Rapid Adaptation of Foreign-accented HMM-based Speech Synthesis

Reima Karhila[1], *Mirjam Wester*[2]

[1]Adaptive Informatics Research Centre, Aalto University, Helsinki, Finland
[2]Centre for Speech Technology Research, University of Edinburgh, UK
rkarhila@cis.hut.fi, mwester@inf.ed.ac.uk

Abstract

This paper presents findings of listeners' perception of speaker identity in synthetic speech. Specifically, we investigated what the effect is on the perceived identity of a speaker when using differently accented average voice models and limited amounts (five and fifteen sentences) of a speaker's data to create the synthetic stimuli. A speaker discrimination task was used to measure speaker identity. Native English listeners were presented with natural and synthetic speech stimuli in English and were asked to decide whether they thought the sentences were spoken by the same person or not. An accent rating task was also carried out to measure the perceived accents of the synthetic speech stimuli. The results show that listeners, for the most part, perform as well at speaker discrimination when the stimuli have been created using five or fifteen adaptation sentences as when using 105 sentences. Furthermore, the accent of the average voice model does not affect listeners' speaker discrimination performance even though the accent rating task shows listeners are perceiving different accents in the synthetic stimuli. Listeners do not base their speaker similarity decisions on perceived accent.

Index Terms: speech synthesis, rapid adaptation

10.21437/Interspeech.2011-701

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`). This objective raises the questions: 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?

Most previous work in S2ST uses mean opinion scores (MOS) to measure speaker similarity [1, 2, 3]. In [4], we showed that similarity MOS scores do not give a complete picture. They do not show whether listeners are able to compare natural and synthetic speech samples and a MOS-task is unable to answer the question whether different speech stimuli are perceived to be from the same speaker. Therefore, instead of using a MOS-task, we evaluate our proposed modelling techniques by carrying out speaker discrimination experiments. In this type of test, subjects listen to two sentences and decide whether the sentences could have been produced by the same speaker.

In our previous work [5, 4], we looked at how well listeners discriminate between speakers across languages [5] and how well they discriminate between speakers when comparing synthetic speech to natural speech [4]. In [5], we investigated listeners' discrimination performance across language boundaries (German-English and Finnish-English) when the stimuli consist of natural speech. These experiments showed that listeners are able to complete this task well, and can discriminate between speakers significantly better than chance. However, on cross-lingual speaker trials listeners perform significantly worse than on matched-language trials.

Listeners' discrimination ability when comparing synthetic speech to natural speech within one language (English) was investigated in [4]. It was found that listeners also complete this task well, with classification results significantly above chance. However, once again, listeners performed significantly worse on mixed trials (synthetic vs natural) than on matched trials (synthetic-synthetic or natural-natural). Furthermore, the degradation in listener's speaker discrimination performance was worse when comparing synthetic vs natural speech, than when comparing speech in different languages.

Another issue that was addressed in [4] was whether using different average voice models would affect listeners' ability to discriminate between speakers. We argued that it is to be expected that a person's accent in a foreign language will influence the perception of their identity. And following on from that, we questioned how the synthetic voice of a person in a foreign language should sound. Basically, there are as many ways of speaking a second language as there are speakers, however, some regional characteristics can be observed, e.g., a type of foreign-accent [6]. We explored this in our synthetic speech by using differently accented average voice models.

In EMIME, speaker adaptation is achieved using a hidden Markov model (HMM) –based synthesis approach [3]. First an average voice model is trained, and then speaker adaptation is performed. This enables us to create synthetic speech with different accents. In [4], Finnish-accented and American-accented average voice models were used to create English synthetic speech stimuli for Finnish speakers.

The results of the speaker discrimination tests in [4] showed no difference between the synthetic speech created with the two differently accented average voice models. We hypothesized that the use of 105 adaptation sentences was overriding any effect of accent in the average voice model, i.e. the two synthetic speech samples simply sounded like the original speaker without any traces of the average voice model's accent. Alternatively, it is also possible that listeners may have been identifying the same speaker but with two different accents.

The study presented in this paper examines which of the two explanations above is more plausible. We investigate the effect of using limited amounts of adaptation data (five and fifteen sentences, i.e. rapid adaptation) on listeners' speaker discrimination performance, once again using the two differently accented average voice models. In addition to a speaker discrimination experiment, we also carried out an accent rating task to ascertain whether listeners perceive different accents depending on which average voice model is used as a basis for the synthetic speech.

2. Rapid Adaptation

2.1. Speaker Adaptation for Speech Synthesis.

HMM-based speech synthesis enables the generation of unique synthetic voices by adapting an average voice model. By using HMMs with explicit duration modelling and by adapting spectral, pitch and duration parameters using sentence-wide phonetic and linguistic context information, it is possible to adapt acoustic, speaking style and prosodic features of the synthetic speech [7, 8]. As a foreign accent can be viewed as a certain type of speaking style, these techniques allow for the adaptation of speaking rhythm, regular mispronunciation patterns and other types of features that are distinctive of foreign accents. The following sections describe the data we used to create average voice models and how the speaker-adapted synthetic speech stimuli were made.

2.2. Average voice models

Two English average voices were trained, one using a Finnish-accented English data set, another using an American-accented English data set. For the Finnish-accented average voice model, 1297 sentences from 6 female and 62 male native Finnish speakers recorded in a classroom setting at Aalto university were used (for more details [4]). The American-accented English data was selected from the WSJ0 si-tr-s set. In order to create average voices of comparable quality, only a subset of 1223 sentences was used. The gender ratio was kept the same with 42 male speakers and 3 female speakers. The sentences were selected to maximise phonetic coverage.

These amounts of data are small for the creation of average voice models, but were considered to be adequate for our adaptation experiments. Full-context labelling for both American-accented English and Finnish-accented English sentences was generated with Festival using the Unilex general American phone set. By using the same context label generation technique, we ensure that differences in prosodic features emerge from the spoken sentences themselves.

The average voice models were trained using the same methods and tools as the EMIME 2010 Blizzard Entry [9]. In short, context-dependent multi-space distribution hidden semi-Markov Models (MSD-HSMMs) were trained on acoustic feature vectors comprising STRAIGHT-analysed Mel-generalised cepstral coefficients, fundamental frequency and aperiodicity features. Speaker-adaptive training was applied to create speaker-adaptive average voice models.

2.3. Rapid Speaker Adaptation

The data used to adapt the speaker independent models were recorded at the University of Edinburgh [10]. Five male native speakers of Finnish who were also fluent in English read sentences in both English and Finnish. In this study, we only used their English speech. In addition to the Finnish speakers, five male native English speakers, recorded in the same conditions, were included.

The two average voices were adapted to the each of the ten speakers (five Finnish and five English), using a set of five or fifteen English sentences. Simultaneous transformation of the cepstral, log F_0 and duration parameters was carried out using CSMAPLR adaptation [7]. To synthesize the test sentences, an excitation signal is generated for each sentence using mixed excitation and PSOLA. From this a synthesised waveform is then generated using the MLSA filter corresponding to the STRAIGHT mel-cepstral coefficients.

3. Evaluation - Listening Test Design

To answer our questions: 1) “What is the effect of limited amounts of adaptation data on the discrimination of speakers?” and 2) “Does accent affect speaker discrimination?”, we designed two listening tests. Both consisted of two tasks: a speaker discrimination task and an accent identification task.

3.1. Speaker Discrimination Task

The first listening test (Exp. I) focused on comparing natural speech to synthetic speech. The second listening test (Exp. II) was concerned with comparing different synthetic speech variants to each other.

For all of the speaker discrimination tasks, listeners were asked to listen to pairs of sentences and to judge whether they thought the sentences were spoken by the same person or by two different people. They were warned that some of the sentences would sound degraded and they were instructed to listen “beyond” the degradation in the signal and concentrate on the identity of the speaker when judging whether the sentences were produced by the same person.

The speaker discrimination portion of Exp. I consists of two parts: one with five Finnish speakers, one with five English speakers. Each part contains 90 trials (i.e. 180 sentences in total). English sentences were selected from a set of 40 news sentences ranging in length from 7 to 10 words. The two sentences within a trial were always different. Each speaker was presented in combination with every other speaker. Within each test there were 45 matched trials and 45 mixed trials, and 45 same trials and 45 different speaker trials. The matched trials were matched in terms of speech type. There were five matched speech type conditions: natural speech and four synthetic speech variants. The four synthetic variants 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: “N” = Natural, “S” = Synthetic, “A” = American,“F” = Finnish and “5”/“15” = number of adaptation sentences. This results in the synthetic variants: SA5, SA15, SF5 and SF15. The mixed trials include a natural speech stimuli and one of the four synthetic stimuli variants.

Exp. II also consisted of two parts: Finnish speakers and English speakers. Each part comprises 80 trials (i.e.160 sentences in total). The focus of this test was to compare different categories of synthetic stimuli. The comparisons are between SA5 and SF5, and SA15 and SF15. Within each test there were 40 matched trials and 40 mixed trials, 40 same speaker trials and 40 different speaker trials.

3.2. Accent Classification Task

In addition to the speaker discrimination task, we designed an accent classification task. In this task, for each speaker one sentence for each of the speech types was selected. Listeners were asked to listen to the sentences and judge whether they thought the accent was mainly American, British or Scandinavian. [1]

[1] 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.

For both Exp. I & II, twenty native English listeners with no known hearing speech and language problems were recruited at the University of Edinburgh (i.e. 40 in total). Each listener completed one of the discrimination tests (English or Finnish from Exp. I or Exp. II) and the accent classification test. On average it took 30 minutes to complete. Due to incompleteness in obtained responses, only eighteen responses could be analysed in Exp. I

4. Results

For both Exp. I discrimination tests, the results of nine listeners are analysed. Individual listener data were pooled for both tests for all speakers. Figure 1 shows boxplots of percent correct per speech type pair. For example, "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 analysis of variance (ANOVA) was conducted with speech type (natural or synthetic) as the within-test factor. The ANOVAs show a significant main effect of speech type: Finnish speakers $[F(8,72) = 7.63, p < 0.001]$ and English speakers $[F(8,72) = 6.75, p < 0.001]$. Tukey HSD (Honestly Significant Difference) multiple comparisons of means with 95% family-wise confidence level were conducted to analyze the effect of speech type in more detail. The Tukey HSD test revealed that listeners perform significantly worse when comparing synthetic speech to natural speech than when the speech types are the same (either synthetic or natural).

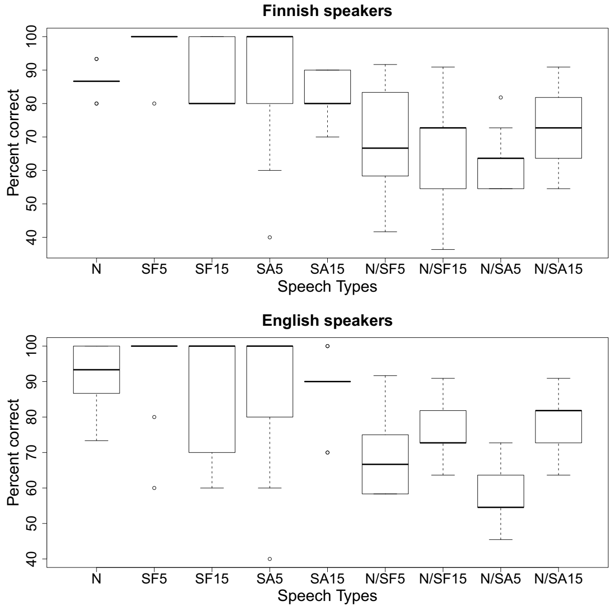

Figure 1: Exp. I – Percent correct discrimination per speech type pair. The median is indicated by a solid bar across a box which shows the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented by circles.

To find out if listeners actually perceive different accents for synthetic speech created using either Finnish-accented or American-accented average voice models an accent classification test was carried out. Figure 2 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 reflect 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.

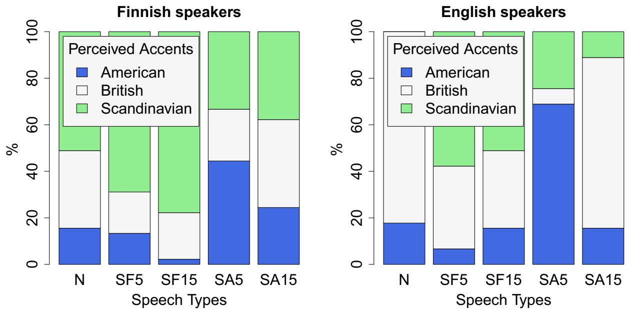

Figure 2: Accent classification of natural and various synthetic sentences for Finnish and English speakers.

For each of the discrimination tests in Exp. II, the results from ten listeners are presented. Individual listener data were pooled for both tests for all speakers. Figure 3 shows the results of judging speaker identity for trials that consist of different variants 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 & SF5, and SA15 & SF15.

ANOVAs with speech type (SF5, SF15, SA5, SA15, SA5/SF5 and SA15/SF15) as the within-test factor were conducted. The ANOVAs showed a significant effect of speech type: Finnish speakers $[F(6,63) = 5.84, p < 0.001]$ and English speakers $[F(6,63) = 6, p < 0.001]$. A TukeyHSD test shows that the significant differences for Finnish speakers are between SA5 and most other speech type pairs. For English speakers, listeners score significantly lower on SA5 trials than on SA15, SF15 and SA15/SF15 trials and significantly lower on SA5/SF5 trials than on SA15 and SF15 matched trials. Basically this means that using only five adaptation sentences when the average voice model is American-accented is affecting the listeners' ability to identify speakers in a negative way. However, comparing the mixed trial data to the matched trial data in Figure 3 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.

A TukeyHSD test comparing the synthetic matched trial data from Figure 1 to the matched trial data in Figure 3 shows that the only significant differences in percent correct across the two experiments is between SA5 (Figure 3) and SF5 and SF15 (Figure 1) for Finnish speakers and between SA5 (Figure 3) and SF5 (Figure 1) for English speakers.

Combining the information from Figures 2 and 3 we can conclude that listeners correctly classify a speaker as themselves even when their synthetic speech is of different accent types.

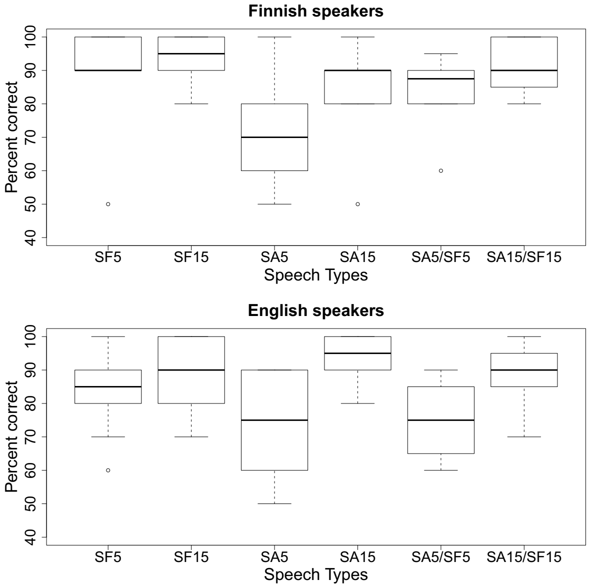

Figure 3: Exp. II: Percent correct discrimination per speech type pair. Comparisons across synthetic speech variants.

5. Discussion & Conclusions

Our first question was what the effect of limited amounts of adaptation data on listeners' ability to discriminate between speakers is. The speaker discrimination task shows that using fifteen sentences leads to slightly better results than using five sentences, but the differences are quite small and we only found a significant difference in listeners' performance for the SA5 condition in Exp. II. Comparing the percentages correct for the Finnish speakers (Figure 1) to the results in [4] shows us a very similar picture. In [4] the results showed that when listeners only compared different synthetic speech variants the average scores were (roughly) between 80 and 90% correct and when it was a comparison between synthetic and natural speech the scores dropped to between 60 and 80% correct. Here, we find that listeners score synthetic stimuli as well as natural speech, and comparisons between natural and synthetic speech result in scores between 60 and 70% for Finnish speakers, and between 60 and 80% for English speakers.

The second question we set out to answer was whether using Finnish- or American-accented average voice models would affect listeners' speaker discrimination performance. Figure 2 clearly shows there is an effect of using Finnish- and American-accented average voice models on the output synthetic speech. The accent of the average voice affects the perceived accent of synthetic stimuli. This effect is especially pronounced for very small amounts of adaptation data (five sentences) but still also quite clear for fifteen adaptation sentences. However, we did not find evidence to support our assumption that a person's accent in a foreign language will influence the perception of their identity (in synthetic speech). Figure 1 showed us that Finnish-accented synthetic speech does not translate into a higher percentage correctly classified Finnish speakers.

The results from the second listening test (Figure 3) in combination with the accent classification results (Figure 2) clearly show that listeners do indeed classify synthetic speech with two different perceived accents as belonging to the same speaker. Or, in other words, listeners do not base their speaker similarity decisions on perceived accent. Although accent is tightly coupled with speaker identity [11] it is not unusual for an individual to use more than one accent (for example, a regional accent and a more general standard accent) [12, 13]. Listeners, as a consequence, will probably encounter individuals using more than one accent on a regular basis and will therefore be accustomed to this phenomenon to a certain extent [14]. This may explain why, in this study, listeners correctly ascribe synthetic speech samples with two different perceived accents as belonging to the same speaker.

6. Acknowledgements

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

7. References

[1] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," in *Proc. Interspeech '06*, 2006.

[2] 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*, vol. 48, pp. 1227–1242, 2006.

[3] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. 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.

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

[5] M. Wester, "Cross-lingual talker discrimination," in *Proc. Interspeech '10*, 2010.

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

[7] 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 Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66 –83, Jan. 2009.

[8] 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, 2007.

[9] J. Yamagishi and O. Watts, "The CSTR/EMIME system for Blizzard Challenge 2010," in *Proc. Blizzard Challenge*, Kyoto, Japan, 2010.

[10] M. Wester, "The EMIME Bilingual Database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1388, 2010.

[11] N. Mendoza-Denton, "Language and identity," in *The Handbook of Language Variation and Change*, J. Chambers, P. Trudgill, and N. Schilling-Estes, Eds. Blackwell Publishing, 2002, pp. 475–499.

[12] P. Howell, W. Barry, and D. Vinson, "Strength of British English accents in altered listening conditions," *Perception and Psychophysics*, vol. 68, no. 1, pp. 139–153, 2006.

[13] W. Labov, *Sociolinguistic Patterns*. Univ. of Pennsylvania Press, 1972.

[14] A. Ikeno and J. Hansen, "Perceptual recognition cues in native English accent variation: "Listener accent, perceived accent and comprehension"," in *Proc. of ICASSP*, 2006, pp. 401–404.