SPEAKER SIMILARITY EVALUATION OF FOREIGN-ACCENTED SPEECH SYNTHESIS USING HMM-BASED SPEAKER ADAPTATION

Mirjam Wester

CSTR, University of Edinburgh, UK

Reima Karhila

Aalto University, Helsinki, Finland

ABSTRACT

This paper describes a speaker discrimination experiment in which native English listeners were presented with natural and synthetic speech stimuli in English and were asked to judge whether they thought the sentences were spoken by the same person or not. The natural speech consisted of recordings of Finnish speakers speaking English. The synthetic stimuli were created using adaptation data from the same Finnish speakers. Two average voice models were compared: one trained on Finnish-accented English and the other on American-accented English. The experiments illustrate that listeners perform well at speaker discrimination when the stimuli are both natural or both synthetic, but when the speech types are crossed performance drops significantly. We also found that the type of accent in the average voice model had no effect on the listeners' speaker discrimination performance.

Index Terms— Similarity Evaluation, Speaker Adaptation, HMM-synthesis

1. INTRODUCTION

The motivation for this study arose from the EMIME speech-to-speech translation (S2ST) task. 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[1]. However, how do we measure whether our modeling attempts are successful or not? That is, how are we to measure whether or not a user sounds similar in two different languages? 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?

Almost all previous studies, for example the S2ST project TC-STAR [1], work by Latorre and colleagues [2] as well as the EMIME project [3] evaluate the success of cross-lingual adaptation or multilingual 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 [4, 5].

In our evaluation experiments, our end goal is to understand the psychoacoustic processes underlying listeners' judgements of speaker similarity in an S2ST system. Before we can look at these processes, however, there are two main issues that we need to consider: 1) How well can listeners judge speaker identity across languages? and 2) How well can listeners judge speaker identity when comparing synthetic speech to natural speech?

The first issue was addressed in a previous study [6] in which we investigated how well listeners are able to recognize speakers across language boundaries (German-English and Finnish-English) when the stimuli consist of natural speech. Those 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.

The main focus of this paper is to address the second issue: how do listeners perform in a discrimination task when asked to compare synthetic speech to natural speech? In order to be able to do this, we disregard the cross-lingual element of S2ST evaluation by restricting the current experiment to discrimination within one language – English. Difficulties associated with comparing synthetic to natural speech have been discussed in detail in [7]. 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.

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 [8]. 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. For example, native English data can be used to create an average voice model which is then adapted using foreign-accented speech, or alternatively foreign-accented English data can be used to create an average voice model which is then also adapted using foreign-accented speech. The second question we want to answer in this study is: Does using different average voice models affect listeners' ability to discriminate between speakers?

2. 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 speaking style and phonetic features of the synthetic speech [9, 10]. As a foreign accent can

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

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

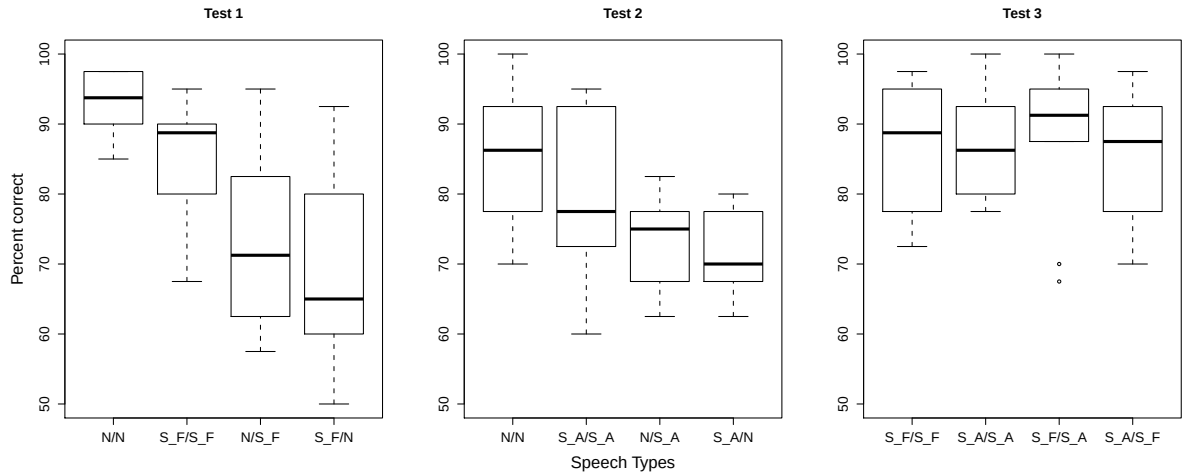

Fig. 1. Percent correct judgements per speech type pair for the three discrimination tasks. N = Natural speech, S = Synthetic speech, _A = American-accented average voice model, _F = Finnish-accented average voice model. The median is indicated by a solid bar across a box which shows the quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented by circles.

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.

2.1. 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.

Finnish-accented English was recorded at Aalto University. In total, 1297 sentences from 6 female and 62 male native Finnish speakers were recorded in a classroom setting. Each speaker read aloud about 20 sentences (some speakers missed out some of the prompts). The sentences were randomly selected from a collection of 200 sentences from the Herald database, 25 sentences from the European Parliament corpus (selected for simplicity and phonetic coverage) and 100 WSJ0 sentences from enrolment and language model test sets (selected for phonetic coverage). Due to the recording conditions, the data was quite noisy. It was cleaned up using channel separation, i.e., by separating speech from noise using a codebook trained from WSJ and Finnish Speecon corpora [11].

The American-accented English data was selected from the WSJ0 si-et-05 set. In order to be able to create average voices of comparable quality, only a subset of the si-et-05 set was used. 42 male speakers and 3 female speakers were included leading to a total of 1223 sentences. These sentences were selected to maximise phonetic coverage.

The 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 [12]. 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. The resulting model sets have very distinct differences in stress and pronunciation style.

2.2. Speaker Adaptation

The data used to adapt the speaker independent models were recorded at the University of Edinburgh [13]. Five male native speakers of Finnish who were also fluent in English recorded sentences in both English and Finnish. In this study, we only used the English speech. The English data includes 25 sentences from the European Parliament corpus, 100 WSJ0 sentences and 20 semantically unpredictable sentences. This data was separated into an adaptation set of 105 sentences and a test set of 40 sentences for each speaker.

The two average voices were adapted to the five University of Edinburgh male speakers, using a set of 105 English sentences. Simultaneous transformation of the cepstral, log F_0 and duration parameters was carried out using CSMAPLR adaptation [9]. To synthesize the 40 test sentences, for each sentence an excitation signal is generated 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

In order to evaluate speaker similarity we designed a discrimination task and a MOS-style rating task. The discrimination task 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

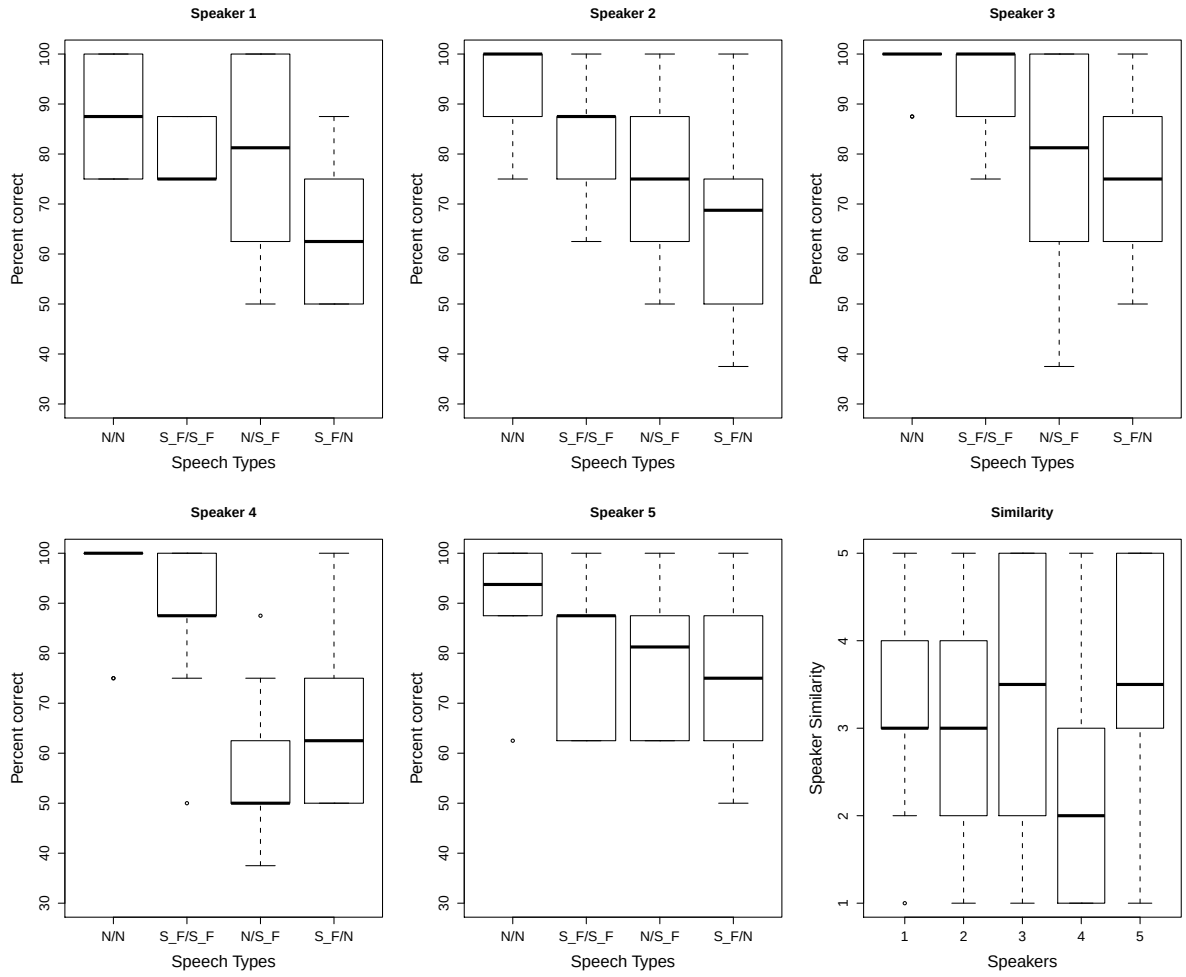

Fig. 2. Percent correct per speech type pair for individual speakers and similarity scores comparing to original speaker. (Fig. 1 more details.)

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. Each test consists of 160 trials (i.e. 320 sentences in total). 40 news sentences, ranging in length from 7 to 10 words, were used per test condition. Each sentence occurred eight times; twice in same-speaker trial (N), twice in different-speaker trial (N), twice in same-speaker trial (S) and twice in different-speaker trial (S). The sentences within a pair were always different. Five speakers were included in the experiment. Each speaker was presented in combination with every other speaker twice and counterbalanced for order. We also ensured there were equal amounts of matched and mixed speech type conditions.

Listeners were asked 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 MOS-style rating task consisted of one test in which 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". All listeners judged all five speakers. In each trial, the natural reference stimulus was played first followed by the synthetic stimulus.

Thirty native English listeners with no known hearing speech and language problems 20-30 years of age, were recruited at the University of Edinburgh. Due to time constraints, each listener completed only one of the discrimination tests and the whole MOS task, which on average took 40 mins to complete.

4. RESULTS

Each discrimination test was done by ten listeners. Individual listener data were pooled for each test for all speakers. Figure 1 shows boxplots of percent correct per speech type pair for each of the three discrimination tests. An analysis of variance (ANOVA) was conducted with speech type (N/N, S/S, N/S, and S/N) as the within-test factor. The ANOVA for Test 1 shows there is a significant main effect of speech type $[F(3, 36) = 11.73, p = 1.65 \times 10^{-5}]$, and for Test 2 it is $[F(3, 36) = 5.29, p = 0.004]$. 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 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 2 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.

5. DISCUSSION & CONCLUSIONS

Our first question was: how do listeners perform in a discrimination task when asked to compare synthetic speech to natural speech? It seems the difficulties listeners have in comparing synthetic speech to natural speech, described in the Introduction, are definitely playing a role in the current study. The results show that when listeners are only comparing different types of synthetic speech the average scores are (roughly) between 80 and 90% correct. However, when it is a comparison between synthetic and natural speech the scores drop to between 60 and 80% correct. These results for comparing different speech types is similar to what we found for cross-lingual speaker discrimination but more pronounced. Overall it can be concluded that when listeners are faced with mixed condition trials (i.e., across languages or across speech types) they perform significantly worse than in matched conditions.

The scores for speaker similarity, shown in the bottom right-hand corner of Figure 2, range between 2 and 3.5. Disregarding Speaker 4, these results are as high as (if not higher than) scores found in other papers on S2ST [1, 9, 14]. The figure also shows a clear relationship between the MOS-task results and the discrimination task results. But, as is apparent from the discrimination data, MOS scores 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 should not solely rely on MOS to draw conclusions about speaker similarity.

To answer our second question, it seems that listeners judge speakers across average voice conditions as well as (or even better than) within average voice conditions (see rightmost boxplot in Figure 1). This finding was somewhat surprising to us, as we expected that the accent in the average voice model would have a larger effect. We conclude that the number of adaptation sentences (105) is large enough to overrule any influence of accent in the average voice model. However, we can't rule out that listeners may be identifying the same speaker but with two different accents. In ongoing work, we are investigating using more limited amounts of adaptation data (5-15 sentences). A discrimination task will show if either of the average voice models is more suitable when using small amounts of adaptation data, and accent rating tests will also be carried out.

To conclude, in the S2ST scenario, we have to contend with both cross-lingual and cross speech type comparisons. The speaker discrimination study in [6] showed us that when comparing stimuli across languages, listeners' performance drops on average 10 percentage points, from 90-100% correct to 80-90% correct. The discrimination study presented here shows that the same is occurring when comparing stimuli across speech types (synthetic versus natural), but to a larger degree: the drop is closer to 20%. The question that remains to be investigated is how well listeners are able to judge speaker identity when they have to deal with stimili pairs that cross both language and speech type boundaries.

6. 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.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.

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

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

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

[7] S.J. Winters and D.B. Pisoni, "Speech synthesis, perception and comprehension of," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed., pp. 31–49. Elsevier, 2005.

[8] J.E. 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.

[9] 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.

[10] 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.

[11] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. Interspeech '10*, 2010.

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

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

[14] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proc. ICASSP*, 2010.