Personalising speech-to-speech translation: unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis

John Dines*, Hui Liang*, Lakshmi Saheer*

Idiap Research Institute, Martigny, Switzerland

Matthew Gibson*, William Byrne*

Cambridge University Engineering Department, Trumpington Street, U.K.

Keiichiro Oura*, Keiichi Tokuda*

Department of Computer Science and Engineering, Nagoya Institute of Technology, Japan

Junichi Yamagishi*, Simon King*, Mirjam Wester*

Centre for Speech Technology (CSTR), University of Edinburgh, U.K.

Teemu Hirsimäki*, Reima Karhila*, Mikko Kurimo*

Adaptive Informatics Research Centre, Aalto University, Finland

Abstract

In this paper we present results of unsupervised cross-lingual speaker adaptation applied to text-to-speech synthesis. The application of our research is the personalisation of speech-to-speech translation in which we employ a HMM statistical framework for both speech recognition and synthesis. This framework provides a logical mechanism to adapt synthesised speech output to the voice of the user by way of speech recognition. In this work we present results of several different unsupervised and cross-lingual adaptation approaches as well as an end-to-end speaker adaptive speech-to-speech translation system. Our experiments show that we can successfully apply speaker adaptation in both unsupervised and cross-lingual scenarios and our proposed algorithms seem to

*Corresponding author
Email address: john.dines@idiap.ch (John Dines)

generalise well for several language pairs. We also discuss important future directions including the need for better evaluation metrics.
Keywords: Speech-to-speech translation, Cross-lingual speaker adaptation, HMM-based speech synthesis, Speaker adaptation, Voice conversion

---

1. Introduction

One of the most elementary and crucial elements of human communication – spoken language – remains a fundamental barrier to economic, cultural and policy exchange both in domestic and international relations. It is clear that a key to breaking down this language barrier is through computer assisted interaction, but the ideal solution in which cross-lingual spoken interaction is instantaneously and seamlessly facilitated by an unobtrusive automated assistant, still remains only a vision for the future. Even so, the critical elements that would comprise such a system – automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) – have made dramatic leaps in performance in the last decade and progress in these fields will continue to bring such a device closer to reality.

Several research and commercially based speech-to-speech translation efforts have been pursued in recent years, to mention only a few: Verbmobil a long-term project of the German Federal Ministry of Education, Science, Research and Technology, Technology and Corpora for Speech to Speech Translation (TC-STAR) FP6 European project, and the Spoken Language Communication and Translation System for Tactical Use (Transtac) DARPA initiative. Ranging from constrained, mobile applications to ambitious systems demanding considerable computing power, these efforts demonstrate that there is a strong demand for such technology across a broad spectrum of applications. One aspect which we take for granted in spoken communication that is largely missing from current technology is a means to facilitate the personal nature of spoken dialog. That is; state-of-the-art approaches lack or are limited in their ability to be personalised in an effective and unobtrusive manner, and so act as a barrier to

natural communication. The authors of this paper are collaborating in an ongoing FP7 European project, *Effective Multilingual Interaction In Mobile Environments* (EMIME), the goal of which is the personalisation of speech-to-speech translation (SST) systems.

The EMIME project aims to achieve its goal of personalised speech-to-speech translation through the use of hidden Markov model based ASR and TTS. Within the last two decades, ASR technology has almost completely converged around this single paradigm and more recently HMM-based TTS is likewise showing a strong concentration of interest from both researchers and industry [1, 2, 3]. The use of a common framework for ASR and TTS provides several interesting research opportunities in the framework of SST, including the development of unified approaches for the modelling of speech for recognition and synthesis that will need to adapt across languages to each user's speaking characteristics. Thus, a core goal of EMIME is the development of unsupervised cross-lingual speaker adaptation for HMM-based TTS.

In this paper we present results from our first experiments on the development of cross-lingual adaptation methods. This work represents a consolidation of several individual research directions currently under investigation by EMIME partners across several targeted language pairs. We show that, using the HMM framework, SST can be posed in two ways: the traditional 'pipeline' approach, where speech input follows a path through independent ASR, MT and TTS modules, or in a 'unified' approach in which ASR and TTS modules are tightly coupled. We present results of cross-lingual speaker adaptation using both pipeline and unified approaches also comparing performance in supervised and unsupervised scenarios. We also present results obtained using a complete end-to-end speaker adaptive SST system. An important conclusion that can be drawn from this work is that conventional speaker adaptation algorithms, long employed by the ASR community and more recently for TTS, are inherently robust when employed in an unsupervised context and provide consistent performance across the language pairs that is only marginally worse than intra-lingual adaptation.

The remainder of the paper is organised as follows: Section 2 we provide a brief overview of speech-to-speech translation with a focus on the pipeline and unified frameworks. Following this, in Section 3 we detail speaker adaptation for HMM-based TTS, drawing together recent work on unsupervised and cross-lingual adaptation. Sections 4 and 5 present our experimental studies to date and a discussion of these results, respectively. Finally, in Section 6 we conclude the paper with a summary of our findings and future directions.

2. Speech-to-speech translation with hidden Markov models

Speech-to-speech translation typically comprises three component technologies: ASR to convert speech in the input language into text in the input language; MT to convert text in the input language into text in the output language; and TTS to convert text in the output language into speech in the output language. Personalisation of SST implies that an additional component is necessary in order to carry out cross-lingual speaker adaptation (CLSA) of the TTS.

In the EMIME project, the major focus of our work is on the personalisation of speech-to-speech translation using HMM-based ASR and TTS, which involves the development of unifying techniques for ASR and TTS as well as the investigation of methods for unsupervised and cross-lingual modelling and adaptation for TTS. Thus, machine translation forms the 'glue' that allows us to link ASR and TTS modules, but is not a subject of investigation in itself. We have developed a modular research framework that can be used to test different configurations of SST systems. The framework accepts modules for feature extraction (FE), ASR, TTS, MT, and CLSA as illustrated in Figure 1. Two typical configurations are what we call the *pipeline* and *unified* SST frameworks, which we detail in the remainder of this section, but first we provide a brief overview of the HMM-based ASR and TTS.

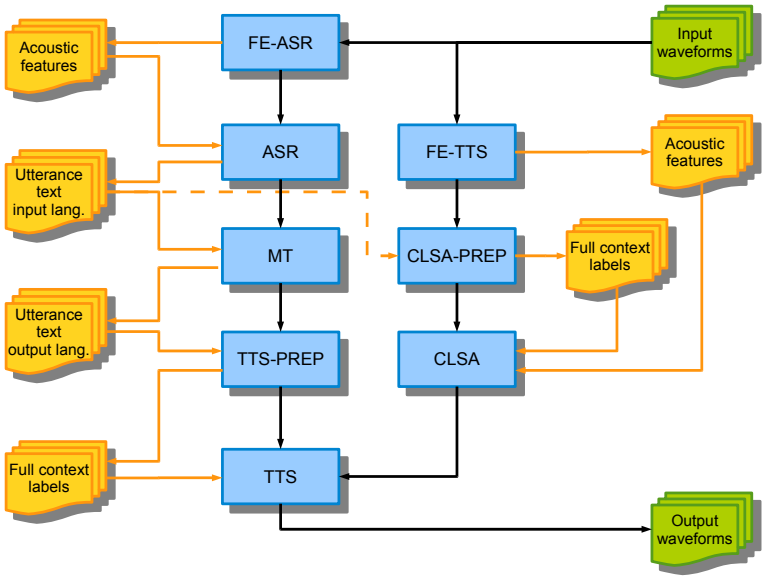

Figure 1: Block diagram of the research system. Blue signifies modules, orange signifies file exchange between modules, and green signifies system input/output files.

2.1. HMM-based ASR and TTS

The central element of our work is the common statistical HMM-framework employed for both ASR and TTS. The adoption of a common modelling approach can be misleading in that it implies a straight-forward means to integrate ASR and TTS. To the contrary, despite the common statistical model the two normally differ significantly [4]. The main differences of consequence to this paper lie at the interfaces between the modules of our SST framework – that is, the acoustic feature extraction and acoustic modelling (see [4] for further details):

Acoustic features

For ASR we normally employ conventional ASR features based on low dimensional short term spectral representations [5, 6] where as in TTS acoustic feature extraction includes mel-cepstrum features derived from STRAIGHT spectrum [7, 8] plus log-pitch and band-limited aperiodic

features for mixed excitation.

Acoustic modelling

ASR acoustic models normally employ a basic HMM topology using phonetic decision tree state tying of triphone context dependent models [9] with Gaussian mixture model (GMM) state emission pdfs. By contrast, TTS acoustic models use multiple stream, single Gaussian state emission pdfs with decision tree state tying of full context models that use a range of contextual information for the prediction of prosodic patterns [10].

2.2. Pipeline translation framework

In the pipeline framework ASR, MT and TTS modules operate largely independently of one another. Figure 1 essentially describes the basis of a possible pipeline configuration in which on the input language side both ASR and TTS modules are used – ASR is necessary to extract text for the machine translator and TTS front-end is required in order to adapt TTS models to the user's voice characteristics (for further details see Section 3.1.1). On the output language side, TTS is once again employed to synthesise the output of the machine translation with voice characteristics of the user. An advantage of the pipeline approach is that it enables simpler integration of components and does not involve any compromises to performance by attempting to combine ASR and TTS modelling. On the other hand, there is a large degree of redundancy in the system.

2.3. Unified translation framework

In contrast to the pipeline approach, a unified translation framework attempts to use common modules for both ASR and TTS. Such a framework is illustrated in Figure 2. It can be seen that the system is conceptually simpler with a minimum of redundancy with respect to feature extraction and acoustic models. Cross-lingual speaker adaptation of TTS is implicit to the ASR, thus a TTS front-end is not required on the input language side (also see Sections 3.1.2 and 3.1.3). The development of such a framework implies the use of common

feature extraction and acoustic modelling techniques for ASR and TTS, however, such unified modelling may come at the expense of reduced performance for ASR and/or TTS. We refer to our previous work on unified modelling for HMM-based ASR and TTS, which show that this is currently the case [11, 12, 4].

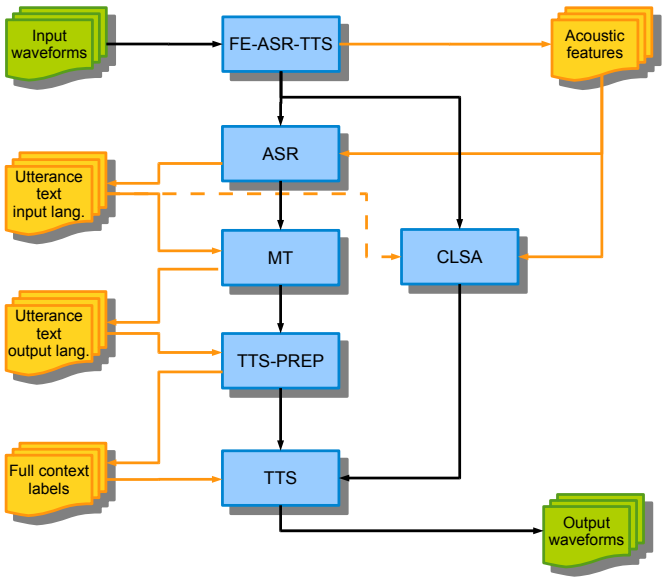

Figure 2: Unified approach to speech-to-speech translation. ASR and TTS modules use the same acoustic features and shared acoustic models (not shown in this diagram).

3. Speaker adaptation for HMM-based TTS

Ideally, in order to build an HMM-based speech synthesizer of high quality for a particular speaker, it is necessary to collect a large amount of speech data from the speaker as training data. Unfortunately, this is often unfeasible as the data collection and annotation is extremely time-consuming and expensive. Speaker adaptation has been proposed as an alternative to overcome this problem by requiring as little as some tens of utterances from a particular speaker as adaptation data. Firstly, an average voice (or speaker-independent) model set is trained on an appropriate multi-speaker speech corpus. Then the average

voice model is transformed to that of the target speaker using utterances read by the particular speaker. Typically, the transofmration of the model is performed using linear transformations estimated by means of maximum likelihood linear regression [13] and/or maximum a posteriori (MAP) adaptation [14]. Such an adapted model set can resemble, to a great extent, a speaker-specific model set [15, 16, 17].

Speaker adaptation plays two key roles in speech-to-speech translation. On the ASR side, it can considerably increase the recognition accuracy, which provides more correct text input for the subsequent machine translation. On the TTS side, it can also be used to personalise the speech synthesised in the output language. We are mostly interested in this latter aspect, i.e., personalisation of output speech.

As mentioned in Section 1, the core of our work is the development of unsupervised cross-lingual speaker adaptation for HMM-based TTS. This implies that we are facing two main challenges: unsupervised adaptation and cross-lingual adaptation of TTS. It follows that in the context of SST, adaptation must normally be performed using the output of the speech recognition system, however, the output of a speech recogniser does not provide the full-context labels [18] normally used for the adaptation of TTS. As a result, TTS models can not be adapted directly from ASR using conventional techniques as mentioned in [19]. Similarly, for cross-lingual adaptation we need to consider how to adapt TTS models of the output language using speech data from the input language. These two challenges will be elaborated in the two remainder of this section.

3.1. Unsupervised adaptation

HMM-based TTS is a parametric approach to speech synthesis, so that we can apply mature and widely used speaker adaptation algorithms from the HMM-based ASR community, for instance, maximum likelihood linear regression (MLLR) or maximum a-posteriori (MAP), and apply them to HMM-based TTS directly. We can achieve unsupervised adaptation of TTS through the use of ASR either by using the noisy text transcription of the speech data with

standard TTS adaptation approaches or using methods that more closely couple ASR and TTS models in so called 'unified' frameworks. These three approaches are described in further detail below.

3.1.1. Using TTS front-end

This is the most straight-forward approach – a combination of a word-based large-vocabulary continuous speech recognition and conventional speaker adaptation for HMM-based TTS. The speech recognition provides word-level recognition results, which are then translated into full-context labels by a TTS front-end. With these full-context labels and corresponding input speech data, adapting the voice identity of TTS models is carried out. The main drawback of such an approach is caused by the noisy text. Full-context labels generated by a TTS front-end may contain many errors due to recognition errors. For instance [20] reports significant differences observed for the quality of synthetic speech using a TTS front-end despite the use of a state-of-the-art six-pass LVCSR systems and confidence scores calculated from confusion networks using word posteriors [21, 22]. Such adaptation is synonymous with the pipeline SST approach previously described since the ASR is largely decoupled from the adaptation of TTS.

3.1.2. Two-pass decision tree construction

In this approach, full-context models are clustered using a decision tree to enable robust estimation of their parameters [23, 24, 10]. Note that the decision tree may have questions related to prosody or lingustic information, which are normally not used for ASR. By imposing constraints upon the decision tree structure, multiple-component triphone mixture models may be derived from single-component full-context models [12]. This constrained decision tree construction process is illustrated in Figure 3.

The first stage, indicated as Pass 1 in Figure 3, uses only questions relating to left, right and central phonemes to construct a phonetic decision tree. This decision tree is used to generate a set of tied triphone contexts, which are easily

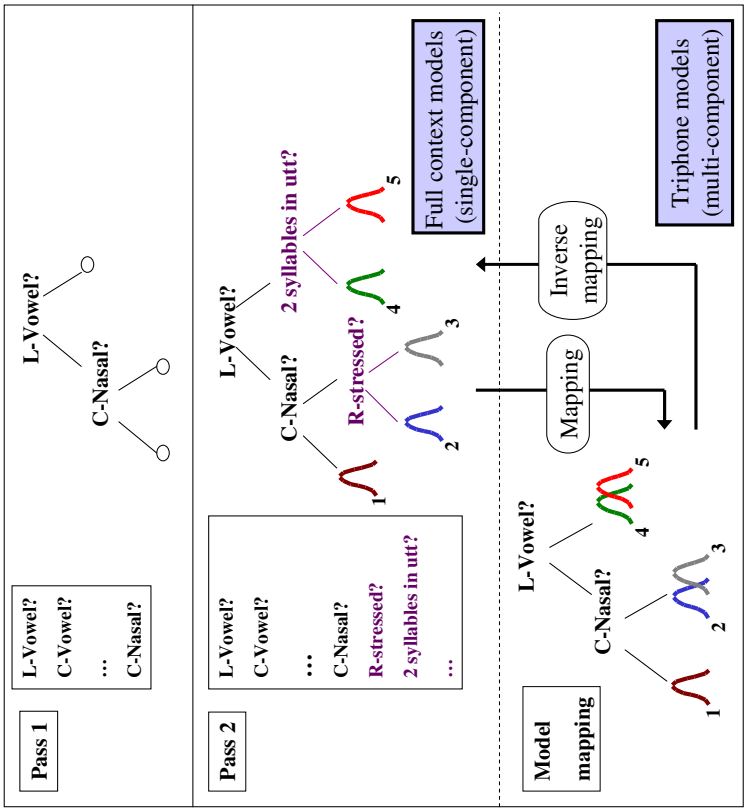

Figure 3: Two-pass decision tree construction. Mapping functions permit sharing of full-context models for TTS and triphone models for ASR.

integrated into the ASR. Pass 2 extends the decision tree constructed in Pass 1 by introducing additional questions relating to supra-segmental information. The output of Pass 2 is an extended decision tree that defines a set of tied full contexts. After this two-pass decision tree construction, single-component

Gaussian state output distributions are estimated for the tied full contexts associated with each leaf node of the extended decision tree. These models are then used for speech synthesis.

A mapping from the single-component full-context models to multiple-component triphone models is defined as follows. Each leaf node of the extended decision tree has a unique 'triphone ancestor' node, namely its ancestor leaf node of the Pass 1 decision tree. Each set of Gaussian components associated with the same 'triphone ancestor' is grouped as components of a multiple-component mixture distribution to model the context defined by the 'triphone ancestor'. The derived triphone models are illustrated at the bottom of Figure 3. The weight of each mixture component is calculated from the occupancies associated with components of the Pass 2 leaf node contexts. The inverse mapping from triphone models to full-context models is obtained by associating each Gaussian component with its original full context. Given this mapping between full-context and triphone models, unsupervised adaptation of full-context acoustic models may be simply achieved via adaptation of triphone models: Triphone models derived from full-context models are used to estimate triphone-level transcriptions of adaptation data. The estimated transcriptions are then used to adapt the triphone models. The adapted triphone models are subsequently mapped back to full-context models using the inverse mapping to enable adaptation of the TTS models without the use of full-context labels.

3.1.3. Decision tree marginalisation

Decision tree marginalization [11] allows the derivation of triphone context models from a full-context speech synthesis model such that the marginalised models can be used in ASR and unsupervised adaptation. Hence, the first stage involves the training of a conventional HMM-based speech synthesis system where each HMM state emission distribution is typically composed of a single Gaussian PDF.

Conventionally, generating a previously unseen model for synthesis is carried out by traversing the decision tree according to the full-context label and

eventually assigning one leaf node to each state of the new model. Decision tree marginalization generates a triphone recognition model from the full context decision tree in almost the same manner. The difference lies in the cases where the questions associated with intermediate nodes are irrelevant to the triphone context. In such cases both children of the intermediate node are traversed, effictively marginalising out contexts associated with that question. A triphone model is thus associated with more than one leaf node resulting in a state emission distribution of multiple Gaussian components. In other words, a triphone model constructed by decision tree marginalization of a synthesis model set can be viewed as a weighted sum of full-context single Gaussian emission distributions whose mixture weights are calculated based on their corresponding occupancies. See Figure 4 for an example.

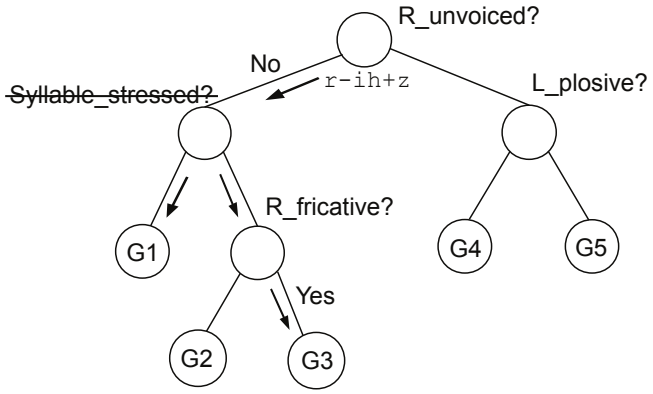

Figure 4: An example of decision tree marginalization, showing how a new recognition model "`r-ih+z`" is derived from a decision tree of a speech synthesis system ("`L_`" / "`R_`": left/right phone; "G?": full-context state emission PDFs)

Once the marginalised triphone models are obtained, transforms are estimated from the adaptation data using the regression class tree of the full-context synthesis model. Thus, subsequent adaptation of the full-context synthesis models involves straight-forward application of the transforms obtained from adaptation of the marginalised triphone models. The decision tree marginalization process described is actually a special case. It can be extended such that any

subset of the full-context labels can be marginalized out. For instance, we can create tonal monophone models by marginalizing out all the contexts that are unrelated to the base phone context and tone information.

3.1.4. Differences between two-pass decision tree construction and decision tree marginalisation

It should be evident from the descriptions in Sections 3.1.2 and 3.1.3 that two-pass and marginalisation approaches are closely related and in fact two-pass is a special case. In light of these similarities it is also worth noting the differences that distinguish the two and possible practical implications. The most evident difference is that two-pass tree construction first clusters HMM parameters according to ASR contexts and then follows with TTS clustering whereas the marginalisation approach, as it has been described, performs the contrary. We may expect then, that the two-pass approach may favour ASR performance over TTS performance and visa-versa for the marginalisation approach.

3.2. Cross-lingual adaptation

Cross-lingual speaker adaptation for HMM-based TTS shares some similarities with the development of ASR systems for resource-poor languages – in both cases well-trained model sets are in a language different from that of given adaptation/training data requiring a means to bridge the gap between the languages of the models and data. Current cross-lingual speaker adaptation can be viewed as being largely based on mapping methods [25] – trying to find correspondence between two different languages, either on the phoneme level using phonetic knowledge or on the HMM state level using data driven approaches. Previous work has shown data driven approaches appear to give better results and as such they have been pursued in this work [26, 27].

3.2.1. State-mapping based approaches to cross-lingual adaptation

Wu *et al.* [27] proposed the state-level mapping approach for cross-lingual speaker adaptation. Establishing state-level mapping rules consists of two steps. Firstly, two average voice models are trained in two languages (say, s and g),

respectively. Secondly, each HMM state, $\Omega_k^s\,(k = 1, \ldots, N^s)$, in the language s is associated with a HMM state $\Omega_j^g\,(j = 1, \ldots, N^g)$ that is the most similar among all the states in the language g. N^s and N^g are the total number of the states in the two respective languages.

Cross-lingual adaptation can then be applied by mapping either the data or speaker transforms. In the *transform mapping* approach, intra-lingual adaptation is first carried out in the input language. Following this, the transforms are applied to the states of the output language acoustic model using the state mappings derived such that the transform associate with states in the input language are applied to their respective mapped state in the output language. Alternatively, a *data mapping* approach was proposed in which states belonging to the input language acoustic model are replaced by states belonging to the output language acoustic model according to the derived state mapping. The 'data mapped' acoustic model may then be adapted in the usual intra-lingual manner and the resulting transformed state emission pdfs can be directly used for synthesis in the output language.

3.2.2. KLD-based state mapping

Since single Gaussian mixture models are used here, let us denote parameters of each state model Ω_k^s including a self-transition probability a_k^s, a mean vector $\boldsymbol{\mu}_k^s$ and a covariance matrix $\boldsymbol{\Sigma}_k^s$. Similarly, we denote the corresponding self-transition probability, mean vector and covariance matrix of the input language as a_j^g, $\boldsymbol{\mu}_j^g$ and $\boldsymbol{\Sigma}_j^g$, respectively.

For each state model Ω_j^g in the input language, we want to find a nearest state model Ω_k^s in the output language, which has the minimum KLD with Ω_j^g. In the case of single Gaussian mixture models, the upper bound of KLD [28] between two state models is calculated as

$$D_{\mathrm{KL}}(\Omega_j^g, \Omega_k^s) \leqslant \frac{D_{\mathrm{KL}}\left(G_k^s || G_j^g\right)}{1 - a_k^s} + \frac{D_{\mathrm{KL}}\left(G_j^g || G_k^s\right)}{1 - a_j^g} + \frac{(a_k^s - a_j^g)\log(a_k^s / a_j^g)}{(1 - a_k^s)(1 - a_j^g)} \qquad (1)$$

where G_k^s denote the Gaussian distribution related to the state model Ω_k^s, which includes the mean vector $\boldsymbol{\mu}_k^s$ and covariance matrix $\boldsymbol{\Sigma}_k^s$, and the KLD between

two Gaussian distributions is calculated as

$$D_{\mathrm{KL}}\left(G_k^s||G_j^g\right) = \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Sigma}_j^g\right|}{\left|\boldsymbol{\Sigma}_k^s\right|}\right) - \frac{D}{2} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_j^{g-1}\boldsymbol{\Sigma}_k^s\right) + \frac{1}{2}\left(\boldsymbol{\mu}_j^g - \boldsymbol{\mu}_k^s\right)^\top \boldsymbol{\Sigma}_j^{g-1}\left(\boldsymbol{\mu}_j^g - \boldsymbol{\mu}_k^s\right) \quad (2)$$

Since we only focus on the distribution of a state model, we ignore the effect of transition probabilities, and calculate the KLD between two state models as

$$D_{\mathrm{KL}}(\Omega_k^s, \Omega_j^g) \approx D_{\mathrm{KL}}\left(G_k^s||G_j^g\right) + D_{\mathrm{KL}}\left(G_j^g||G_k^s\right) \quad (3)$$

Based on the above KLD measurement, the nearest state model $\Omega_{k'}^s$ in the output language for each state model Ω_j^g in the input language is calculated as

$$k_j' = \arg\min_k D_{\mathrm{KL}}(\Omega_j^g, \Omega_k^s). \quad (4)$$

Finally, we map all the state models in the input language to the state models in the output language, which can be formulated as

$$\Omega_j^g \Rightarrow \Omega_{k_j'}^s, \quad j = 1, \ldots, N^g. \quad (5)$$

Here we establish a state mapping from the model space of an input language to that of an output language. In this case, all the state models in the input language have a mapped state model in the output language. However, it should be noted that not all the state models in the output language have a corresponding state model in the input language, and that the mapping direction can be reversed, namely, from the model space of an output language to that of an input language. The KLD-based transform mapping process is illustrated in Figure 5.

3.2.3. Probabilistic state mapping

The state mapping approaches previously described generate a deterministic mapping between HMM states in the input and output languages. An alternative is to derive a stochastic mapping which could take the form of a mapping between states, $P(\Omega_j^g|\Omega_k^s)$, or from states directly to the adaptation

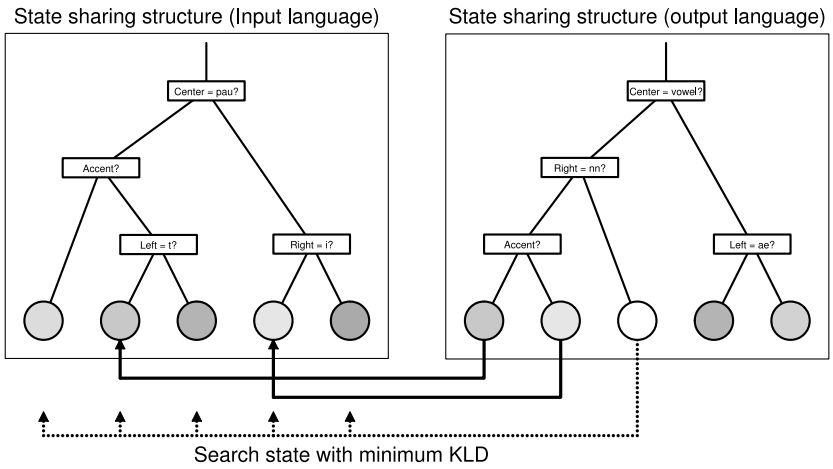

Figure 5: The state-mapping is learned by searching for pairs of states that have minimum KLD between input and output language HMMs. Linear transforms estimated with respect to the input language HMMs are applied to the output language HMMs, using the mapping to determine which transform to apply to which state in the output language HMMs.

data, $P(\Omega_k^s|o_t^g)$, where o_t^g is an observation from input language g at time t. The simplest such way of deriving this mapping is by performing ASR on the adaptation data using an acoustic model of the output language. The resulting sequence of recognised phonemes provides the mapping from data in the input language to states in the output language, though the phoneme sequence itself is meaningless.

4. Experimental studies

The HMM-based speaker adaptation techniques presented in the previous section provide the necessary means to achieve personalised speech-to-speech translation. More specifically, through the combination of unsupervised and cross-lingual adaptation we are able to realise either pipeline or unified SST frameworks. The experiments presented in this section describe independent studies that investigate SST in the context of different combinations of the above approaches. The number of possible combinations is quite significant,

hence, each study concentrates only on a subset as briefly summarised in Table 1.

Table 1: Summary of experimental studies. US: unsupervised adaptation (1: TTS front-end, 2: two-pass decision tree, 3: decision tree marginalisation). CLSA (TM: transform mapping, DM: data mapping, PM: probabilistic mapping)

Study	Framework[a]	Adaptation				CLSA		
		S	US1	US2	US3	TM	DM	PM
1	Unified	■		■	■			■
2	Unified	■		■	■	■	■	■
3	Pipeline	■	■			■		

[a] Only system 3 investigates an end-to-end SST framework. Studies 1 and 2 are focused on the speaker adaptation component only.

Our aim in presenting these studies is to show the range of techniques that have been developed and evaluated to date and to discuss their relative benefits and disadvantages. In so doing we are primarily concerned with assessing the preservation of speaker identity in the speech output. This could also include consideration of complex issues including the human perception of speaker identity, further compounded by the cross-lingual scenario. Such considerations lie outside the scope of our initial investigations and are discussed in more detail elsewhere [25].

We have not set out to provide an exhaustive comparative study, thereby discovering which is the 'best' approach. However, it would also not be appropriate to make direct comparisons of results of systems presented in different studies, thus we instead aim to characterise the effectiveness of the approaches presented with respect to three main criteria using conventional objective and subjective metrics:

Generality across languages

We would like to know whether CLSA performs equivalently across lan-

guages or if some languages are more challenging than others.

Supervised vs unsupervised adaptation

Personalised SST not only relies on ASR to provide input to the MT, but also for unsupervised speaker adaptation of the TTS. Hence, we should know whether the use of noisy transcripts is detrimental to CSLA.

Cross-lingual versus intra-lingual adaptation

Several cross-lingual adaptation schemes have been proposed in the course of our work. We would like to know which of these shows the most promise and compare this against intra-lingual adaptation.

4.1. Study 1: Finnish – English

In this study we use a simple unsupervised probabilistic mapping technique using two-pass decision tree construction that avoids the need to train synthesis models in the input language.

4.1.1. Setup

Full context English average voice models are estimated using speaker adaptive training (SAT, [16]) and the Wall Street Journal (WSJ) SI84 dataset. Acoustic features used are STRAIGHT-analysed Mel-cepstral coefficients [8], fundamental frequency, band aperiodicity measurements, and the first and second order temporal derivatives of all features. The acoustic models use explicit duration models [29] and multi-space probability distributions [30].

Decision trees (one per state and stream combination) are constructed using the two-pass technique of Section 3.1.2. Adapted TTS systems are derived from the average voice models using the two-pass decision tree method ([31]) and constrained maximum likelihood linear regression. Speech utterances are generated from models via feature sequence generation [32] and resynthesis of a waveform from the feature sequence [8].

4.1.2. Adaptation and evaluation datasets

The adaptation datasets comprise 94 utterances from a corpus of parallel text of European parliament proceedings [33]. English and Finnish versions of this dataset are recorded in identical acoustic environments by a native Finnish speaker also competent in English. Statistics relating to these datasets are provided in Table 2. The evaluation dataset comprises English utterances (distinct

Language	# utterances	# minutes	# words
English	94	12.3	1546
Finnish	94	10.9	1066

Table 2: Europarl adaptation datasets.

from the adaptation utterances) from the same Europarl corpus.

4.1.3. Evaluation details

The following systems are evaluated.

- System A: average voice.
- System B: unsupervised cross-lingual adapted.
- System C: unsupervised intralingual adapted.
- System D: supervised intralingual adapted.
- System E: vocoded natural speech.

System B is the result of applying unsupervised cross-lingual adaptation to the average voice models using the Finnish adaptation dataset. System C results from unsupervised adaptation using the English adaptation dataset. System D is identical to System C with the exception that the correct transcription is used during adaptation. System E analyses and resynthesises the evaluation utterances using STRAIGHT[8].

All systems were evaluated by listening to synthesised utterances via a web browser interface, as used in the Blizzard Challenge 2007. The evaluation comprised four sections. In the first pair of sections, listeners judged the naturalness

of an initial set of synthesised utterances. In the second pair of sections, listeners judged the similarity of a second set of synthesised utterances to the target speaker's speech. Four of the target speaker's natural English utterances were available for comparison. Each synthetic utterance was judged using a five point psychometric response scale, where '5' and '1' are respectively the most and least favourable responses.

Twenty-four native English and sixteen native Finnish speakers conducted the evaluation. Different Latin squares were used for each section to define the order in which systems were judged. Each listener was assigned a row of each Latin square, and judged five different utterances per section, each synthesised by a different system.

4.1.4. Results

Figure 6 summarises listener judgements of 'similarity to target speaker' and 'naturalness' using boxplots [34] while Table 3 displays the average mean opinion scores (MOS) of these judgements for each system in the columns labelled 'av'. Analysis of these judgements by listener native language is provided in the columns labelled 'En' and 'Fi', respectively denoting English and Finnish.

Sys	Source lang.	Sup?	MOS similarity			MOS naturalness		
			En	Fi	av	En	Fi	av
A	-	-	1.2	1.1	1.1	2.3	2.4	2.3
B	Fi	N	2.3	1.5	2.0	2.4	2.4	2.4
C	En	N	2.6	1.7	2.2	2.6	2.7	2.7
D	En	Y	2.7	2.0	2.4	2.5	2.8	2.6
E	-	-	4.6	4.6	4.6	3.7	4.1	3.8

Table 3: Mean opinion scores of evaluated systems.

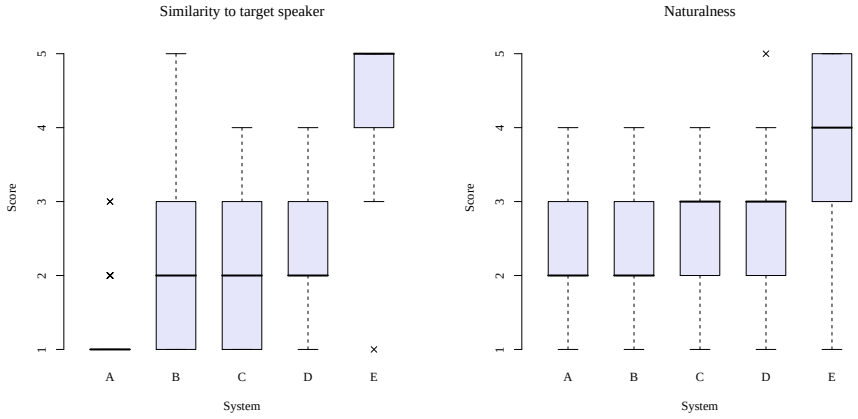

Figure 6: Listener opinion scores for similarity to target speaker and naturalness.

4.2. Study 2: Chinese – English

This study is concerned with comparing different cross-lingual speaker adaptation schemes in supervised and unsupervised settings. Unsupervised adaptation is achieved using the decision tree marginalisation method. Decision tree marginalisation is also used to perform supervised cross-lingual adaptation using only the output language acoustic models. Rather than adapting pitch stream using decision tree marginalisation, we use simple mean shift of the pitch according to the input speech.

4.2.1. Setup

The experiments were conducted using the Mandarin Chinese - English language pair. We trained two average voice, single Gaussian synthesis model sets on the corpora SpeeCon (Mandarin) and WSJ SI84 (English) [35]. We collected bilingual adaptation data from two Chinese students (H and Z) who also spoke English well. The Mandarin and English test prompts, which were not included in the training data, were also selected from SpeeCon and WSJ, respectively. Mandarin and English were defined as input ($L1$) and output ($L2$) languages, respectively, throughout our experiments.

We evaluated four different cross-lingual adaptation schemes each in supervised and unsupervised modes, making a total of eight systems. These systems (S2, S1-M, S1-T, S1-D, U2, U1-M, U1-T and U1-D) are described as follows, according to the labelling scheme in Table 4:

S2 purely built on the English side

S1-M We marginalized out all the English-specific contexts first. As a result, a Mandarin full-context label was associated with more than one English state-cluster. Then Mandarin adaptation data could be treated as English data for "intra-lingual" speaker adaptation.

S1-T & S1-D as described in Section 3.2.1

U2 purely built on the English side; as described in Section 3.1.3

U1-M We marginalized out all the non-triphone contexts and then recognized Mandarin adaptation data with English models. Mandarin adaptation data was thus associated with the English average voice model set.

U1-T & U1-D Combination of unsupervised adaptation with transform and data KLD-based mapping

3 Speech features were 39th-order mel-cepstra, $\log F_0$, five dimensional band aperiodicity, and their delta and delta-delta coefficients. The CSMAPLR [16] algorithm and 40 adaptation utterances were used. Global variances were calculated on adaptation data. A simple phoneme loop was adopted as a language model for recognition. The average phoneme error rate was around 75%.

4.2.2. Results

We first evaluated system performance using objective metrics. For this we calculated RMSE of mel-cepstrum (MCEP) and F_0, as well as correlation coefficients and voicing error rates of F_0. See Table 5.

Our formal listening test consisted of two sections: naturalness and speaker similarity. In the naturalness section, a listener was requested to listen to a

System name format: (**S**/**U**) (**1**/**2**) **-** (**D**/**T**/**M**)	
S/U	supervised / unsupervised
1/2	cross-lingual / intra-lingual
D/T	data/transform version of HMM state mapping
M	Decision tree marginalization was used instead of HMM state mapping. The average voice model set of Mandarin ($L1$) was therefore unnecessary.

Table 4: Labelling of CLSA systems for Study 2

natural utterance first and then utterances synthesized by the eight systems each as well as vocoded speech in a random order. Having listened to each synthesized utterance, the listener was requested to score what he/she heard on a 5-point scale of 1 through 5, where 1 meant "completely unnatural" and 5 meant "completely natural". The speaker similarity section was designed in the same fashion, except that a listener was requested to listen to one more utterance which was synthesized directly by the average voice models and the 5-point scale was such that 1 meant "sounds like a totally different person" and 5 meant "sounds like exactly the same person". Twenty listeners participated in our listening test. Because of the anonymity of our listening test, only two native English speakers can be confirmed. The results are shown in Figures 7 – 10.

4.3. Study 3: English – Japanese

Although our focus up until now has been on the evaluation of cross-lingual speaker adaptation, we have also performed some experiments with an end-to-end speech-to-speech translation system.

4.3.1. Setup

We performed experiments on unsupervised English-to-Japanese speaker adaptation for HMM-based speech synthesis. An English speaker-independent model for ASR and average voice model for TTS were trained on the pre-defined

	MCEP		F_0			
	RMSE (/frm)		RMSE (Hz/frm)		CorrCoef	
	H	Z	H	Z	H	Z
AV	1.39	1.43	26.0	35.9	0.46	0.49
S2	1.04	1.04	11.8	9.6	0.46	0.56
U2	1.06	1.08	13.0	14.0	0.47	0.54
S1-T	1.23	1.22	20.0	12.6	0.47	0.51
U1-T	1.24	1.26	21.1	16.5	0.48	0.53
S1-D	1.13	1.14	19.5	12.6	0.47	0.51
U1-D	1.13	1.13	22.7	17.3	0.48	0.55
S1-M	1.10	1.11	25.9	22.3	0.48	0.54
U1-M	1.10	1.11	25.1	21.0	0.48	0.53

Table 5: Objective evaluation results ("AV" means "average voice")

training set "SI-84" comprising 7.2k sentences uttered by 84 speakers included in the "short term" subset of the WSJ0 database (15 hours of speech). A Japanese average voice model for TTS was trained on 10k sentences uttered by 86 speakers from the JNAS database (19 hours of speech). One male and one female American English speaker, not included in the training set, were chosen from the "long term" subset of the WSJ0 database as target speakers. The adaptation data comprised 5, 50, or 2000 sentences selected arbitrarily from the 2.3k sentences available for each of the target speakers.

Speech signals were sampled at a rate of 16 kHz and windowed by a 25 ms Hamming window with a 10 ms shift for ASR and by an F_0-adaptive Gaussian window with a 5 ms shift for TTS. ASR feature vectors consisted of 39-dimensions: 13 PLP features and their dynamic and acceleration coefficients. TTS feature vectors comprised 138-dimensions: 39-dimension STRAIGHT mel-cepstral coefficients (plus the zeroth coefficient), $\log F_0$, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 3-state left-to-right triphone HMMs for ASR and 5-state left-to-right context-dependent

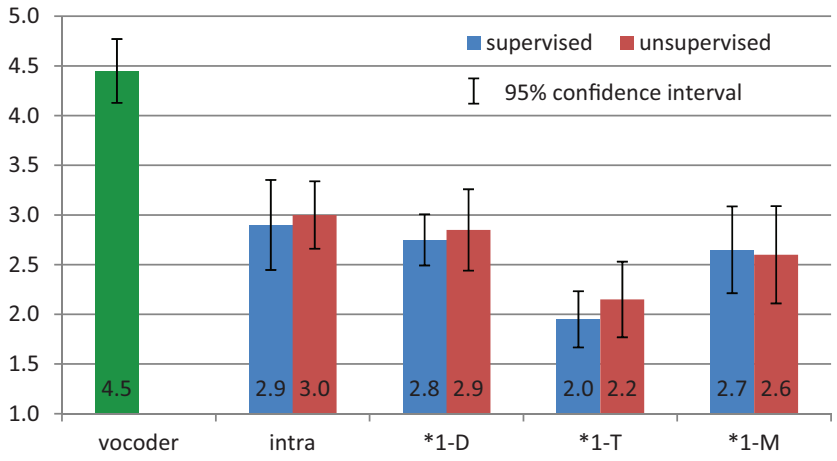

Figure 7: Naturalness score (speaker H)

multi-stream MSD-HSMMs for TTS. Each state had 16 Gaussian mixture components for ASR and a single Gaussian for TTS. For speaker adaptation, the linear transforms $\boldsymbol{W}_i$ had a tri-block diagonal structure, corresponding to the static, dynamic, and acceleration coefficients. Since automatically transcribed labels for unsupervised adaptation contain errors, we adjusted a hyperparameter (τ_b in [16]) of CSMAPLR to higher-than-usual value of 10000 in order to place more importance on the prior (which is a global transform that is less sensitive to transcription errors).

4.3.2. Results

Synthetic stimuli were generated from 7 models: the average voice model and supervised or unsupervised adapted models each with 5, 50, or 2000 sentences of adaptation data. 10 Japanese native listeners participated in the listening test. Each listener was presented with 12 pairs of synthetic Japanese speech samples in random order: the first sample in each pair was a reference original utterance from the database and the second was a synthetic speech utterance generated from one of the 7 models. For each pair, listeners were asked to give an opinion score for the second sample relative to the first (DMOS), expressing how similar the speaker identity was. Since there were no Japanese speech data

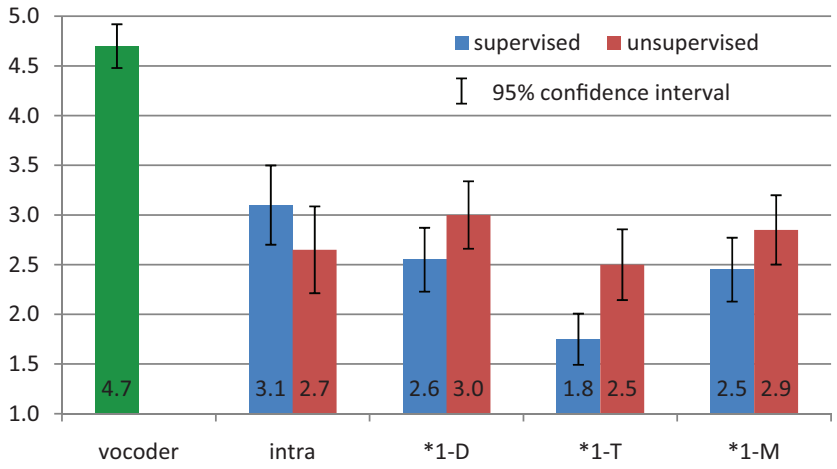

Figure 8: Naturalness score (speaker Z)

available for the target English speakers, the reference utterances were English. The text for the 12 sentences in the listening test comprised 6 written Japanese news sentences randomly chosen from the Mainichi corpus and 6 spoken English news sentences from the English adaptation data that had been recognized using ASR then translated into Japanese text using MT. The average WERs of these recognized English sentences were 11.3%, 10.0%, and 11.4% when using 25, 50, and 100 sentences of adaptation data, respectively.

Figure 11 shows the average DMOS and their 95% confidence intervals. First of all, we can see that the adapted voices are judged to sound more similar to target speaker than the average voice. Next, we can see that the differences between supervised and unsupervised adaptation are very small. This is a very pleasing result. However, the effect of the amount of adaptation data is also small, contrary to our expectations.

Figure 12 shows the average scores using Japanese news texts from the corpus and English news texts recognized by ASR and translated by MT. It appears that the speaker similarity scores are affected by the text of the sentences.

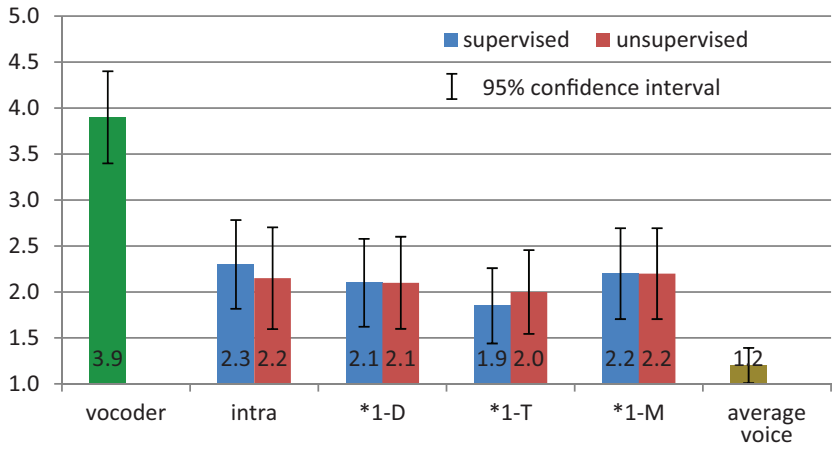

Figure 9: Similarity score (Mandarin reference uttered by speaker H)

5. Discussion

Based on the three studies we have conducted we can draw several conclusions concerning unsupervised cross-lingual adaptation of TTS and its application to personalised speech-to-speech translation.

5.1. Unsupervised versus supervised adaptation

In our three studies we compared supervised and unsupervised adaptation using several approaches. All three studies showed that the adapted voices sound more similar to the target speaker than the average voice and that differences between supervised and unsupervised cross-lingual speaker adaptation are small. In study 2 we note that differences in perceived speaker similarity between supervised and unsupervised adaptation were generally larger when the reference speech was in the same language as the synthesised speech and this also varied depending on the cross-lingual speaker adaptation approach. It appears that the probabilistic mapping approaches from studies 1 and 2 show the least difference between supervised and unsupervised adaptation.

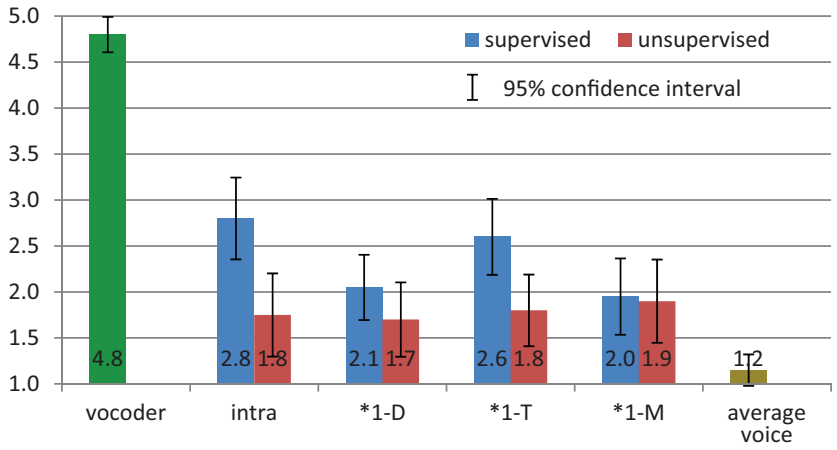

Figure 10: Similarity score (English reference uttered by speaker H)

5.2. Cross-lingual versus intra-lingual adaptation

In study 2 we conducted a comparison of various unsupervised CLSA approaches, including KLD based mappings (both transform and data) and probabilistic mapping based on decision tree marginalisation. We provide both objective and subjective measures. The objective measures indicate that data mapping and probabilistic mapping provide the best results, close to that of intralingual adaptation with transform mapping trailing somewhat behind. This is confirmed by the subjective results for both naturalness and speaker similarity, though we note that when reference speech was in the output language the intra-lingual adaptation was perceived as being somewhat better. In study 1 a different probabilistic mapping-based cross-lingual adaptation approach was undertaken, but similar results were observed.

5.3. Generality across languages

In these three studies we have presented results for three language pairs: Finnish – English, Chinese – English and English – Japanese. Despite the distinct differences between these languages we see that overall unsupervised cross-lingual adaptation has been successful in all cases. Thus we can hypothesise that personalisation of SST based on HMM-adaptation is relatively robust

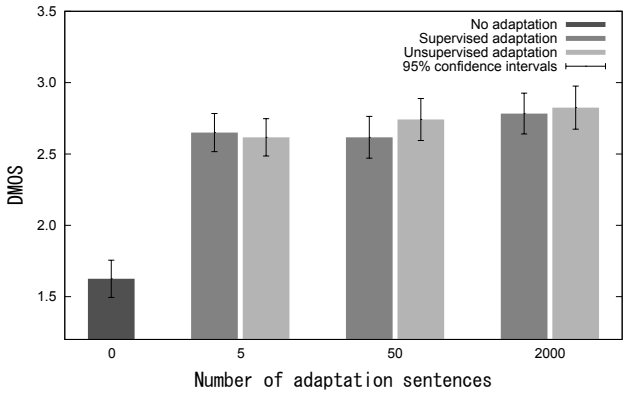

Figure 11: Experimental results (English - Japanese): comparison of supervised and unsupervised speaker adaptation. "0 sentences" means the unadapted average voice model for the output language.

although it may be that some CLSA methods may be more or less susceptible to language differences than others.

5.4. End-to-end system evaluation

In study 3 an end-to-end speech-to-speech system was evaluated. The results from this experiment show that overall speaker similarity is likewise maintained in the end-to-end system compared to the more controlled experiments conducted in studies 1 and 2, though some additional observations could be made with the inclusion of the recognition and machine translation errors in the synthesised output. Most significantly, it appears that the speaker similarity scores are affected by the text of the sentences and the gap between the translated and source language text increases with more adaptation data. These issues will require further investigation.

5.5. Regarding evaluation criteria

In these studies we have used conventional evaluation metrics to judge speaker similarity and naturalness of unsupervised cross-lingual adaptation. It is clear to the authors that further effort also needs to be devoted to the development

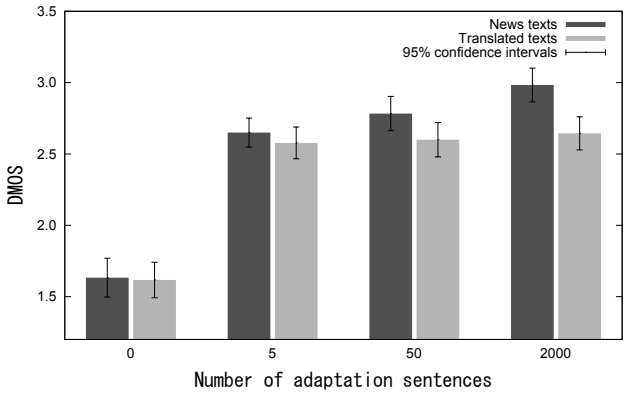

Figure 12: Experimental results (English - Japanese): comparison of Japanese news texts chosen from the corpus and English news texts which were recognized by ASR then translated into Japanese by MT. "0 sentences" means the unadapted average voice model for the output language.

of alternative and more effective evaluation for this type of work. For instance, our current evaluation framework only compares the synthesised output to a given reference – we can imagine that a more appropriate measure might ask listeners to assess speaker similarity in terms of a speaker line up where other competing test utterances would be presented. Our initial results from study 2 that demonstrated the importance of the language of the reference speech on the perception of speaker similarity also highlights the SST application of CLSA may be less demanding that than more general evaluation scenarios where we can provide reference speech in the same language of the synthesised speech.

6. Conclusions

We have presented detailed experiments on cross-lingual speaker adaptation for speech-to-speech translation. Our results show that using HMM-based ASR and TTS we can personalise speech-to-speech translation systems and the challenges of adapting HMM-based TTS in an unsupervised and cross-lingual setting can be addressed using both conventional and novel adaptation frame-

works. Most importantly, speaker similarity is preserved compared to conventional supervised intra-lingual TTS.

Our work towards a new unified translation approach has also shown good progress, with adaptation of TTS showing similar performance to conventional pipeline approaches, though without the additional overhead and complexity. We still need to extend our work on unified models to the analysis of ASR performance.

Finally, our results provide insights into new research directions. Two important directions include the development of better subjective evaluation metrics and also the investigation of methods to adapt supra-segmental speaker properties including pitch and duration statistics, since our studies to date have concentrated mostly adapting the spectrum.

7. 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). SK holds an EPSRC Advanced Research Fellowship. JY is partially supported by EPSRC. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). Simplified descriptions of this research are introduced in papers that will appear in the proceedings of ICASSP 2010 [31, 36, 37].

[1] M. Ostendorf, I. Bulyko, The impact of speech recognition on speech synthesis, in: Proc. IEEE Workshop on Speech Synthesis, Santa Monica, USA, 2002, pp. 99–106.

[2] M. Gales, S. Young, The application of hidden Markov models in speech recognition, Foundations and Trends in Signal Processing 1 (3) (2007) 195–304.

[3] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis, Speech Communication 51 (11) (2009) 1039–1064.

[4] J. Dines, J. Yamagishi, S. King, Measuring the gap between HMM-based ASR and TTS, IEEE Journal of Special Topics in Signal Processing 4 (6) (2010) 1046 – 1058.

[5] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, Journal of Acoustical Society of America 87 (4) (1990) 1738–1752.

[6] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, Acoustics, Speech and Signal Processing, IEEE Transactions on 28 (4) (1980) 357–366.

[7] K. Koishida, G. Hirabayashi, K. Tokuda, T. Kobayashi, Mel-generalized cepstral analysis —a unified approach to speech spectral estimation, in: Proc. ICSLP, Vol. 3, Yokohama, Japan, 1994, pp. 1043–1046.

[8] H. Kawahara, I. Masuda-Katsuse, A. Cheveigne, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds, Speech Communication 27 (1999) 187–207.

[9] J. J. Odell, The use of context in large vocabulary continuous speech recognition, Ph.D. thesis, Queens College, University of Cambridge (1995).

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in: Proc. of Eurospeech, 1999, pp. 2347–2350.

[11] J. Dines, L. Saheer, H. Liang, Speech recognition with speech synthesis models by marginalising over decision tree leaves, in: Proc. of Interspeech, 2009, pp. 1395–1398.

[12] M. Gibson, Two-pass decision tree construction for unsupervised adaptation of hmm-based synthesis models, in: Proc. of Interspeech, 2009, pp. 1791–1794.

[13] M. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Computer Speech and Language 12 (2) (1998) 75–98.

[14] J. Gauvain, C. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process. 2 (1994) 291–298.

[15] J. Yamagishi, T. Kobayashi, Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training, IEICE Trans. Inf. & Syst E90-D (2) (2007) 533–543.

[16] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, J. Isogai, Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm, IEEE Trans. Speech, Audio & Language Process. 17 (1) (2009) 66–83.

[17] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, S. Renals, A robust speaker-adaptive HMM-based text-to-speech synthesis, IEEE Trans. Speech, Audio & Language Process. 17 (6) (2009) 1208–1230.

[18] K. Tokuda, H. Zen, A. W. Black, HMM-based approach to multilingual speech synthesis, in: S. Narayanan, A. Alwan (Eds.), Text to speech synthesis: New paradigms and advances, Prentice Hall, 2004, pp. 135—153.

[19] S. King, K. Tokuda, H. Zen, J. Yamagishi, Unsupervised adaptation for HMM-based speech synthesis, in: Proc. Interspeech 2008, Brisbane, Australia, 2008, pp. 1869–1872.

[20] J. Yamagishi, M. Lincoln, S. King, J. Dines, M. Gibson, J. Tian, Y. Guan, Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework, in: Proc. Blizzard Challenge Workshop, Edinburgh, U.K., 2009.

[21] G. Evermann, P. Woodland, P. C. Woodl, Large vocabulary decoding and confidence estimation using word posterior probabilities, in: Proc. ICASSP 2000, 2000, pp. 2366–2369.

[22] L. Mangu, E. Brill, A. Stolcke, Finding consensus in speech recognition: word error minimization and other applications of confusion networks, Computer Speech & Language 14 (4) (2000) 373 – 400. doi:DOI: 10.1006/csla.2000.0152.

[23] S. J. Young, J. J. Odell, P. C. Woodland, Tree-based state tying for high accuracy acoustic modeling, in: Proc. ARPA Human Language Technology Workshop, Plainsboro, NJ, 1994, pp. 307–312.

[24] K. Shinoda, T. Watanabe, MDL-based context-dependent subword modeling for speech recognition, J. Acoust. Soc. Japan (E) 21 (2000) 79–86.

[25] M. Wester, et al., Speaker adaptation and the evaluation of speaker similarity in the emime speech-to-speech translation project, in: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Vol. submitted, 2010.

[26] Y.-J. Wu, S. King, K. Tokuda, Cross-lingual speaker adaptation for HMM-based speech synthesis, in: Proc. of ISCSLP, 2008, pp. 1–4.

[27] Y.-J. Wu, Y. Nankaku, K. Tokuda, State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis, in: Proc. of Interspeech, 2009, pp. 528–531.

[28] P. Liu, F. Soong, J.-L. Zhou, Divergence-based similarity measure for spoken document retrieval, in: Proc. of ICASSP, 2007, pp. 89–92.

[29] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Hidden semi-Markov model based speech synthesis, in: Proc. of Interspeech, 2004, pp. 1393–1396.

[30] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, Multi-space probability distribution HMM, IEICE Trans. Inf. & Syst. E85-D(3) (2002) 455–464.

[31] M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo, W. Byrne, Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using

two-pass decision tree construction, in: Proc. of ICASSP, 2010, pp. 4642 – 4645.

[32] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, Speech parameter generation algorithms for HMM-based speech synthesis, in: Proc. of ICASSP, 2000, pp. 1315–1318.

[33] P. Koehn, Europarl: A Parallel Corpus for Statistical Machine Translation, in: Machine Translation Summit, 2005.

[34] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, Statistical analysis of the Blizzard Challenge 2007 listening test results, in: Proceedings of the Blizzard challenge workshop, 2007.

[35] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, S. Renals, A robust speaker-adaptive HMM-based text-to-speech synthesis, IEEE Trans. Speech, Audio & Language Process. 17 (6) (2009) 1208–1230.

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

[37] K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis, in: Proc. of ICASSP, 2010, pp. 4642 – 4645.