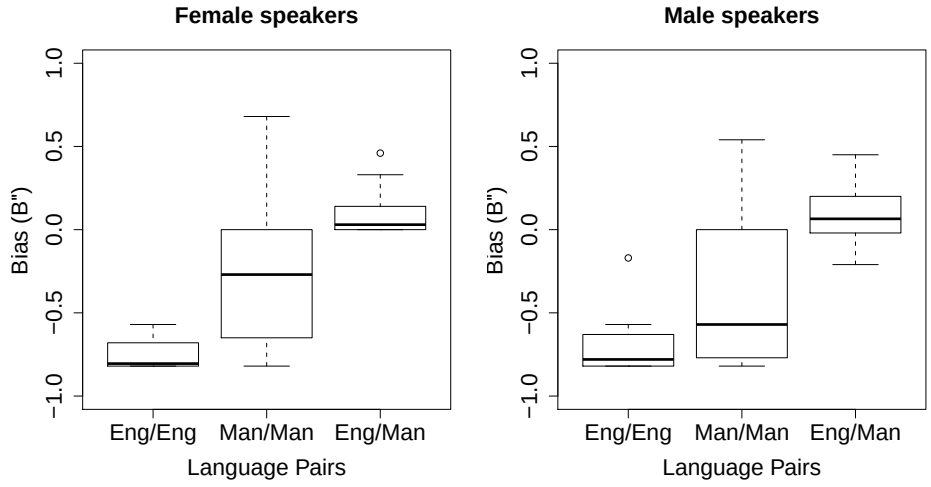

Figure 2j: Exp. IV –Bias (B'') for each language pair condition, per test condition.

2.6.5 Exp. V: Natural and Synthetic Speech (within-language adaptation) - within-language discrimination

In this section, the results for speaker discrimination *within* a language and across speech types are presented. In this case, the results of listener's comparisons of natural and synthetic speech are shown (taken from [15, 4]). In addition to comparing natural and synthetic speech, the effect of accent on speaker identity was investigated by using differently accented average voice models: one trained on Finnish-accented English and the other on American-accented English.

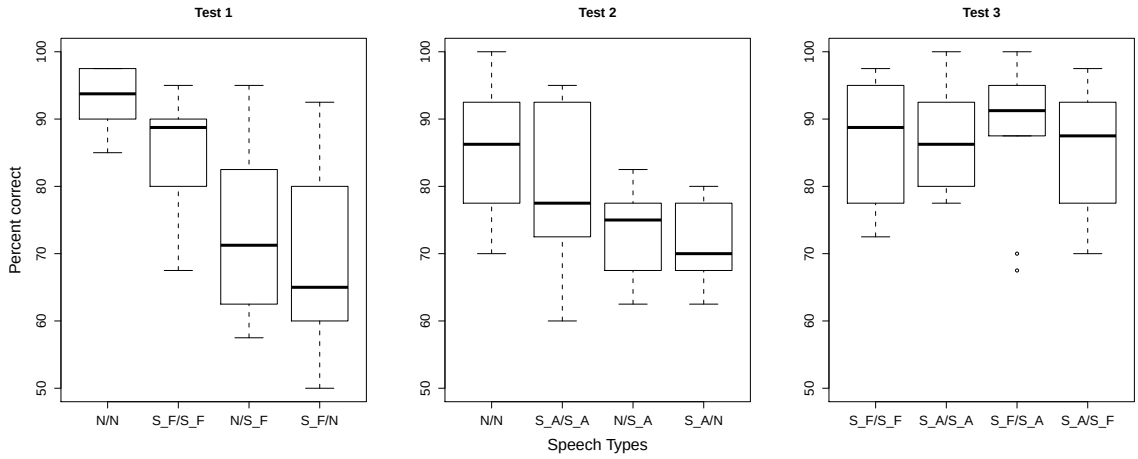

Figure 2k: Exp. V-A – Percent correct discrimination per speech type pair for the three discrimination tasks. N = Natural speech, S = Synthetic speech, _A = American-accented average voice model, _F = Finnish-accented average voice model.

Exp V-A looks at speaker discrimination tests in which the synthetic speech was created using the two different average voice models and 105 adaptation sentences per speaker. Also, MOS style scores are compared to speaker-discrimination test results. Exp. V-B, was done as a follow-up to Exp. V-A. In Exp. V-B, the effect of using limited amounts of adaptation data on speaker identity (i.e. rapid adaptation), once again using the two differently

accented average voice models, was investigated. In addition to looking at the effect of limited amounts of speaker adaptation data on speaker identity, we wanted to determine whether using 105 adaptation sentences as in Exp. V-A overrides the effect of accent in the average voice model, or whether, under this condition, listeners actually identify a person as the same speaker but with two different accents.

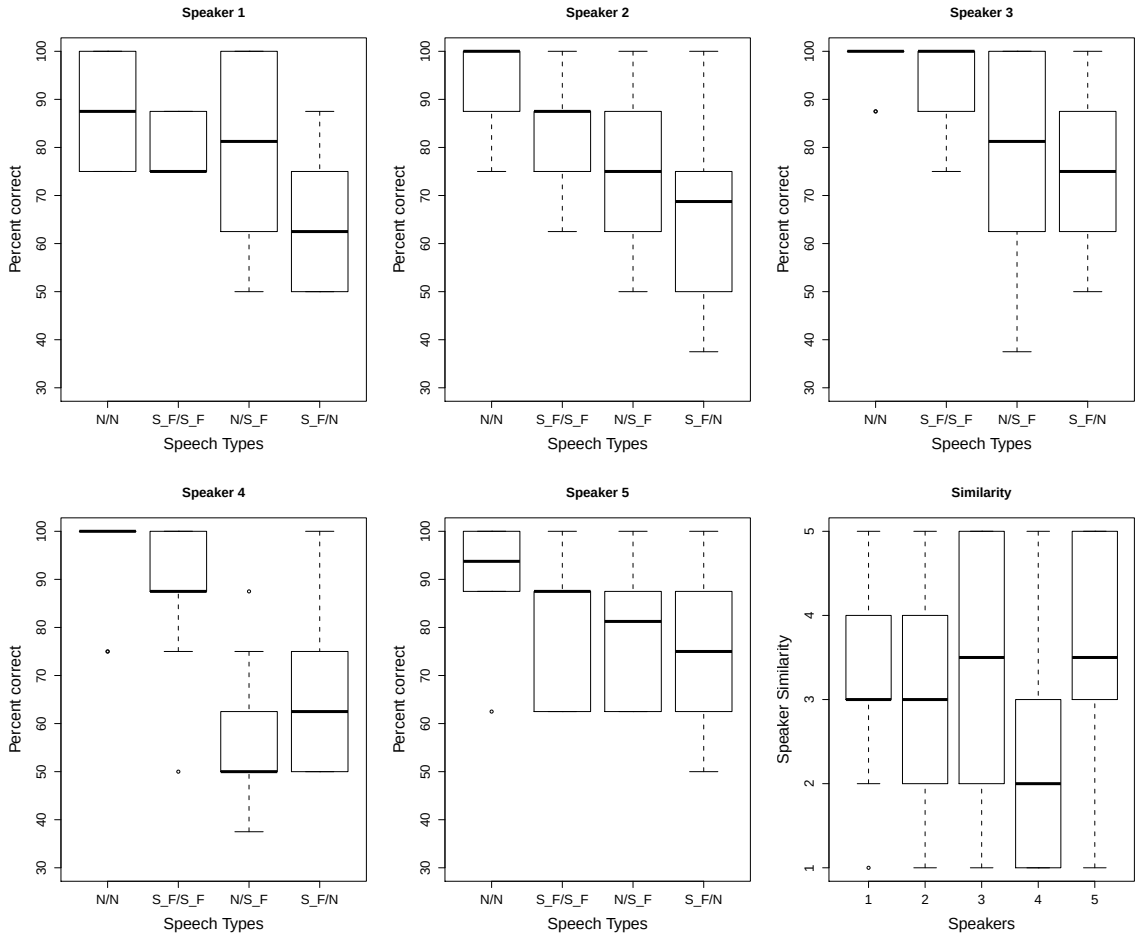

Figure 2l: Exp. V-A – Percent correct per speech type pair for individual speakers and similarity scores comparing to original speaker.

The discrimination task for Exp. V-A consisted of three tests: Test 1 compares natural speech (N) to synthetic speech based on the Finnish-accented average voice model (S_F), Test 2 compares natural speech to synthetic speech based on the American-accented average voice model (S_A) and Test 3 compares the two types of synthetic speech, S_F and S_A. Figure 2k shows boxplots of percent correct per speech type pair for each of the three discrimination tests. An ANOVA was conducted with speech type (N/N, S/S, N/S, and S/N) as the within-test factor. ANOVAs show a significant main effect of speech type: Test 1 $[F(3, 36) = 11.73, p < 0.001]$ and Test 2 $[F(3, 36) = 5.29, p = 0.004]$. A Tukey HSD test revealed that listeners perform significantly worse when comparing synthetic speech to natural speech than when the speech type is of one type (either synthetic or natural). The ANOVA for Test 3 with speech type (S_A/S_A, S_F/S_F, S_F/S_A and S_A/S_F) as the within-test factor shows there are no significant differences between any of the speech type pairs: listeners correctly identify speakers as an individual irrespective of the accent in the average voice model.

In order to compare the MOS task results to the discrimination task results, individual speaker results for Test

1 have been calculated. Figure 2l shows percent correct scores for each of the five speakers from Test 1, as well as the similarity scores. It shows that high percent correct discrimination for mixed speech type trials (Speakers 1, 2, 3 & 5 $> 70\%$) seems to translate to high speaker similarity scores and low percent correct discrimination (Speaker 4 $< 70\%$) corresponds to low speaker similarity scores.

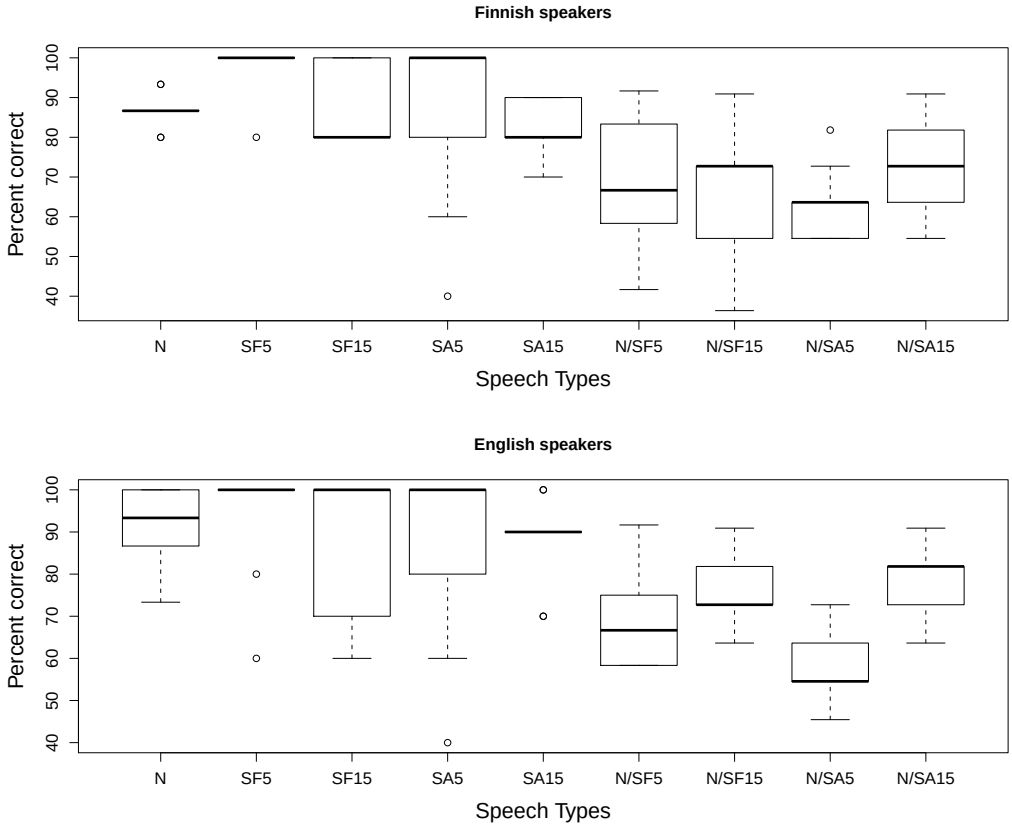

Figure 2m: Exp. V-B – Percent correct discrimination per speech type pair. N = Natural speech, S = Synthetic speech, A = American-accented average voice model, F = Finnish-accented average voice model, 5 = 5 adaptation sentences, 15 = 15 adaptation sentences.

In Exp. V-B we were interested to find out the effect of using limited amounts of adaptation data on the identification of speakers and whether listeners identify an individual with two different synthetic accents as the same person. The same five native Finnish speakers as in Exp. V-B were included and five male native English speakers were added. Figure 2m shows boxplots of percent correct per speech type pair for a speaker discrimination task in which the synthetic speech was based on 5 or 15 adaptation sentences per speaker. Individual listener data were pooled for both tests for all speakers. In this case, "N" indicates a trial in which two natural utterances were compared to each other. "N/SA15" indicates a trial in which a natural utterance was compared to a synthetic utterance based on the American-accented average voice model and for which 15 adaptation sentences were used. An ANOVA was conducted with speech type (natural or synthetic) as the within-test factor. ANOVAs show a significant main effect of speech type: Finnish talkers $[F(8, 72) = 7.63, p < 0.001]$ and English talkers $[F(8, 72) = 6.75, p < 0.001]$. Again, the Tukey HSD test revealed that listeners perform significantly worse when comparing synthetic speech to natural speech than when the speech type is of one type (either synthetic or natural). These results are very similar to the results presented in Figure 2k.

To find out if listeners actually perceive different accents for synthetic speech created using either Finnish-

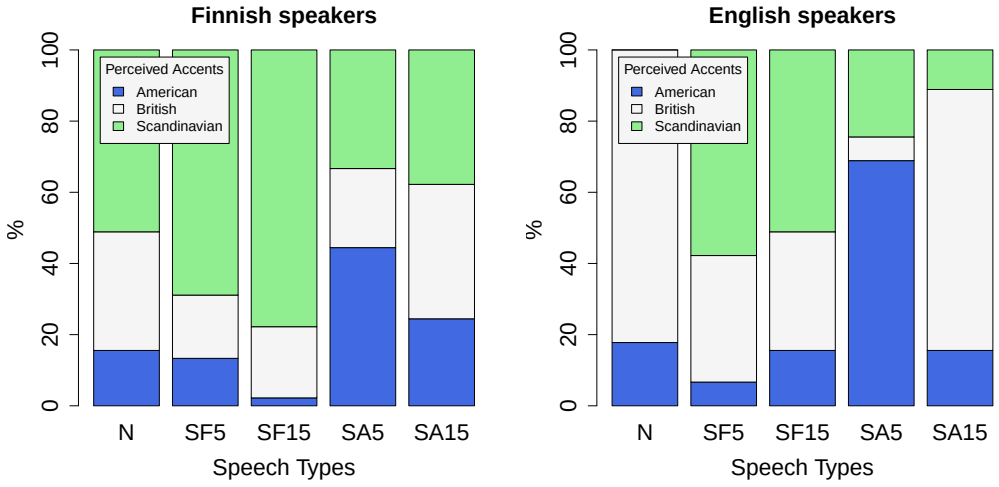

Figure 2n: Exp. V-B – Accent rating of natural and various synthetic sentences for Finnish and English speakers.

accented or American-accented average voice models we carried out an accent classification test. Figure 2n shows the accent classification results. In this figure, the accent is presented as the percentage of sentences that were classified as either American, British or Scandinavian. The first bar, labelled "N" (for natural) shows the results for English speakers. One of the speakers is an American, the other four are British, the actual percentages of perceived accents – roughly 20% British and 80% American – correctly reflects this. For both Finnish and English speakers, using a Finnish-accented average voice model (SF5, SF15) leads to large increases in the percentages of perceived Scandinavian accent and using an American-accented average voice model (SA5, SA15) leads to increases in the percentages of perceived American accent. It is clear from this figure that listeners perceive synthetic speech based on different average voice models with different amounts of adaptation data as belonging to different accent categories.

Figure 2o shows the results of judging speaker identity for trials that consist of different types of synthetic stimuli. In this speaker discrimination experiment, the matched condition trials were the four types of synthetic stimuli compared to themselves and the mixed condition trials consisted of SA5 and SF5, and SA15 and SF15. A TukeyHSD test comparing the matched trial data from Figure 2o to Figure 2m showed the only significant differences across the two experiments was SF5 (from Figure 2o) and SA5 (Figure 2m), for English talkers and SF15, SF5 (Figure 2o) and SA5 (Figure 2m), for Finnish talkers. A further TukeyHSD test comparing the mixed trial data to the matched trial data in Figure 2o shows that SA5/SF5 is not significantly different to the matched conditions SA5 and SF5. And likewise, SA15/SF15 is not significantly different to SA15 and SF15.

Combining the information from Figure 2o and Figure 2n we can conclude that listeners are happy to classify a speaker as themselves even when their synthetic speech is of different accent types.

2.7 Conclusions

Exp. I shows that listeners perform well on a speaker discrimination task. Discrimination accuracy is significantly higher than chance. However, listeners perform significantly better in matched-language conditions than in mixed-language conditions: in matched-language conditions, percent correct is significantly higher. We expected the across-language condition to be the most difficult for listeners and this is corroborated by the results. On average the across-language trials are scored *incorrectly* 8-10% (absolute) more often than the matched-language trials (20% even in the case of the Mandarin female test set).

There is no clear indication that Finnish talker discrimination is more difficult for English native listeners than

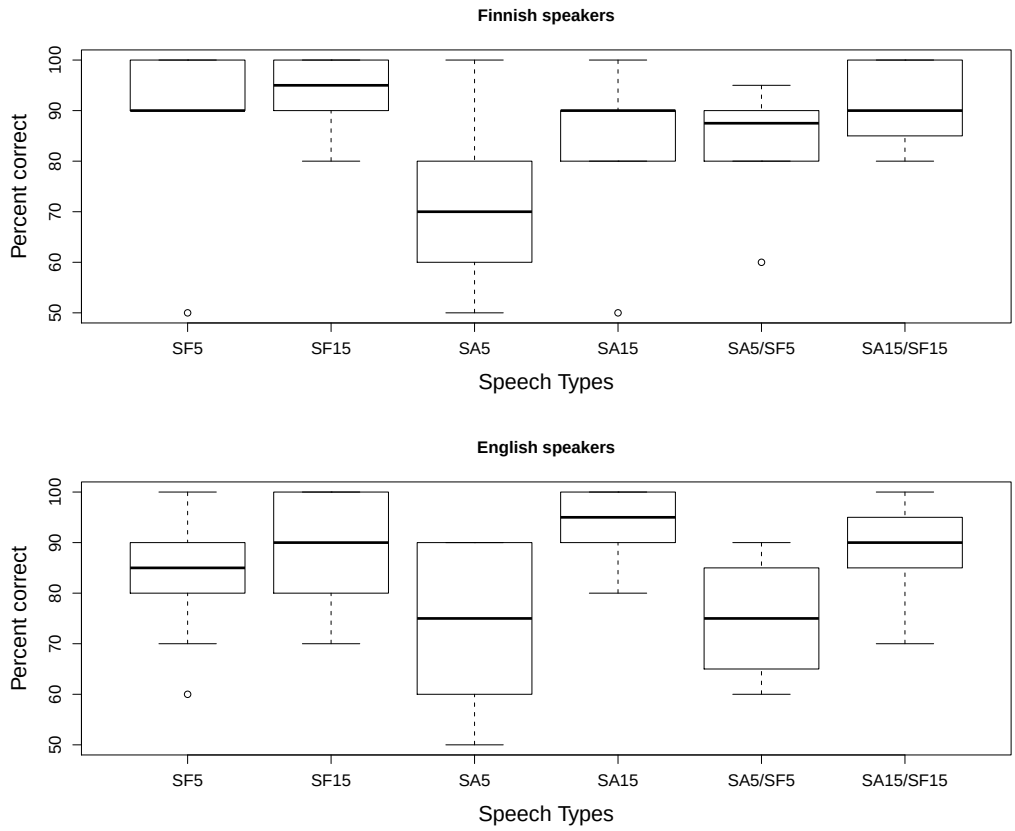

Figure 2o: Exp. V-B. Percent correct discrimination per speech type pair. Comparisons across synthetic speech types.

German talker discrimination. In the matched-language condition results are 2% lower on Finnish than on English, however in the across-language condition results are comparable. Mandarin taker discrimination also does not seem to be more difficult for native English listeners when we look 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 is that the set of five female Mandarin speakers are intrinsically more confusable than the other sets of speakers. All the results that we found for Mandarin females are somewhat lower than for the other speaker sets, but the trends all look to be the same as for the other speaker sets.

Our next question was: How do listeners perform in a discrimination task when asked to compare synthetic speech to natural speech? First, lets look at the results for matched speech type trials, i.e., both sentences in the trial are either synthetic or natural speech. The Mandarin experiments showed us that comparing synthetic speech trials instead of natural speech trials leads to reductions in percentage correct of about 4-9%. In Exp. V-A –Finnish talkers– we found that listeners perform as well on synthetic matched speech type trials as on natural trials. Test 3, in Figure 2k illustrates this best. In Exp. V-B, we even found that listeners manage to identify speakers across accents in synthetic speech almost as well as in natural speech. Therefore, we can conclude that the step going from natural speech to synthetic speech leads to only slight decreases in listeners' performance, if any.

Next, what is the effect of mixed speech type trials within one language? Exp. V shows that the difficulties listeners have in comparing synthetic speech to natural speech, described in [19], are definitely playing a role. The results show that when listeners are only comparing different types of synthetic speech within a language

the average scores are (roughly) between 80 and 90% correct. However, when asked to compare synthetic and natural speech, the scores drop to between 60 and 80% correct. These experiments comparing across speech types, have shown us that it is very important to keep in mind the extra cognitive load listeners experience when hearing synthetic speech.

In Exp V-A, the results from a MOS-style evaluation task were compared to results from a speaker discrimination task. It was clearly shown that MOS do not give the full picture. They do not show whether listeners are able to compare the natural and synthetic speech samples, and the MOS-style task does not answer the question: is this the same speaker or not? Studies investigating speaker similarity in synthesis should not solely rely on MOS to draw conclusions about the success of their methods.

Finally, when in addition to comparing different speech types listeners also have to contend with across-language trials their ability to correctly identify speakers suffers quite substantially. For the female Mandarin speakers, listeners perform at chance level when identifying speakers across languages and across speech types. For the male Mandarin speakers, listeners' ability to identify speakers is still above chance, but there is a large drop (18%) in performance when the synthetic-synthetic results are compared to synthetic-natural results.

This work has shown us that listeners are well able to carry out talker discrimination tasks – deciding whether or not a talker in L2 sounds similar to the original talker in L1 is an achievable task for listeners. The fact that listeners do not seem to experience Mandarin as any more difficult than Finnish or German in a speaker discrimination task is a very interesting finding. In our experiments, we found that the synthetic speech created in EMIME both using within-language and across-language adaptation leads to speaker identities that are individual. Our speaker discrimination tests show that the results for synthetic speech are very similar to the results found for natural speech. However, we also found that care must be taken when comparing different speech types. Speaker discrimination across speech types and across languages is a very difficult task for listeners to perform. We have shown that our speaker discrimination task is more suited to measuring whether or not listeners perceive a speaker as him/herself than a MOS-style rating task in which listeners are asked to judge speaker similarity. Our final results on Mandarin –in which listeners had to judge speaker identity for trials containing natural Mandarin speech and synthetic English speech– show there is still room for improvement. Future research will need to concentrate on further improving a speaker's identity in synthetic speech when the goal is to sound like the original speaker.

3 Real-time demonstrator evaluation

3.1 Introduction

The real-time demonstrator described in D4.6 "Final version of real-time demonstrator" was evaluated internally using a task-based evaluation. Following the recommendations of the second year review, less emphasis was placed on this evaluation than originally anticipated and a relatively small-scale test has been conducted in order to demonstrate that we have achieved the goal of a functioning real-time system which incorporates cross-lingual speaker adaptation.

Three pairs of users were given a simple scenario which they were asked to complete using their own languages. In the following sections we first describe the experimental set-up, i.e., the subject pairs, the scenarios and the type of evaluation that was carried out. This is followed by the results of the evaluations and our conclusions.

3.2 Subject pairs

Three user pairs were asked to participate in the evaluation. One male native US-English speaker interacted with each of three Mandarin speakers (two male and one female). All subjects were recruited at Nokia in Beijing. The subjects were paid for their participation.

For each speaker about 100 sentences in their native language were recorded in a session which took place a number of days prior to the real-time demonstrator evaluation session. These recordings were necessary in order to enable cross-lingual speaker adaptation (CLSA) (for details see D4.6).

3.3 Scenarios

Three scenarios were used. In the first one, the US-English speaker visits China and asks one of the male Mandarin speakers how to get to the subway. In the second scenario, the native US-English speaker checks into a hotel, the female Mandarin speaker is the hotel receptionist. In the final scenario, the US-English speaker has just had dinner in a restaurant and wants to pay the bill.

3.4 Questionnaire

Subjective measures were solicited by means of the questionnaire shown in Figure 3a.

Questionnaire EMIME Real-time Demonstrator Evaluation

Name: _____ **Native language**: _____
Gender: _____ **Other languages**: _____
Age: _____

Please answer the following questions:

1. **How easy did you find it to use the system overall?**
   very easy □—□—□—□—□ very difficult

2. **How well could you understand the sentences in your own language?**
   could understand well □—□—□—□—□ could not understand

3. **Did you think your translated speech sounded like you?**
   very much so □—□—□—□—□ not at all

4. **Did you think the translated speech of the other person sounded like them?**
   very much so □—□—□—□—□ not at all

5. **Were the sentences sensible? Did they make sense to you?**
   very sensible □—□—□—□—□ very nonsensical

6. **What did you think of the speed of the system (i.e waiting for translation)?**
   very fast □—□—□—□—□ much too slow

7. **How easy was it to navigate on the phone?**
   very easy □—□—□—□—□ very difficult

8. **How likely would you be to use this system in real life?**
   definitely □—□—□—□—□ definitely not

9. **Please tell us anything else you would like to about your experience in using this system:**
   _____
   _____
   _____
   _____

Figure 3a: Questionnaire filled in by the subjects.

3.5 Results

The raw results are shown in Table 3a.

Table 3a: *Subjects replies to the questionnaire.*

	Subjects				
Question	Mandarin 1 (M)	Mandarin 2 (M)	Mandarin 3 (F)	English (M)	Average
1:	3	2	3	3	2.75
2:	3	3	3	3	3
3:	3	4	3	5	3.75
4:	2	3	2	5	3
5:	2	3	3	2	2.5
6:	2	2	3	5	3
7:	2	3	1	3	2.3
8:	3	4	3	4	3.5

To summarize, subjects found the system reasonably easy to use. They were able to understand the synthetic speech, although they did not think it was easy. They also did not think the translated speech sounded like themselves, but found that the other person's translated speech sounded a bit more like them. Most of the subjects were satisfied with the system's speed and they found the phone easy to navigate. However, in real life they probably would not choose to use the system.

Note that the native US-English subject seems to have higher expectations for the quality and speed of the translation system. He found the translated speech totally unlike himself, or the others, and also thought the system was much too slow, quite different to the Mandarin subjects' opinions. The native US-English speaker also commented on the system: Lots of potential, but still long way to go

3.6 Conclusions

The real-time demonstrator evaluation showed us that the system works, there is an EMIME end-to-end system which naive subjects are able to operate. The results show us that there is room for improvement in the CLSA component before it is convincing to users that the translated synthetic speech sounds like the original speaker. This finding is very much in line with the results presented for the research system in Section 2.

[1] T.L. Eadie and P.C. Doyle. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *JASA*, 112:3014–3021, 2002.

[2] J.E. Flege. Second language speech learning, theory, findings and problems. In W. Strange, editor, *Speech Perception and Linguistic Experience: Issues in Crosslanguage Research*. York Press, 1995.

[3] B.R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G.S. Berke. Comparing internal and external standards in voice quality. *J. Sp. Hear. Res.*, 36:14–20, 1993.

[4] R. Karhila and M. Wester. Rapid adaptation of foreign-accented HMM-based speech synthesis. In *submitted to Interspeech 2011*, 2011.

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

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

[7] 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*, 2010.

[8] L.C. Nygaard and D.B. Pisoni. Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3):355–376, 1998.

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

[10] D. Sündermann, H. Höge, A. Bonafonte, and J. Ney, H.and Hirschberg. Text-independent cross-language voice conversion. In *Proc. Interspeech '06*, pages 2262–2265, Pittsburgh, USA, 2006.

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

[12] M. Wester. Cross-lingual talker discrimination. In *Proc. Interspeech '10*, 2010.

[13] M. Wester. The EMIME Bilingual Database. Technical Report EDI-INF-RR-1388, The University of Edinburgh, 2010.

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

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

[16] M. Wester and H. Liang. Cross-lingual speaker discrimination using natural and synthetic speech. In *submitted to Interspeech 2011*, 2011.

[17] M. Wester and H. Liang. The EMIME Mandarin Bilingual Database. Technical Report EDI-INF-RR-xxxx, The University of Edinburgh, 2011.

[18] S.J. Winters, S.V. Levi, and D.B. Pisoni. Identification and discrimination of bilingual talkers across languages. *J. Acoust. Soc. Am.*, 123:4524, 2008.

[19] S.J. Winters and D.B. Pisoni. Speech synthesis, perception and comprehension of. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages 31–49. Elsevier, 2005.