Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project

Mirjam Wester[1], *John Dines*[2], *Matthew Gibson*[4], *Hui Liang*[2], *Yi-Jian Wu*[5], *Lakshmi Saheer*[2], *Simon King*[1], *Keiichiro Oura*[5], *Philip N. Garner*[2], *William Byrne*[4], *Yong Guan*[6], *Teemu Hirsimäki*[3], *Reima Karhila*[3], *Mikko Kurimo*[3], *Matt Shannon*[4], *Sayaka Shiota*[5], *Jilei Tian*[6], *Keiichi Tokuda*[5], *Junichi Yamagishi*[1]

[1] University of Edinburgh, UK, [2] Idiap Research Institute, Switzerland, [3] Aalto University, Finland,
[4] University of Cambridge, UK, [5] Nagoya Institute of Technology, Japan,
[6] Nokia Research Center Beijing, China

mwester@inf.ed.ac.uk

Abstract

This paper provides an overview of speaker adaptation research carried out in the EMIME speech-to-speech translation (S2ST) project. We focus on how speaker adaptation transforms can be learned from speech in one language and applied to the acoustic models of another language. The adaptation is transferred across languages and/or from recognition models to synthesis models. The various approaches investigated can all be viewed as a process in which a mapping is defined in terms of either acoustic model states or linguistic units. The mapping is used to transfer either speech data or adaptation transforms between the two models. Because the success of speaker adaptation in text-to-speech synthesis is measured by judging speaker similarity, we also discuss issues concerning evaluation of speaker similarity in an S2ST scenario.

Index Terms: speech-to-speech translation

1. Introduction

EMIME (Effective Multilingual Interaction in Mobile Environments) is a European FP7 project concerned with speech-to-speech translation (S2ST)[1]. The main goal of EMIME is to develop a mobile device that performs personalised 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.

The idea for the EMIME project arose out of recent advances in hidden Markov model-based speech synthesis. In recent years, speech synthesis systems based on hidden Markov models (HMMs) have reached performance levels comparable to state-of-the-art unit selection systems [1, 2]. This has sparked interest in whether unified models for both recognition and synthesis are possible. One of the motivations for pursuing such unified models is the success of speaker adaptation techniques, developed for automatic speech recognition (ASR), in text-to-speech (TTS) synthesis, including for unsupervised speaker adaptation [3]. Figure 1 gives an overview of the EMIME system.

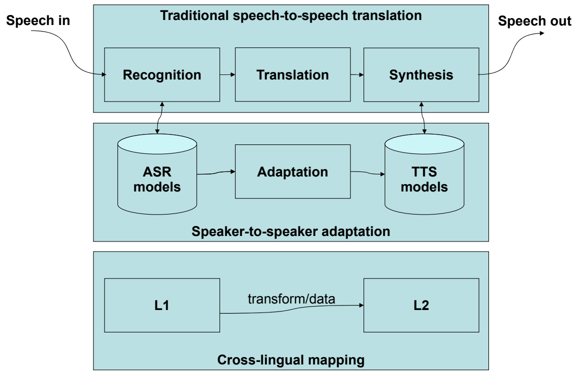

Figure 1: Overview of the system.

However, despite a common HMM statistical framework there remain substantial differences between HMM-based ASR and TTS. This should not be surprising: ASR is concerned with minimising speaker (and environment) specific effects, aiming to be as general as possible. This is characterised by systems that use as few parameters as possible, smearing out speaker differences. By contrast, synthesis is concerned with producing intelligible speech, that sounds like a specific speaker, and that is as natural as possible. This is characterised by parametrically rich systems that model specific speaker traits.

We have explored the gap between ASR and TTS [4], and concluded that many of the techniques or configurations used in ASR (or TTS) cannot simply be used in TTS (or ASR) without negative consequences. In particular, a single acoustic feature type and order that worked well for ASR and TTS could not be found. We have also investigated the transfer of adaptation transforms from ASR models to TTS models within one language – which enables unsupervised speaker adaptation for TTS – with good results. Initial experiments on cross-language speaker adaptation have demonstrated the feasibility of the approach and recently this has been extended to multilingual acoustic modelling. When describing cross-lingual experiments in this paper, we will refer to the *input language* (*L1* – the language of the adaptation data) and the *output language* (*L2* – the language of the synthesized speech); we will avoid the terms *source language* and *target language* since these may be ambiguous in some situations.

An important issue that arises when speaker adaptation is carried out across languages is whether or not a speaker is recognisable as him/herself in the output language (L2). A

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

few experiments have been conducted to explore this issue, but many questions are still open – the issues surrounding the evaluation of speaker similarity are discussed towards the end of the paper.

2. Basic system

This section gives a brief description of the general set-up of the EMIME system. We are targeting four languages: English, Finnish, Japanese and Mandarin.

2.1. ASR

The initial work of the project has included the development of baseline speech recognition systems for all languages. These are broadly similar in design, being based on HMMs and using large word or sub-word vocabularies. The Hidden Markov Toolkit (HTK), one of the *de-facto* standards used for training HMMs for ASR was used, as well as HTS (H triple-S, or HMM speech synthesis system) a significant set of patches for HTK that allow it to train models suitable also for synthesis. The acoustic models are speaker-independent, but they can be adapted to new speakers. The baseline systems differ in the acoustic features extracted from the speech signal, the set of phonemes used in acoustic models, and the construction of the n-gram language models [5].

2.2. Machine translation

In the first instance, general machine translation architectures are used in EMIME since the focus is on recognition, synthesis and cross-lingual speaker adaptation. In the demonstration S2ST system, existing systems and resources [6] are used to translate from L1 to L2. However, there are interesting challenges in speech translation that will be addressed. For example, in Finnish-to-English translation there are complex issues in reconciling morphological analyses used in translation with those used in ASR. For statistical machine translation, the role of analysis is to make it easier to map Finnish to English whereas in ASR the role is to make it easier to build and integrate language and acoustic models. Rather than try to find a Finnish morphological analysis scheme which is ideal for both ASR and statistical machine translation (SMT), combination procedures which allow for considering multiple analyses in translation have been developed: [7] found that translation performance and robustness is improved.

2.3. TTS

The TTS systems in EMIME are built using the framework from the HTS-2007 system [3], which is a speaker-adaptive system that was entered for the Blizzard Challenge 2007. The HTS-2007 system consists of four main components: speech analysis, average voice training, speaker adaptation and speech generation. Speech analysis generates an acoustic feature vector comprising high-order STRAIGHT-analysed Mel-generalised cepstral coefficients (MGCEP), fundamental frequency and a 5-band representation of the aperiodic component of the signal. The “average voice model” approach to speaker adaptation is used in which HMMs are trained on multiple speakers’ data using speaker adaptive training (SAT) [8]. These HMMs are then adapted to a single target speaker’s data. To synthesise an utterance, an excitation signal is generated using mixed excitation and PSOLA and then a synthesised waveform is generated using the MLSA filter corresponding to the STRAIGHT mel-cepstral coefficients.

2.4. Evaluation

ASR performance is calculated using the standard word error rate (WER) measure. TTS performance is measured through formal listening tests as well as a number of objective measures which are used to pre-select systems for listening evaluation. The key properties of synthetic speech that are evaluated are: naturalness, intelligibility and similarity to the original speaker. To evaluate naturalness and similarity to the target speaker, 5-point mean opinion score (MOS) scales are used. To evaluate intelligibility, the subjects are asked to transcribe semantically unpredictable sentences; the average WER is calculated from these transcripts. Evaluation is discussed in Section 4.

2.5. Data

The main data sets used for ASR and TTS are the following. For English it is the Wall Street Journal corpus (WSJ) [9] and CMU-ARCTIC (`http://festvox.org/cmu_arctic/`). For Finnish and Mandarin Speecon [10] was employed, and for Japanese the Japanese Newspaper Article Sentences (JNAS) database [11] and a 6-speaker corpus from Nitech were used.

3. Speaker adaptation

All of the speaker adaptation experiments conducted in EMIME can be described within a common framework. This framework involves the construction of a mapping to link two acoustic models (whether that is an ASR model and a TTS model for the same language, or models for different languages). Once the mapping is constructed, it is used to adapt one of the models by either a) learning the adaptation using the other model, then transferring the adaptation transforms through the mapping, or b) transferring data associated with one model through the mapping, and then using it to learn the adaptation transform for the other model. This is illustrated in the bottom tier of Figure 1.

The mapping can be formed in a number of ways, linking either acoustic model states or linguistic units in the two models (Section 3.2). Since our goal is unsupervised adaptation (i.e., the only external input to the system is a speech waveform), the adaptation data must be automatically transcribed (Section 3.3). Once a transcription has been obtained, adaptation transforms are learned. Techniques used for adaptation are derived from the Maximum Likelihood Linear Regression (MLLR) and *maximum a posterior* (MAP) methods [8]. The mapping can be thought of of having a root node where vocal tract length normalisation (VTLN) sits and leaves where the MLLR transforms are. We summarise VTLN (Section 3.1) and then intra-lingual (Section 3.4) and cross-lingual (Section 3.5) experiments.

3.1. VTLN

Vocal tract length normalisation (VTLN) is a simple adaptation technique based on the facts that different people have different sized vocal tracts, and that vocal tract lengths are directly related to formant frequencies. VTLN warps the feature spectrum. Whilst it is not capable of adapting to the extent available from MLLR, it can be done using much fewer data; as little as a few tens of seconds worth. VTLN also fits well with the MGCEP feature extraction used in HTS-based synthesis.

In the framework presented here, VTLN forms the root of the adaptation hierarchy, being class agnostic. It is basically a feature transform which effectively provides a prior (initialisation) for the more powerful adaptation techniques. It was shown in [12] that VTLN can quickly produce a synthesised output voice matching rough gender characteristics of the input voice, and sounding more natural than other transforms for small amounts of adaptation data. It was also confirmed to have additive effects with respect to MLLR.

3.2. Linking two acoustic models via a mapping

Two acoustic models, each comprising a set of context-dependent models of sub-word units (e.g., phones), can be linked by a mapping which associates one or more parameters in one model with one or more parameters in the other model. Mappings can be defined in terms of acoustic model states or linguistic units. The intra-lingual experiments discussed in this paper used either phoneme mappings constructed using expert knowledge or state mappings derived from state-clustering decision trees (Section 3.4). The cross-lingual experiments used either phoneme mappings constructed using expert knowledge or state mappings learned automatically using the Kullback-Leibler (KL) divergence between the state's probability density functions (Section 3.5).

3.3. Transcribing the adaptation data

In order to carry out unsupervised speaker adaptation, we need to automatically obtain a transcription of the adaptation speech waveforms. This transcription might be words, phones (or triphones), or "full context" labels. "Full context" is the term used to refer to the type of models used for HMM-based synthesis, which are typically quinphones (phones conditioned on two left and two right phones of context) conditioned on suprasegmental prosodic features such as syllable structure, word and phrase boundaries, etc. These full context transcriptions are usually predicted from the word transcription by using the front-end of the TTS system, which predicts the segmental and suprasegmental information using a variety of methods (pronunciation dictionary, letter-to-sound model, intonation model etc.).

An obvious way to obtain a transcription is by using an ASR system. A TTS front-end can be used to generate full context labels from the words. A potential drawback with this approach is that errors present in the word level transcription might cause wide-ranging errors in the full-context labelling, because of the wide context taken into account in these labels and in the TTS front end. Phone labels can be produced easily using an ASR system, either by deriving them from the words, or by running the ASR models as a phoneme recogniser. Another way to transcribe the data is to decode the speech using TTS-style full context acoustic models. However, this is infeasible as a first-pass decoding due to the very long context dependencies in these models and the consequently vast decoding network. The explicit duration model used in TTS models also adds significant complexity. These problems can be worked around, by limiting the span over which context constraints are obeyed in the decoding network, or by running a multi-pass system.

All of the above methods assume that we are attempting to transcribe the input (L1) speech using L1 labels (words, phones, etc), prior to either learning adaptation transforms or sending the data through the mapping. It is also possible to directly obtain an L2 transcription of the input speech, by running an L2 ASR system on the data (e.g., run Mandarin ASR on English speech waveforms). Of course, it will not produce meaningful output, but the sub-word labels may still be used to learn adaptation transforms directly with respect to the L2 model.

3.4. Intra-lingual speaker adaptation

In EMIME, intra-lingual adaptation is not really the target. However, it formed a very useful means for baseline experiments, especially before multilingual data was available. Generally, a technique that does not work on intra-lingual adaptation is unlikely to work on cross-lingual adaptation. In particular, it was of interest whether adaptation would work on ASR data, and (hence) would be suitable for unsupervised adaptation in TTS. This section describes this background and baseline work, and re-interprets it in terms of the common framework introduced earlier. King et al. [13] reports the first (intra-lingual) unsupervised speaker adaptation experiments, which used a trivial mapping between full context and triphone models. In [14] and [15], we reported two approaches in which the mapping was derived from the decision trees used to perform state-clustering for both the ASR model and the TTS model.

King et al. [13] used a simple phone recogniser, based on triphone acoustic models and a bigram phone language model, to transcribe the adaptation data. These ASR acoustic models used TTS-type acoustic features (high order spectral features + source features) and were derived from the full context synthesis models by untying all states, then reclustering them into triphone models. After transcribing the adaptation data in terms of triphone labels, adaptation transforms were learned with respect to the ASR acoustic models. The construction of the mapping was trivial: each full context synthesis model was mapped to the corresponding triphone. Transferring the adaptation transform through this mapping resulted in adapting the TTS full context models using these transforms.

Gibson [14] used a two-pass decision tree construction method which allowed the same set of underlying models to be used to transcribe the adaptation data and to generate adapted synthetic speech. Pass 1 only asks questions about left, right and central phonemes to construct a phonetic decision tree. The tree is used to generate a set of tied-state triphone models which are used for ASR. Pass 2 extends the tree by asking additional questions about suprasegmental information, to arrive at full context models for TTS. The mapping between full context TTS models and triphone ASR models is defined by the decision tree: every full context model has, somewhere further towards the tree's root, an ancestor triphone model.

The adaptation data are transcribed using triphone models. These ASR models are adapted and the transforms are mapped to the full context models via the tree. This method was compared to a second approach in which the word transcription is expanded to full context labels using the TTS front end. These labels are then used to adapt the full context models. The mapping in the second approach is defined by the prediction of full context labels from the word sequence, and it is the adaptation data that is being transferred via this mapping to the TTS model.

Gibson [14] found that, with regard to similarity to the target speaker, significant improvements over an average voice system were observed for all adapted systems. No significant performance degradation or improvement was observed when using direct adaptation of the full context models compared to triphone model adaptation followed by transfer of the transforms via the mapping.

Dines et al. [15] focused on how to share parameters between ASR and TTS models. Like Gibson [14], the approach uses the state clustering tree to define a mapping between ASR and TTS models. Dines et al. [15] generated the ASR model by starting with a trained TTS model, then marginalising over the leaves of the state clustering decision tree. ASR results show some degradation (compared to a conventionally-built baseline) as a result of the decision tree marginalisation procedure. Since the ASR and TTS models share parameters, it is then possible to perform intra-lingual speaker adaptation by using the adaptation transforms generated during ASR.

3.5. Cross-lingual speaker adaptation

Ultimately the goal of EMIME is to perform unsupervised speaker adaptation *across languages*: what we are seeking to achieve is a system where the synthesised speech at the output in L2 sounds like it was spoken by the same person who spoke the L1 input. The mapping between the ASR and TTS models must now take differences in the languages (such as their phone inventory) into account. A number of experiments have been conducted on this so far. In the first two studies the correct labels for the input data were known (i.e., a supervised setting) and the mappings were based on either manually-constructed cross-language phone mappings or a learnt state-level mapping. In more recent work [16, 17, 18], unsupervised cross-lingual speaker adaptation has been investigated.

3.5.1. Supervised

In the first series of experiments, Wu et al. [19] performed Mandarin to English speaker adaptation. Two different phonetic label mapping schemes were employed, both manually constructed: a one-to-one mapping and a one-to-sequence mapping. The Mandarin phone labels for the Mandarin adaptation data were mapped to English phone labels, using one of the mapping schemes. The mapping was in terms of linguistic units, and it was speech data that was transferred between the models via this mapping. Once the Mandarin speech had been transcribed using English labels, adaptation transforms for a set of English triphone models were estimated, which were then transferred to the English full context TTS models using the method from King et al. [13] described above.

In a subsequent paper [20], Wu et al. introduced a state-based mapping for cross-lingual speaker adaptation, this time for English and Japanese. The mapping is learned by examining the KL divergence between each state in an L1 average voice model and each state in an L2 average voice model. The mapping can be used to transfer either transforms or data from L1 to L2. In the case of transforms, these are estimated using L1 data with respect to the L1 model, then transferred via the mapping and used to adapt the L2 model. In the case of data, the L1 data is aligned with the L1 model, the result of which is to associate frames of data with states in the L1 model – a 'state transcription' of the data. These L1 states are mapped to L2 states, with the result that frames of L1 data are now associated with L2 states. Adaptation transforms can now be estimated with respect to the L2 model.

In a final experiment, using Japanese adaptation data manually transcribed with full context labels, the state mapping approach outperformed the phone mapping approach [20] and, in terms of speaker similarity, the data sharing method resulted in higher mean opinion scores than transform sharing, although transform sharing was better in terms of naturalness.

Mapping at the state-level outperforms knowledge-based phone-level mapping, presumably because it is finer grained and because it is learned from data. In state mapping, it is more likely that acoustically-similar units in L1 and L2 will be mapped onto one another. One disadvantage of the state-level mapping is that average voice models for the two languages are needed. The approach also assumes that the model space in L1 is similar to the model space in L2, which is not necessarily the case, particularly if the training data are from different corpora. This may be why the transform mapping approach degrades speaker similarity.

Our preliminary conclusions regarding the relative merits of mapping transforms or data are that transform mapping minimises the chance that the L2 output will be "L1 accented", but at the cost of the speaker identity characteristics being less strong, so perceived speaker similarity suffers. On the other hand, data mapping can achieve better speaker similarity but at the cost of imposing an L1 accent on the L2 output.

3.5.2. Unsupervised

In order to perform unsupervised cross-lingual speech adaptation a transcription of the adaptation data is needed. Oura et al. [16] extended supervised adaptation using the state level transform mapping [20] to unsupervised adaptation by automatically transcribing the adaptation data using ASR HMMs. All acoustic models were trained on ASR databases both for ASR and TTS. Listening tests showed that the adapted speech was a bit more similar to the original speaker than the average voice model.

Gibson's two-pass decision tree construction method was extended from intra-lingual to cross-lingual speaker adaptation by treating the L2 adaptation data as if it were uttered in L1 [18]. Listening tests show that there is no significant difference in similarity nor naturalness for any of the tested systems. Both intra-lingual and cross-lingual unsupervised adaptation deliver performance approaching that of supervised adaptation.

Liang et al. [17] combined decision tree marginalization [15] and state level mapping [20] using both data and transform mapping to perform unsupervised cross-lingual speaker adaptation. An important finding in [17] is that the difference between similarity scores for supervised and unsupervised adaptation is larger when the reference utterance and test utterance are in the same language. A language mismatch, i.e., reference utterance is Mandarin, test utterance is English, leads to listeners judging both systems (supervised and unsupervised) as equally similar/dissimilar to the original speaker. There is also no difference between data and transform mapping in the mismatched condition whereas transform mapping outperforms data mapping in a matched language condition. Furthermore, in contrast to [20] Liang et al. [17] found that transform mapping resulted in lower naturalness scores than data mapping.

The three approaches described here all show that unsupervised cross-lingual adaptation achieves comparable results - in terms of similarity and naturalness - to supervised cross-lingual speaker adaptation. However, similarity scores are all around 2 on MOS scales ranging from 1-5 indicating that although adaptation does make the synthesis sound more similar to the original speaker it is not yet recognisable as the original speaker. Naturalness scores range from 2 - 3. Section 4 discusses the issues that exist with these similarity scores and alternatives are suggested.

4. Speaker similarity evaluation

We now discuss issues of evaluation, focusing on how to measure the success of cross-lingual speaker adaptation. Measuring speaker similarity in a meaningful way is obviously a key aspect of the evaluation of any type of voice conversion or speaker-adaptive text-to-speech synthesis. However, a survey of previous cross-lingual voice conversion research shows that none of the studies give a precise explanation of what is being measured. It also appears that no current techniques achieve what we could call "good" cross-lingual speaker similarity.

4.1. Voice conversion literature

Research in voice conversion, including across languages, has a longer history than cross-language speaker-adaptive HMM-based text-to-speech synthesis, so it is worthwhile surveying the literature in this area to see what it can tell us about evaluation. [21] used bilingual data (Japanese/English) and measured similarity by calculating mutual information between speaker pairs. [22] also used bilingual data (Japanese/English) and used the objective measure Mel Cepstral Distortion (MCD) to evaluate speaker individuality. In the S2ST project TC-STAR [23] data from monolingual speakers was used in a unit selection system. Evaluation was carried out using mean opinion scores (MOS) for similarity and quality. The work of Latorre and colleagues [24] has a slightly different focus: multilingual synthesis, which is the ability to generate utterances in more than one language, or utterances of mixed language, from a single system. They also use MOS, for intelligibility, similarity and native accent.

A common technique, used in several of these studies, is to compare cross-lingual voice conversion to intra-lingual voice conversion. However, this does not directly measure how similar the speech sounds to that of the original speaker. Using mean opinion scores to evaluate similarity, although a widely-used technique, is not without problems: judging how similar utterances are on a scale from 1 to 5 may be too difficult for listeners, especially if the utterances are in different languages. The results in [17] support this. Judgements of speaker similarity are also strongly correlated with the overall quality or naturalness of the synthetic speech: listeners are probably unlikely to rate an utterance as sounding like the target speaker if the quality is poor.

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

4.2. Re-evaluation of speaker similarity

Clearly, more research focussed on the evaluation of speaker similarity is needed. The research needs to address two important questions: does a speaker actually sound the like the same person in L1 and L2, and can listeners judge speaker similarity across languages? Our own impressions are that, although there are voice quality differences for an individual speaker when speaking L1 vs L2, it is still possible to identify them (e.g., pick them out from a group of speakers) in either language.

With regard to the second question, speech perception literature suggests that this is not necessarily a straightforward task for listeners. For instance, language familiarity plays a significant role in voice identification [25]. Furthermore, [26] found that, when asked to focus on voice quality to judge voice similarity in a foreign language, monolingual listeners were not able to ignore language characteristics. In [27] it was found that monolingual English speakers can identify English speakers significantly more successfully than they can identify Spanish speakers speaking Spanish. Winters et al. [28] describe two experiments which look at respectively, identification and discrimination of bilingual talkers across languages (English and German). The results of these experiments indicate that there is sufficient language-independent speaker-specific information in speech for listeners to generalize knowledge of speakers' voices across English and German and to successfully discriminate between bilingual speakers regardless of the language they are speaking. The lessons we draw from these studies are that it is important to pay attention to which language(s) the *listeners* speak and how familiar they are with the pairs of languages they are listening to.

An in-depth investigation of whether a speaker sounds the same in two different languages goes beyond the scope of EMIME, but we have investigated the rephrased question "Do these two sentences (in L1 and L2) sound like they were spoken by the same person?". In Wester [29], native English listeners were presented with two sentences spoken by bilingual speakers (English/German and English/Finnish) and were asked to judge whether the sentences were spoken by the same person or not. The results showed that listeners perform well on this task, they are able to discriminate between speakers significantly better than chance. However, we also found that listeners are significantly less accurate on cross-lingual trials than on matched-language pairs. These initial findings are promising for EMIME as listeners are at least able to discriminate between bilingual talkers when the stimuli are natural speech.

Ongoing work is expanding on the experiments in [29] by running the same listening test, using the same data, but in synthesized form, i.e. the bilingual data is used as adaptation data to create synthetic stimuli. These synthetic speech stimuli will be used in a listening test rather than natural speech stimuli. If listeners also perform well on this task, it will bring us one step closer to the EMIME scenario in which cross-lingual speaker adaptation comparing natural and synthetic speech must be assessed.

5. Discussion

We have provided an overview of several speaker adaptation experiments, and have brought them together under a common descriptive framework: first form a mapping between two acoustic models, then use the mapping to transfer either speech data or adaptation transforms from one model to the other.

The experiments using phone recognisers or LVCSR to transcribe the adaptation data, the two pass decision tree method [14] and the decision tree marginalisation approach [17] all suggest that errors in transcription do not lead to poorer quality speech synthesis. The precise reasons for this require further investigation, but the explanation is probably that erroneous transcriptions are still "close enough" to be useful for learning adaptation transforms. For example, a word error will still contain many correct phones, a phone error will often be of a similar class to the correct phone, and very often in the same adaptation regression class.

Adaptation transforms learned with respect to triphone models perform as well as adaptation of full context models [13, 14] even though this neglects prosody. That is not to say that the F0 and duration parameters are not adapted, it just means than the adaptation classes for these prosodic parameters are formed on a phonetic basis only. Performing linguistic analysis (i.e., the TTS front end), to predict prosodic information from the word sequence, on estimated word transcriptions containing errors does not seem to affect synthesis in terms of naturalness or speaker similarity. In general, unsupervised systems were not significantly different from supervised comparison systems. Overall these are very positive results: unsupervised adaptation of a TTS voice is possible using ASR models.

We should always keep in mind that the data on which the average voice models are trained also have noisy full context labels, because there is no guarantee that the TTS-frontend predicts full context labels that exactly correspond to the acoustic signal (i.e., to how the speaker read that sentence). This is particularly true of the suprasegmental labels but also applies to phenomena such as vowel reduction and segment deletions. It may seem obvious that better average voice models could be built if the training data full context labels were a more accurate reflection of the speech signal. However, this may result in better synthetic speech only if the full context used during synthesis were equally accurate. It is possible that consistency between training and synthesis (i.e., using the same front-end to produce the labels in both cases) is more important than accurate labelling of the training data.

We hope that the research being conducted in the EMIME project has implications extending beyond the narrow goals of speaker-adaptive speech-to-speech translation, to areas such as unsupervised adaptation more generally, unified models for ASR and TTS, voice conversion, multilingual ASR and multilingual TTS. Specific challenges common to several applications include inter-corpus normalisation and the investigation of new parameter sharing and labelling schemes for multilingual modelling.

6. Acknowledgements

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) grant agreement 213845 (the EMIME project: `www.emime.org`).

7. References

[1] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard 2007 (in Proc. Sixth ISCA Workshop on Speech Synthesis)*, 2007.

[2] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, 2008.

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

[4] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," *IEEE Journal of Selected Topics in Signal Processing*, 2010, accepted.

[5] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, pp. 724–732, 2009.

[6] G. Blackwood, A. de Gispert, J. Brunning, and W. Byrne, "European language translation with weighted finite state transducers: The CUED MT system for the 2008 ACL workshop on statistical machine translation," in *Proc. of the Third Workshop on Statistical Machine Translation*, 2008.

[7] A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne, "Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions," in *Proc. NAACL-HLT*, 2009.

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

[9] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992.

[10] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002.

[11] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn.(E)*, vol. 20, no. 3, pp. 199–206, 1999.

[12] L. Saheer, P. Garner, J. Dines, and H. Liang, "VTLN adaptation for statistical speech synthesis," in *Proc. ICASSP '10*, 2010.

[13] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. Interspeech '08*, 2008.

[14] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proc. Interspeech '09*, 2009.

[15] J. Dines, L. Saheer, and H. Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves," in *Proc. Interspeech '09*, 2009.

[16] K. Oura, K. Tokuda, J. Yamagishi, S. King, and M. Wester, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proc. ICASSP '10*, 2010.

[17] 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. ICASSP '10*, 2010.

[18] M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo, and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," in *Proc. ICASSP '10*, 2010.

[19] Y.-J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proc. ISCSLP '08*, 2008.

[20] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. Interspeech '09*, 2009.

[21] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech of cross-language voice conversion," *J. Acoust. Soc. Am.*, vol. 90, no. 1, pp. 76–82, July 1991.

[22] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and STRAIGHT," in *Proc. Eurospeech '01*, 2001.

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

[24] 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*, no. 48, pp. 1227–1242, 2006.

[25] J. Goggin, C. Thompson, G. Strube, and L. Simental, "The role of language familiarity in voice identification," *Memory and Cognition*, vol. 19, no. 5, pp. 448–458, 1991.

[26] V. Stockmal, Z. Bond, and D. Moates, "Judging voice similarity in unknown languages," in *Proc. of the 17th Congress of Linguists*, Prague, 2004.

[27] C. Thompson, "A language effect in voice identification," *Applied Cognitive Psychology*, vol. 1, pp. 121–131, 1987.

[28] S. Winters, S. Levi, and D. Pisoni, "Identification and discrimination of bilingual talkers across languages," *J. Acoust. Soc. Am.*, vol. 123, p. 4524, 2008.

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