Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping

Keiichiro Oura[a], Junichi Yamagishi[b], Mirjam Wester[b], Simon King[b], Keiichi Tokuda[a]

[a] *Department of Computer Science and Engineering, Nagoya Institute of Technology,*
Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
[b] *The Centre for Speech Technology Research, University of Edinburgh,*
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK

Abstract

In the EMIME project, we developed a mobile device that performs personalized speech-to-speech translation such that a user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice. We integrated two techniques into a single architecture: unsupervised adaptation for HMM-based TTS using word-based large-vocabulary continuous speech recognition, and cross-lingual speaker adaptation (CLSA) for HMM-based TTS. The CLSA is based on a state-level transform mapping learned using minimum Kullback-Leibler divergence between pairs of HMM states in the input and output languages. Thus, an unsupervised cross-lingual speaker adaptation system was developed. End-to-end speech-to-speech translation systems for four languages (English, Finnish, Mandarin, and Japanese) were constructed within this framework. In this paper, the English-to-Japanese adaptation is evaluated. Listening tests demonstrate that adapted voices sound more similar to a target speaker than average voices and that differences between supervised and unsupervised cross-lingual speaker adaptation are small. Calculating the KLD state-mapping on only the first 10 mel-cepstral coefficients leads to huge savings in computational costs, without any detrimental effect on the quality of the synthetic speech.

Keywords: HMM-based speech synthesis, unsupervised speaker adaptation, cross-lingual speaker adaptation, speech-to-speech translation

1. Introduction

The goal of speech-to-speech translation research is to "enable real-time, interpersonal communication via natural spoken language for people who do not share a common language" (Liu *et al.*, 2003). Several research and commercial speech-to-speech translation efforts have been pursued in recent years, for example: *Verbmobil*, a long-term project of the German Federal Ministry of Education, Science, Research and Technology[1], *Technology and Corpora for Speech to Speech Translation* (TC-STAR), an FP6 European project[2], and the *Global Autonomous Language Exploitation* (GALE) DARPA initiative[3]. In the European FP7 project EMIME[4], we developed a mobile device that performs personalized speech-to-speech translation, such that a user's spoken input in one language is used to produce spoken output in another language, while continuing to sound like the user's voice.

In contrast to previous "pipeline" speech-to-speech translation systems that combined isolated automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) components, EMIME places the main emphasis on coupling ASR with TTS, specifically to enable speaker adaptation for HMM-based ASR (Woodland, 2001) and TTS (Yamagishi *et al.*, 2009a) in cross-lingual scenarios. Other work that has investigated coupling components of the speech-to-speech translation systems are, for example, Gao (2003) and Ney (1999) which investigated the coupling of ASR and MT, or Noth *et al.* (2000) in which natural language processing and prosody processing were connected. The principal modeling framework of speaker-adaptive HMM-based speech synthesis is conceptually and technically similar to conventional ASR systems (although without discriminative training) making it possible for both ASR and TTS systems to be built from the same corpora (Yamagishi *et al.*, 2010). This enables the sharing of Gaussians, decision trees or linear transforms between the two (Dines *et al.*, 2010).

In the EMIME project, we conducted extensive experiments exploring the possibilities for combining ASR and TTS models and for achieving unsupervised speaker adaptation (Wester *et al.*, 2010). For example, unsupervised adaptation techniques for HMM-based TTS using either a phoneme recognizer (King *et al.*, 2008) or a word-based large-vocabulary continuous speech recognizer (LVCSR) (Yamagishi *et al.*, 2009b) were explored. In addition, map-

[1] `http://verbmobil.dfki.de/overview-us.html`
[2] `http://www.tc-star.org/`
[3] `http://www.darpa.mil/ipto/programs/gale/gale.asp`
[4] `http://www.emime.org/`

ping between ASR and TTS acoustic models was investigated using 2-pass decision trees (Gibson, 2009) or by the marginalization of decision trees (Dines *et al.*, 2009; Liang *et al.*, 2010). In addition to this, various cross-lingual adaptation techniques for HMM-based TTS were developed. For instance, Wu and Tokuda (2009) proposed a mapping algorithm which maps either the adaptation data or transforms based on the Kullback-Leibler divergence (KLD) between the HMM states of input and output languages. This mapping approach has also been explored by Qian *et al.* (2009); Liang *et al.* (2010).

This paper describes the integration of these developments into a single architecture which achieves unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. We demonstrate an end-to-end speech-to-speech translation system built for four languages – American English, Mandarin, Japanese, and Finnish. Although all language pairs and directions are possible in our framework, only the English-to-Japanese adaptation is evaluated in the perceptual experiments presented here; these experiments focus on measuring the similarity of the output Japanese synthetic speech to the speech of the original English speaker in order to assess and evaluate the performance of the proposed unsupervised cross-lingual speaker adaptation technique. In addition, we investigated whether restricting the features on which the KLD is calculated affects the quality of the output speech. Instead of using 120 mel-cepstral coefficients (including statics, deltas and delta-deltas), only the first 10 static mel-cepstral coefficients were used.

The article is organized as follows. Section 2 gives details of the EMIME speech-to-speech translation system using HMM-based ASR and TTS. In Section 3, an overview of the unsupervised cross-lingual speaker adaptation method adopted is given. Section 4 describes the experimental set-up that we used to analyze and evaluate the system. The analysis of the proposed cross-lingual speaker adaptation method, i.e., an analysis of the KLD output is given in Section 5. This is followed in Section 6 by the results of the listening tests. Finally, Section 7 summarizes our findings and gives suggestions for future work.

2. Overview of the EMIME speech-to-speech translation system

Figure 1 shows a diagram of the EMIME speech-to-speech translation system. It comprises HMM-based ASR, HMM-based TTS, MT, and cross-lingual speaker adaptation (CLSA). A short description of each of these components is given here.

All acoustic models, for both HMM-based ASR and TTS, are trained on large conventional speech databases, comprising speech from hundreds of speakers, which were originally intended for ASR: Wall Street Journal (WSJ0/1) databases for English (Paul and Baker, 1992), Speecon databases for Mandarin and Finnish (Iskra *et al.*, 2002), and the JNAS database for Japanese (Itou *et al.*, 1998). Details of the front-end text processing used to derive phonetic-prosodic labels from the word transcriptions can be found in Yamagishi *et al.* (2010).

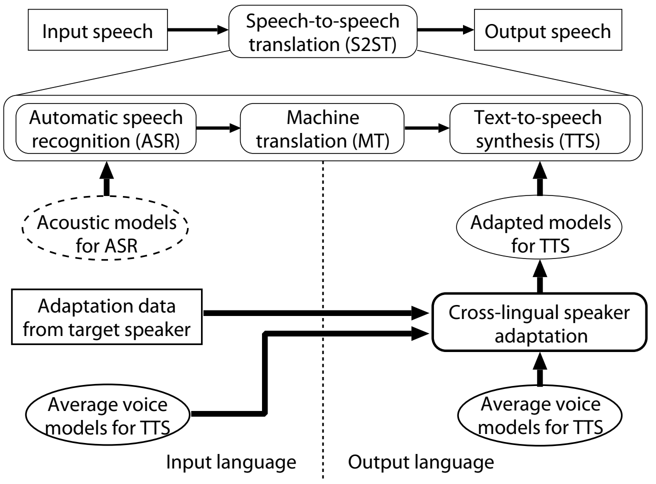
Input speech
Speech-to-speech translation (S2ST)
Output speech
Automatic speech recognition (ASR)
Machine translation (MT)
Text-to-speech synthesis (TTS)
Acoustic models for ASR
Adapted models for TTS
Adaptation data from target speaker
Cross-lingual speaker adaptation
Average voice models for TTS
Average voice models for TTS
Input language
Output language

Figure 1: Overview of the EMIME Speech-to-Speech Translation system using HMM-based ASR and TTS.

For ASR of each language, 3-state no-skip triphone speaker-independent HMMs are trained. Either MFCCs or Perceptual Linear Predictive (PLP) cepstral coefficients (Hermansky, 1990) can be used as the acoustic features for ASR. The ASR language models used for English, Mandarin and Japanese each contain about 20k bi-grams; the language model for Finnish is a word 10-gram plus a morph bi-gram (Hirsimäki *et al.*, 2009). They are smoothed using the standard Kneser-Ney method (Kneser and Ney, 1995).

For TTS of each language, 5-state no-skip context-dependent speaker-independent MSD-HSMMs (Tokuda *et al.*, 2002; Zen *et al.*, 2007b) are trained as "average voice models" using speaker-adaptive training (SAT) (Anastasakos *et al.*, 1996; Gales, 1998). For the state tying (Young *et al.*, 1994), minimum description length (MDL) automatic decision tree clustering is used (Shinoda and Watanabe, 2000). TTS acoustic features comprise the spectral and excitation features required for the STRAIGHT (Kawahara *et al.*, 1999) mel-cepstral vocoder (Tokuda *et al.*, 1994) with mixed excitation (McCree and Barnwell III, 1995; Kawahara *et al.*, 2001).

For unsupervised cross-lingual speaker adaptation and decoding, a multi-pass framework is used:

1. In the first pass, initial transcriptions are obtained using "Juicer" (Moore *et al.*, 2006), a weighted finite state transducer (WFST) decoder with speaker independent (SI) HMMs.
2. In the second pass, constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation (Yamagishi *et al.*, 2009a) is applied to SAT-HMMs (ASR) using the hypotheses obtained in the first pass.

3. In the third pass, using these adapted models, the speech is decoded again and the transcriptions are refined.
4. In the final pass, CSMAPLR transforms are estimated for SAT-HSMMs (TTS) with the refined transcriptions.
5. Finally, these transforms are applied to the SAT-HSMMs for the output language, by employing a state-level mapping that has been constructed based on the Kullback-Leibler divergence (KLD) between pairs of states from the input and output TTS HMMs (Wu and Tokuda, 2009). Details of this state-mapping are given in the next section.

Note that EMIME did not focus on translation technology research. This was a deliberate choice, to allow us to concentrate on ASR and TTS research. Therefore, for the MT module, we simply used Google translation provided via their AJAX language APIs[5]. This translator only provides the 1-best result.

Finally, the speech waveform is output in the TTS module. Acoustic features (spectral and excitation features) are generated from the adapted HSMMs in the output language using a parameter generation algorithm that considers the global variance (GV) of a trajectory (Toda and Tokuda, 2007). Then, mixed excitation signals are produced using a mel-logarithmic spectrum approximation (MLSA) filter (Fukada *et al.*, 1992) which corresponds to the generated STRAIGHT mel-cepstral coefficients. These vocoder modules are the same as Zen *et al.* (2007a).

3. Cross-lingual speaker adaptation based on a state-level transform mapping learned using minimum KLD

A cross-lingual adaptation method based on a state-level mapping, learned using the KLD between pairs of states, was proposed by Wu and Tokuda (2009) and is summarized here. We call this approach "state-level transform mapping." 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.

3.1. Learning the state-mapping

The mapping between the input language and output language states are learned as follows. For each state $^{\forall}j \in [1, J]$ in the output language HMM λ_{output}, we search for the state $\widehat{i}$ in the input language HMM λ_{input} with the minimum symmetrized KLD to state j in λ_{output}:

$$\widehat{i} = \underset{1 \leq i \leq I}{\operatorname{argmin}} \, D_{\text{KL}}\left(j, i\right), \tag{1}$$

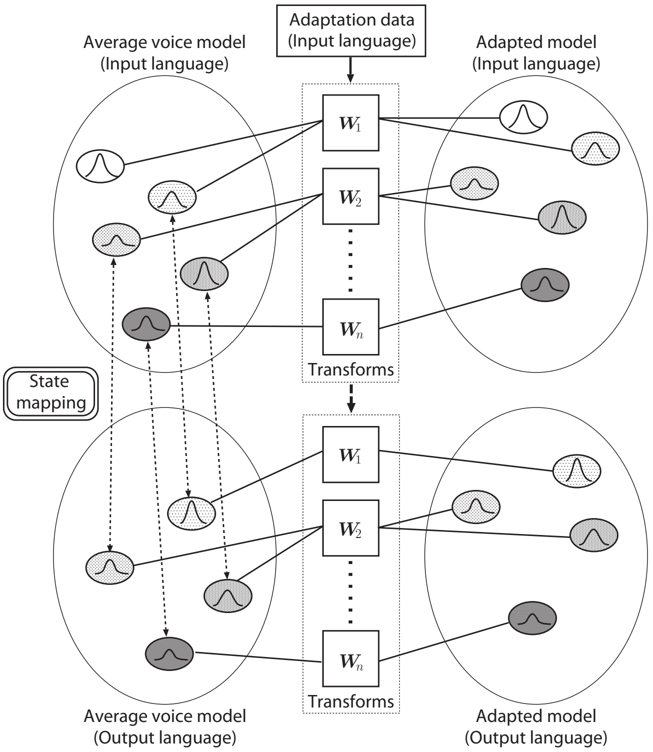

Figure 2: Graphical representation of the state-level mapping using minimum KLD between input and output language HMMs.

where λ_{output} has J states and $D_{\text{KL}}\left(j, i\right)$ represents the KLD between state i in λ_{input} and state j in λ_{output} (Figure 2). $D_{\text{KL}}\left(j, i\right)$ is calculated as in Qian *et al.* (2009):

$$\begin{aligned} D_{\text{KL}}\left(j, i\right) \approx & D_{\text{KL}}\left(j \mid\mid i\right) + D_{\text{KL}}\left(i \mid\mid j\right), & (2) \\ D_{\text{KL}}\left(i \mid\mid j\right) = & \frac{1}{2} \ln \left(\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} \right) - \frac{D}{2} + \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}_j^{-1} \boldsymbol{\Sigma}_i\right) & \\ & + \frac{1}{2} \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right)^{\top} \boldsymbol{\Sigma}_j^{-1} \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right), & (3) \end{aligned}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent the mean vector and covariance matrix of the Gaussian pdf associated with state i.

3.2. Estimating the input language HMM transforms

Next, we estimate a set of *state-dependent* linear transforms $\widehat{\Lambda}$ for the input language HMM λ_{input} in the usual way:

$$\begin{aligned} \widehat{\Lambda} &= \left(\widehat{\boldsymbol{W}}_1, \ldots, \widehat{\boldsymbol{W}}_I\right) \\ &= \underset{\Lambda}{\operatorname{argmax}} \, P\left(\boldsymbol{O} \mid \lambda_{\text{input}}, \Lambda\right) P\left(\Lambda\right), \end{aligned} \tag{4}$$

where $\boldsymbol{W}_i$ represents a linear transform for state i, I is the number of states in λ_{input}, and $\boldsymbol{O}$ represents the adaptation data. $P\left(\Lambda\right)$ represents the prior distribution of the linear transform for CSMAPLR (Yamagishi *et al.*,

[5] http://code.google.com/intl/ja/apis/ajaxlanguage/

2009a). Note that the linear transforms will usually be tied (shared) between groups of states known as regression classes, to avoid over-fitting and to enable adaptation of all states, including those with no adaptation data.

3.3. Applying the transforms to the output language HMM

Finally, these transforms are mapped to the output language HMM. The Gaussian pdf in state j of λ_{output} is transformed using the linear transform for state $\widehat{i}$, which is transform $\widehat{\boldsymbol{W}}_{\widehat{i}}$. By transforming all Gaussian pdfs in λ_{output} in this way, cross-lingual speaker adaptation is achieved.

3.4. Unsupervised cross-lingual adaptation

We can extend this method to unsupervised adaptation simply by automatically transcribing the input data using ASR-HMMs. For supervised adaptation, λ_{input} and λ_{output} are both TTS-HMMs (for the input and output languages, respectively). For unsupervised adaptation of HMM-based speech synthesis, λ_{input} may be either a TTS-HMM, or an ASR-HMM that utilizes the same acoustic features as TTS. When the ASR-HMM uses Gaussian mixtures, we can use an approximated KLD (Goldberger *et al.*, 2003). No other constraints need to be placed on the ASR-HMM. In particular, it does not need to use prosodic-context-dependent-quinphones (which are necessary for TTS models).

3.5. Efficient methods for calculating the KLD

The state-mapping is learned by searching for pairs of states that have minimum KLD between input and output language HMMs. The computational cost is huge because KLD calculation of all combinations of states in both language HMMs is required.

In Dines *et al.* (2010) it was shown that the use of the lower dimensional part of mel-cepstral STRAIGHT coefficients (e.g., 1st to 13-th dimensions of a 40-dimensional mcep) is sufficient for recognizing phonemes in an ASR system. It was also found that using the higher dimensional mel-cepstral coefficients results in higher word error rates for ASR. For TTS it was shown that the use of the higher dimensional mel-cepstral coefficients increases naturalness, as evaluated using mean opinion scores (MOS). From Dines *et al.* (2010) it can be concluded that the higher mel-cepstral dimensions mainly contribute to speaker identity and naturalness rather than phoneme identity.

The KLD state-mapping is calculated between the average voice models of input and output languages, i.e., it is learning the mapping between two languages. This type of mapping concerns phoneme identity rather than speaker identity and naturalness, therefore, it seems that disregarding the higher dimensional mel-cepstral coefficients may be possible without affecting the state-mapping outcome in a negative way. To investigate this and as a solution to the computational cost associated with KLD on the full feature vector, we restrict the number of mel-cepstral dimensions for KLD calculation. The proposed method eliminates delta and high dimensional mel-cepstral coefficients as phoneme identity information is available in the static and low dimensional mel-cepstral coefficients.

Although log F0 and aperiodicity features are used for speaker adaptation in the same way as the mel-cepstral coefficients, this technique of reducing computational cost were used for only mel-cepstral coefficients. We explored the effect of the following KLD calculations:

- KLD calculation using 120-dimensions (40-dim static, 40-dim delta, 40-dim delta-delta)
- KLD calculation using only the first 20 of the 40 static dimensions
- KLD calculation using only the first 10 of the 40 static dimensions

The low dimensional mel-cepstral coefficients (i.e. the first 10) contain more information than higher dimensional mel-cepstral coefficients (Imai, 1983). Furthermore, the static features also contain more information than the dynamic features (Yu *et al.*, 2008).

4. Experimental setup

We performed experiments on English-to-Japanese speaker adaptation for HMM-based speech synthesis. First, specifics on the data that was used to analyze the KLD state-mapping are given. Next, the set-up of the perceptual experiments is described.

4.1. Models and data for KLD analysis

The objective of the analysis is to illustrate the effectiveness of the KLD state-mapping in phonetic and speaker similarity terms. KLD simply measures divergences between HMM states. No explicit linguistic or phonetic knowledge is used. In order to get an idea of the phonetic appropriateness of the mapping, we compare the vowel triangle in Japanese for the average voice model, a male and a female voice. Next, we compare the vowel spaces for cross-lingual speaker adapted Japanese and speaker-dependent American-English TTS for a single male speaker. We also present a comparison between this male speaker and a group of 60 male American speakers. F1 vs F2 - F1 space (whose dimensions are the first formant vs the difference between the second and first formants) is used to examine phonetic properties. F1 vs F2 - F1 space results in a closer visual correspondence between the formant plots and the IPA vowel chart than F1 vs F2 space (Ladefoged and Maddieson, 1996). F0 vs F1 space is used to examine speaker identity.

Table 1: The word accuracy of the ASR systems (second pass)

	Number of sentences		
speaker	5	50	2000
001	85.32	87.54	85.63
002	86.88	87.08	84.78

4.2. Training and adaptation data

An English speaker-independent model for ASR and average voice model for TTS were trained on the pre-defined 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). The two average voice models were used to learn the KLD state-mapping.

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. They are named 001 and 002 in the following experiments, respectively.

2000 randomly chosen English sentences (about 2 hours in duration) uttered by 001 and 002 were selected from the "long term" subset of the WSJ0 corpus. These 2000 sentences were used as the two speaker's adaptation data. This data was used as adaptation data to adapt the Japanese average voice model to speaker 001 and 002. This data was also used to create speaker-dependent acoustic English TTS models for speaker 001 and 002. This made it possible to compare their English and Japanese synthetic vowel spaces to each other.

4.3. Features and acoustic models

Speech signals were sampled at a rate of 16 kHz and windowed by a 25ms Hamming window with a 10 ms shift for ASR and by an F0-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 zero-th coefficient), log F0, 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 multi-stream MSD-HSMMs for TTS. Each state had 16 Gaussian mixture components for ASR and a single Gaussian for TTS. The word recognition accuracy in the second pass of the ASR system, which is used for TTS unsupervised speaker adaptation in the third pass, is shown in Table 1. Although the accuracy is not very high, the ASR system uses only very standard techniques and is an adequate benchmark system for comparing the differences between supervised and unsupervised adaptation for TTS.

4.4. Speaker adaptation

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 hyper-parameter (τ_b in Yamagishi *et al.* (2009a)) of CSMAPLR to a 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).

We applied the CSMAPLR transforms $\boldsymbol{W}_i$ to the Gaussian pdfs of the output language HMMs using the proposed KLD-based state-level mapping. For the transform mapping in the MSD streams that have both voiced and unvoiced spaces for the F0 modelling, the KLD calculation was conducted between a pair of Gaussian pdfs in the voiced space; Qian *et al.* (2009) calculates KLD using both voiced and unvoiced spaces.

5. Analysis of KLD state-mapping

5.1. Speech material for KLD analysis

Japanese synthetic speech was generated using the Japanese average voice model and the two speakers, in other words the cross-lingual adapted speaker 001 and 002 models. As we were interested in measuring vowel formants, 50 sentences containing each of the Japanese vowels in Table 2 were generated. This gave us about 2000 vowel tokens per vowel to analyze. For each of the speakers, 2000 sentences of English adaptation data were used.

The F1 and F2 values of the vowels were measured using the Snack Sound Toolkit (Sjölander *et al.*, 1998; Sjölander and Beskow, 2000). The algorithm for formant extraction used in Snack applies dynamic programming to select and optimize a formant trajectory from multiple candidates which are obtained by solving for the roots of the linear predictor polynomial (poles of a filter).

For the speaker-dependent comparison between English and Japanese vowel spaces, 001's synthetic English and Japanese was used. The same vowels as above for Japanese were used and 50 English sentences were generated. This gave us 5315 English vowel tokens to analyze.

Our final analysis looking at the KLD output is a comparison between male speaker 001 and a group of 60 other male American speakers, in F0 vs F1 vowel space. The 60 male speakers were selected from the "short term" subset of the WSJ0 corpus. For each speaker, approximately 150 sentences were available. The sentences were all manually transcribed and the vowels were segmented using forced-alignment. The speakers all utter different sentences. F0 and F1 values for each of the 60 speakers were calculated at the midpoint of each vowel and we took the mean over all vowel tokens (once again using the Snack toolkit). We used the synthetic speech from Yamagishi *et al.* (2010) for the 60 speakers.

Table 2: General American English and Japanese monophthongs in the two TTS systems. (IPA notation)

	English		Japanese	
Features	IPA	Example	IPA	Example/meaning
close front	i	fleece	i	oj**i**san (uncle)
near close front	ɪ	kit		
close-mid front	e	waist	e	s**e**to (city in Japan)
open-mid front	ɛ	dress		
near open front	æ	trap		
mid central	ə	comm**a**		
open central			a	ob**a**san (aunt)
close back	u	goose	ɯ	y**u**ki (snow)
near close back	ʊ	foot		
close-mid back			o	t**o**ri (bird)
open-mid back	ʌ, ɔ	strut, cloth		
open back	ɑ	lot		

We also calculated F0 and F1 values for speaker 001's English synthetic speech and Japanese synthetic speech which was achieved by cross-lingual speaker adaptation based on his English speech data (2000 adaptation sentences). 50 Japanese sentences were generated, the vowels were segmented and measured in the same way as described above.

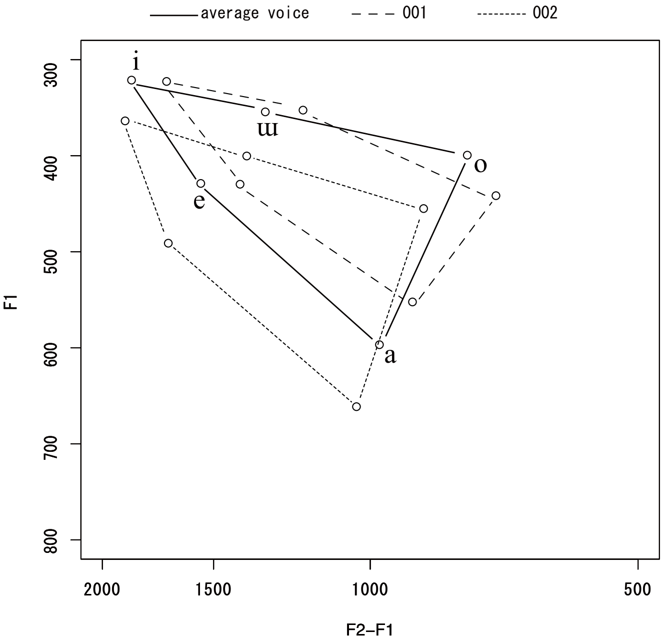

Figure 3: Japanese vowel triangles of the average voice, male speaker 001 and female speaker 002. For the male and female speaker, cross-lingual speaker adaptation is applied. Their adaptation data is English speech data.

5.2. Phonetic analysis – Vowels

One of great advantages of the state-mapping cross-lingual adaptation used in these experiments is that the technique is applicable to any acoustic models that use the same acoustic features, regardless of phoneme and contextual differences. We can effortlessly apply the mapping of linear transforms between English and Japanese acoustic models even though the models are based on completely different TTS text-processing modules and the languages share only a limited amount of similar sounds.

Table 2 shows the English and Japanese monophthongs which are used in the two text-processing modules (represented here in IPA notation) (Fitt, 2000; Yoshimura *et al.*, 1999). It can be seen that there appears to be little overlap between Japanese and English vowel sets, according to the IPA.

Not shown in Table 2, but certainly also of interest, is the different way in which diphthongs and long vowels are described in the two languages. From a phonological perspective, generally speaking, short vowels are counted as one unit and long vowels and diphthongs as two. However in phonetics, this is arguably not the case. For English listeners, long vowels and diphthongs in English are perceived as one indivisible unit. Moreover, English listeners also treat diphthongs in Japanese as one indivisible unit (Yoneyama, 2004). In Japanese, however, due to the influence of mora (rather than syllable structure) long vowels and diphthongs are divisible, and can be perceived as two units by Japanese listeners (Yoneyama, 2004; Tsujimura, 2006). Consequently, in English front-end TTS processing diphthongs and long vowels are treated as distinct phonemes (i.e., additions to the vowel set, not combinations of monophthongs) whereas in Japanese front-end TTS processing diphthongs and long vowels are described by sequences of monophthongs.

It is interesting to confirm that the transforms estimated for the input language data are being applied to the output language vowels in a phonetically appropriate way. If this is indeed the case, it indicates that the KLD state-mapping is functioning as intended. Figure 3 shows the shift in Japanese vowel triangles in F1 vs F2 - F1 space after applying cross-lingual speaker adaptation (for details of formant measurements see Section 5.1).

The normal line represents that of the average voice, the broken lines represent that of male speaker 001 and female speaker 002. Figure 3 shows that the vowel triangles are roughly similar in size and shape even after the application of cross-lingual speaker adaptation. This is despite the large differences between English and Japanese vowel spaces.

To get an indication of the distance between a single target speaker's vowels in the two languages, we measured F1 and F2 values for speaker 001's synthetic Japanese and English vowel tokens. The mean values per vowel are shown in Figure 4 for English and Japanese synthetic vowels in F1 vs F2 - F1 space. 50 sentences including 5315 vowels were used for calculating the average. Circle markers indicate synthetic Japanese vowels after applying cross-lingual speaker adaptation using male speaker 001's English adaptation data. Cross markers show speaker 001's

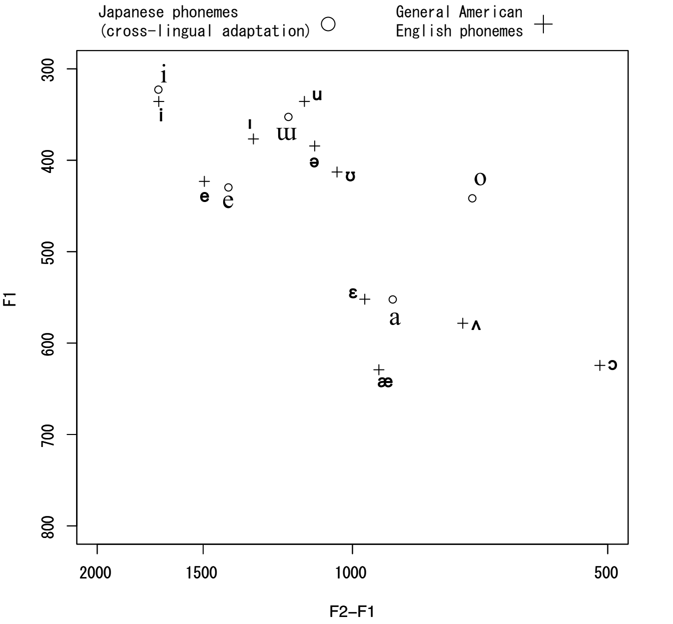

Figure 4: Japanese and English vowels of male speaker 001 in F1 vs F2 - F1 space. Japanese vowels were created by applying cross-lingual speaker adaptation using English speech data.

English synthetic vowels which were generated using the speaker-dependent English TTS acoustic models.

From this figure, we can first see that the two phonemes which are represented by the same IPA symbol in Japanese and English – /i/ and /e/ – are also located near to each other in the F1 vs F2 - F1 vowel space. Recall that the KLD mapping algorithm does not utilize any phonetic or linguistic knowledge at all: it simply measures the KLD between two Gaussians pdfs and then the linear transform estimated from the English acoustic models is applied to the corresponding states of the Japanese model, thus performing an affine transform of mel-cepstral acoustic space in Japanese. We see that the affine transform results in close F1 vs F2 - F1 values for these two vowels. This is an indication that the state-mapping cross-lingual adaptation behaves in a way that is consistent with phonetic knowledge.

Figure 4 also shows that the other Japanese vowels – which do not have a direct match in English (in the IPA representation) are transformed to phonetically reasonable places. For instance, we see that the Japanese vowel /a/ achieved by cross-lingual adaptation lies between the English vowels /æ/, /ʌ/ and /ɛ/, which closely mirrors the IPA vowel chart.

5.3. Comparison with 60 different English speakers in the F0 vs F1 space

An F0 vs F1 vowel space can be viewed as a low dimensional perceptual space which matches listeners discrimination between different speakers to a certain extent (Baumann and Belin, 2010). It can also be used to illustrate degree of speaker similarity between different speakers. To illustrate the effectiveness of our cross-lingual speaker adaptation from English to Japanese we compared speaker 001's English and Japanese speech in F0 vs F1 space to 60 other male English speakers. Note that our HMMs have both mel-cepstral features and log F0 and these are simultaneously transformed into those of the target speaker by our cross-lingual speaker adaptation.

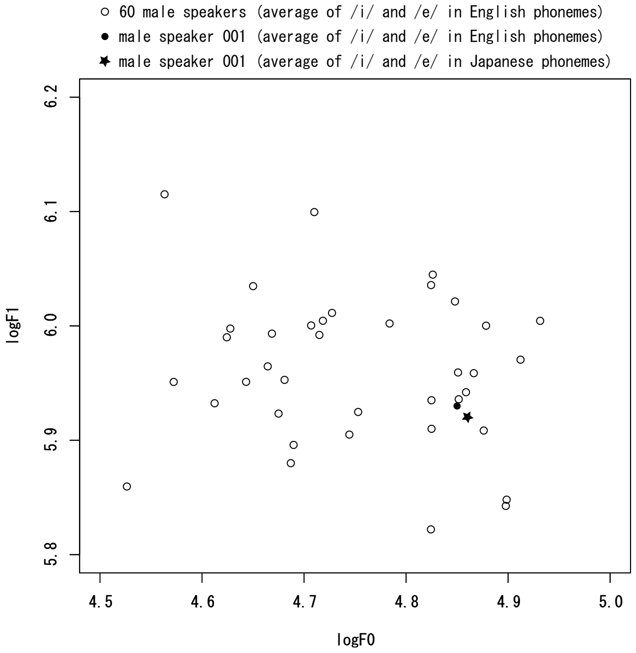

Figure 5: 60 male English speakers and male speaker 001 in English and Japanese in the log F0 vs F1 space. These points were calculated only from two common phonemes of English and Japanese – /i/ and /e/ –. The 60 different speakers are represented by white points and speaker 001 in English and Japanese are represented by black points and black stars, respectively.

The method we used to measure F0 vs F1 data points for each of the speakers is described in Section 5.1. These points were calculated only from the phonemes which English and Japanese have in common – /i/ and /e/ –.

The results are shown in Figure 5 where we can see English and Japanese versions of the 001 synthetic voices are close to each other, compared to the data points for the other 60 speakers. Note that these points represent the averages of log F0 and log F1 values calculated from the two common phonemes. This result supports our claim that the state-mapping cross-lingual adaptation achieves a high degree of speaker similarity between the synthetic speech of a targeted speaker in two different languages at the segmental level (as far as vowels are concerned).

5.4. Phonetic analysis – Consonants

Table 3 shows the consonants used in our experiments. In contrast to vowels, where only two phonemes are shared

Table 3: English and Japanese consonants according to the IPA consonant chart.

Features	English	Japanese
voiceless bilabial plosive	p	p
voiced bilabial plosive	b	b
voiceless coronal plosive	t	t
voiced coronal plosive	d	d
voiceless velar plosive	k	k
voiced velar plosive	g	g
voiced labiodental nasal	ɱ	
voiced bilabial nasal	m	m
voiced coronal nasal	n ɳ	n
voiced velar nasal	ŋ	
voiced coronal tap or flap	ɾ	
voiceless labiodental fricative	f	f
voiced labiodental fricative	v	
voiceless dental fricative	θ	
voiced dental fricative	ð	
voiceless alveolar fricative	s	s
voiced alveolar fricative	z	z
voiceless postalveolar fricative	ʃ	ʃ
voiced postalveolar fricative	ʒ	
voiceless glottal fricative	h	h
voiced coronal approximant	ɹ	ɹ
voiced palatal approximant	j	j
voiced velar approximant		ɰ
voiceless labialized velar approximant	ʍ	
voiced labialized velar approximant	W	
voiced coronal lateral approximant	l	
voiceless alveolar affricate		t͡s
voiceless postalveolar affricate	ʧ	
voiced postalveolar affricate	ʤ	
voiceless alveolo-palatal affricate		ʨ
voiced alveolo-palatal affricate		ʥ

between English and Japanese, there are relatively many shared consonants. We calculated phoneme level KLDs for the phonemes shared across languages and verified whether the pairs with minimum KLD corresponded to the same consonants in both languages. The accuracy achieved was 45%. This means that about half of the mapping 'rules' automatically learned by the KLD without any linguistic knowledge are phonetically plausible. One might feel that this accuracy is not good enough for cross-lingual speaker adaptation. We therefore analyzed the errors. We checked the N-best results of the KLD mapping and found that most of the errors happened due to misjudgment of voiced and unvoiced categories such as unvoiced bilabial plosive /p/ and voiced bilabial plosive /b/.

This can be explained perfectly well by the theoretical limitations of the current KLD calculation strategy: We calculated the KLD per Gaussian per feature. In other words, we did not utilize F0 values and voicing information for the KLD calculation of spectral features. Therefore the mapping rules learned from the KLD between Gaussians for spectral features cannot represent any voicing categories and as a consequence the confusion between voiced and unvoiced categories happens frequently. Development of better learning methods for mapping 'rules' across not just spectral but also source and voicing features is an important future task.

6. Results for the listening tests

6.1. Perceptual Experiments

To assess and evaluate our method perceptually, we performed several perceptual experiments. The aims of the perceptual experiments are 1) to confirm that the state-mapping approach to cross-lingual adaptation can improve speaker similarity in the output language, 2) to assess the differences between supervised and unsupervised adaptation and 3) to confirm that the proposed method for efficient KLD estimation does not reduce the quality of the synthetic speech. The first and second aims were investigated in the first listening test. The third aim was assessed separately but using the same stimuli.

First, experiments on supervised and unsupervised English-to-Japanese speaker adaptation for HMM-based speech synthesis were performed. Synthetic stimuli were generated from seven models: the average voice model and supervised or unsupervised adapted models each with 5, 50, or 2000 English adaptation sentences.

Ten Japanese native listeners participated in the two listening tests for speaker similarity judgement and intelligibility tasks. In the speaker similarity judgement task, each listener was presented with 12 pairs of sentences in random order: the first sample in each pair was a reference original utterance from the database (English) and the second was a synthetic speech utterance generated using one of the seven models (Japanese). 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 on a 5-point scale. As no Japanese speech data were available for the target English speakers, the reference utterances were in English. The text for the 12 Japanese sentences in the listening test for the speaker similarity task comprised six written Japanese news sentences randomly chosen from the Mainichi corpus and six spoken English news sentences from the English adaptation data that had been recognized using ASR and then translated into Japanese text using MT. In the intelligibility task, the listeners heard semantically unpredictable sentences (SUS) (Benoit *et al.*, 1996) and were asked to type in what they heard. Typographical errors and spelling mistakes were allowed for in the scoring procedure.

6.2. Subjective evaluation results for cross-lingual speaker adaptation – speaker similarity

Figure 6 shows the average DMOSs and 95% confidence intervals. First of all, we can see that the adapted voices are judged to sound more similar to the target speakers than the average voice. However, the figure also shows

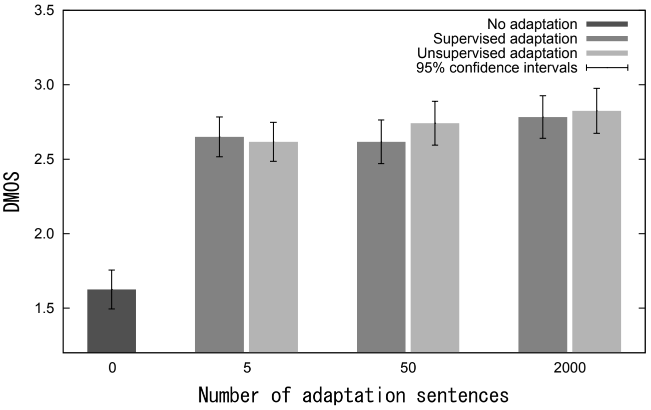

Figure 6: DMOS results for supervised and unsupervised speaker adaptation. “0 sentences” refers to the unadapted average voice model for Japanese.

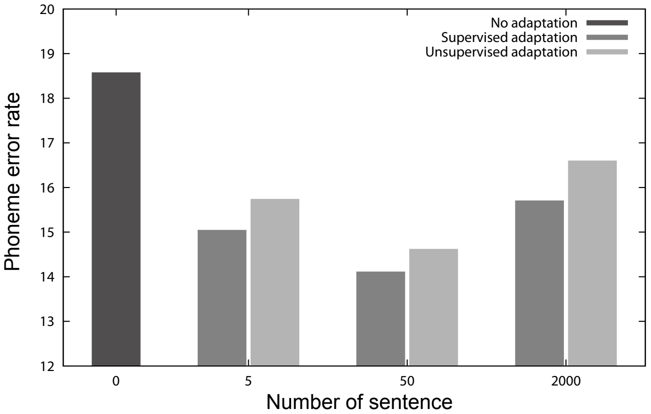

Figure 8: Phoneme error rates for supervised and unsupervised speaker adaptation. “0 sentences” refers to the unadapted average voice model for Japanese.

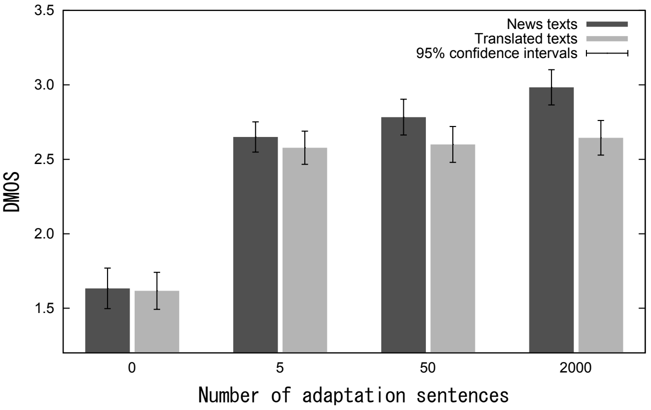

Figure 7: DMOS results for Japanese news texts chosen from the corpus and English news texts which were recognized by ASR then translated into Japanese by MT. “0 sentences”refers to the unadapted average voice model for Japanese.

that the maximum scores are less than three. Our earlier analysis on vowels (Section 5.2) showed that the state-mapping cross-lingual adaptation does seem to change the speaker similarity of synthetic speech in F0 vs F1 space to match that of a target speaker well at the segmental level (for vowels). We hypothesize that the reason this does not translate to higher speaker similarity scores in this experiment is 1) the gap between natural speech and synthetic speech and 2) the gap between English and Japanese. As references for judging the degree of speaker similarity of the synthetic speech to the original speaker, we used natural speech. However, it has been shown that there is a significant degradation in a listener’s ability to decide on speaker similarity when comparing natural and synthetic speech stimuli (Wester and Karhila, 2011). The task here is further made more complex by requiring the listeners to rate speaker similarity across languages. This has also been shown to affect speaker similarity rating significantly (Wester, 2010). These two factors combined explain why speaker similarity scores were not higher.

Next, we can see that the differences between supervised and unsupervised adaptation are very small. This is a positive outcome because real-world applications of these techniques would most likely need to use unsupervised adaptation. A somewhat puzzling result however is that the amount of adaptation data has a relatively small effect. This requires further investigation in future work.

In the DMOS test, two different types of sentence were synthesised and presented to the subjects: fluent sentences chosen from the Japanese news text corpus; sentences that had been recognized using ASR and then had been translated from English into Japanese. To clarify the effect of the text types used in speech synthesis, we then analyze the scores of Figure 6 in a different way. Figure 7 shows the average scores using Japanese news texts from the Mainichi 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. Interestingly the gap becomes larger as the number of adaptation sentences increases; a parallel investigation was performed and we found that fluency of translated texts affect synthetic speech. For details, refer to Hashimoto *et al.* (2011a,b).

6.3. Subjective evaluation results for cross-lingual speaker adaptation – Intelligibility

Figure 8 shows the phoneme error rates of the SUS sentences used in the intelligibility test. First of all, very interestingly, we can see that all the adapted voices have better phoneme error rates than that of the average voice. To investigate this initially surprising result, we compared the adapted voices with the average voice and found out that the adapted models always have smaller variance than that of the average voice model. Note that CSMAPLR transforms not only the mean vectors but also the variance matrices of Gaussian pdfs of the average voice model. Figure 9 shows the average of the diagonal components

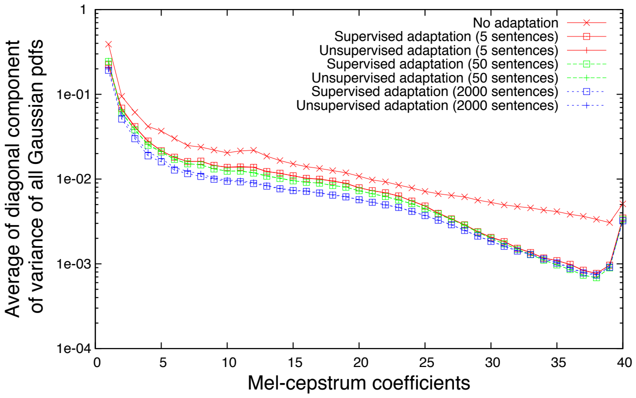

Figure 9: Comparison of the average of dialogue components of variance matrices of all Gaussian pdfs.

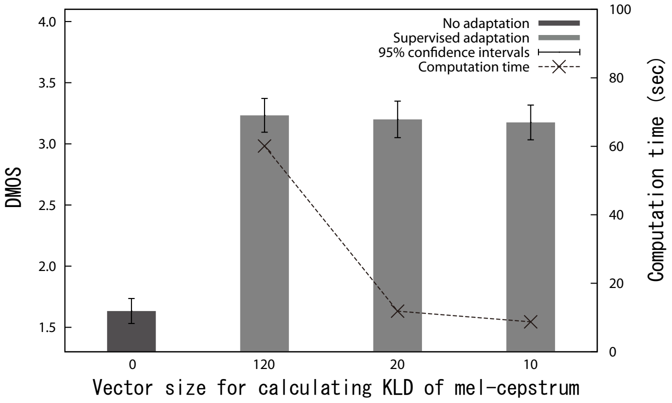

Figure 10: Experimental results: DMOS comparison of dimension for KLD calculation.

of the covariance matrices of all Gaussian pdfs for mel-cepstra of the average voice model and adapted models using 5, 50 and 2000 sentences, in supervised and unsupervised manners. We can see that for all dimensions of the mel-cepstra, variance becomes smaller after speaker adaptation. We hypothesize that this smaller variance causes the generated mel-cepstral trajectories to be more ‘prototypical’ and hence more intelligible, as reflected in the better phoneme error rates.

We can also see that voices using unsupervised adaptation always have worse phoneme error rates than ones using supervised adaptation. This is not surprising because we adapted the voices using automatically transcribed sentences that have typically 13% to 15% word error rate. However, it is worth emphasising that the increase in phoneme error rate is just 1% absolute.

6.4. Results of listening test for efficient KLD calculation

Experiments investigating the effect of using restricted order mel-cepstral coefficients on the KLD calculation of state-mapping were performed. Although the number of mel-cepstral coefficients for calculation of KLD was different, the number of log F0 coefficients for calculation of KLD was not different. Ten Japanese native listeners participated in the listening test. They carried out a DMOS test: after listening original speech and synthetic speech, the subjects were asked how similar to the target speaker’s identity. Experimental methods were described in section 3.5. Once again the reference utterances were English. In the speaker similarity judgement and intelligibility tasks, Japanese news sentences randomly chosen from the Mainichi corpus and SUSs, respectively, were used as the sentences in this listening test. Synthetic stimuli were generated from the average models and the supervised adapted models with 2000 utterances.

Figure 10 plots DMOSs. The results show that speaker similarity of the speech samples with low dimensional state-mapping achieve the same level as the baseline. Comparing “120” and “10” in the figure, we see that computational cost required for mel-cepstrum state-mapping was reduced by about 90 percent without any detrimental effect on the quality of the synthetic speech. To further underpin this result, we analysed and compared the KLD state-level mapping rules before and after we restricted the order of mel-cepstral coefficients. We found that when we restrict the order of mel-cepstral coefficients to be used for KLD to “20” or “10”, 21% and 16% of state-mapping rules acquired are identical to those using all dimensions, respectively. Although the number of pairs shared between the mappings generated by the baseline and the proposed methods is small because of the different criterion, it can be seen from the figure that an appropriate state-mapping was still found by the proposed method.

Figure 11 shows the subjective phoneme error rates of the SUS sentences in the intelligibility test. We can see that using restricted order mel-cepstral coefficients for the KLD calculation of state-mapping does not degrade the intelligibility of the adapted voices. In fact, restricting the order of the mel-cepstral coefficients to 20 was found to slightly increase intelligibility.

7. Conclusions

In this paper, several developments have been integrated into a single architecture which achieves unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. We demonstrate an end-to-end speech-to-speech translation system built for four languages (English, Finnish, Mandarin, and Japanese). The phonetic analysis supports the finding the state-mapping cross-lingual adaptation achieves a high degree of speaker similarity between the synthetic speech of a targeted speaker in two different languages at the segmented level. The listening tests for English-to-Japanese adaptation demonstrate that the adapted voices sound more similar to the target speaker than the average voice and that differences

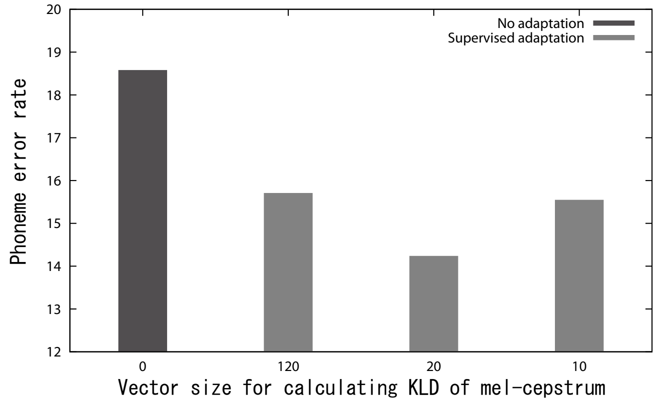

Figure 11: Experimental results: intelligibility comparison of dimension for KLD calculation.

between supervised and unsupervised cross-lingual speaker adaptation are small in terms of both the quality and intelligibility of the synthetic speech. Using the proposed efficient KLD calculation method, the computational cost of finding the mel-cepstrum state-mapping is significantly reduced without any detrimental effect on the quality and intelligibility of the synthetic speech.

We have not addressed the question of whether the cross-lingual adapted voices should sound like a true bilingual or an adult second language learner (with an obvious 'foreign accent'). Our instructions to the subjects were simply to rate how similar they thought the synthesized speech was to the original speaker. In parallel with the research presented in this paper, other research has been investigating the above issues. For more details, please refer to Wester (2010); Wester and Karhila (2011); Tsuzaki *et al.* (2011).

Since December 2002, we have made regular public releases of an open-source software toolkit named "HMM-based speech synthesis system (HTS)" to provide a research and development platform for statistical parametric speech synthesis. Various organizations currently use it to conduct their own research. HTS version 2.2 was released in December 2010 and supports the cross-lingual adaptation method based on state-level mapping, learned using the KLD between pairs of states.

Future work includes unsupervised cross-lingual speaker adaptation using linear transforms estimated directly by ASR-HMMs, which must therefore use the same acoustic features as TTS-HSMM and to use an approximated KLD to efficiently measure the distance between pairs of Gaussian mixtures, necessitated by the fact that ASR-HMMs typically use Gaussian mixture output densities (Goldberger *et al.*, 2003). Spectral state mapping that uses the voicing information to improve the linguistic accuracy of the KLD mapping, especially for consonants, as found in Section 5.4, is also part of our future work. In this paper, we had performed experiments only on English-to-Japanese speaker adaptation. The other language pairs should also be evaluated. Speech samples used in this experiments are available online (`http://www.sp.nitech.ac.jp/~uratec/clsa.html`).

8. Acknowledgements

The authors thank Kaori Yutani and Xiang-Lin Peng of the Nagoya Institute of Technology for their help with the experiments reported in this paper. Special thanks go to Akinobu Lee and Yoshihiko Nankaku for their technical support and helpful discussions.

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), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

References

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proc. ICSLP-96*, pages 1137–1140, Philadelphia, PA.

Baumann, O. and Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, **74**, 110–120. 10.1007/s00426-008-0185-z.

Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**(4), 381–392.

Dines, J., Saheer, L., and Liang, H. (2009). Speech recognition with speech synthesis models by marginalising over decision tree leaves. In *Proc. Interspeech 2009*, Brighton, UK.

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

Fitt, S. (2000). Documentation and user guide to unisyn lexicon and post-lexical documentation and user guide to UNISYN lexicon and post-lexical rules. Technical report, Centre for Speech Technology Research, University of Edinburgh.

Fukada, T., Tokuda, K., and Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. ICASSP 1992*, pages 137–140, San Francisco, CA.

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

Gao, Y. (2003). Coupling vs. unifying: Modeling techniques for speech-to-speech translation. In *Proc. EUROSPEECH 2003*, pages 365–368, Geneva, Switzerland.

Gibson, M. (2009). Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models. In *Proc. Interspeech 2009*, pages 1791–1794, Brighton, U.K.

Goldberger, J., Gordon, S., and Greenspan, H. (2003). An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *ICCV'03*, pages 487–493.

Hashimoto, K., Yamagishi, J., Byrne, W., King, S., and Tokuda, K. (2011a). An analysis of machine translation and speech synthesis in speech-to-speech translation system. In *Proc. ICASSP 2011*, pages 5108 – 5111, Prague, Czech Republic.

Hashimoto, K., Yamagishi, J., Byrne, W., King, S., and Tokuda, K. (2011b). Impacts of machine translation and speech synthesis on speech-to-speech translation. *Speech Communication*. (under review).

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Amer.*, **87**(4), 1738–1752.

Hirsimäki, T., Pylkkönen, J., and Kurimo, M. (2009). Importance of high-order n-gram models in morph-based speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing,*, **17**(4), 724–732.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. ICASSP-83*, pages 93–96.

Iskra, D., Grosskopf, B., Marasek, K., Van den Heuvel, H., Diehl, F., and Kiessling, A. (2002). SPEECON – speech databases for consumer devices: Database specification and validation. In *Proc. LREC 2002*, pages 329–333, Canary Islands, Spain.

Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., and Itahashi, S. (1998). The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proc. ICSLP 1998*, pages 3261–3264, Sydney, Australia.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999). 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**, 187–207.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. 2nd MAVEBA*, Firenze, Italy.

King, S., Tokuda, K., Zen, H., and Yamagishi, J. (2008). Unsupervised adaptation for HMM-based speech synthesis. In *Proc. Interspeech 2008*, pages 1869–1872, Brisbane, Australia.

Kneser, R. and Ney, H. (1995). Improved backing-off for N-gram language modeling. In *Proc. ICASSP 1995*, pages 181–184, Detroit, Michigan, USA.

Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*. Blackwell.

Liang, H., Dines, J., and Saheer, L. (2010). A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Proc. ICASSP 2010*, pages 4598–4601, Dallas, TX, USA.

Liu, F., Gu, L., Gao, Y., and Picheny, M. (2003). Use of statistical N-gram models in natural language generation for machine translation. In *Proc. ICASSP 2003*, pages 636–639, Hong Kong.

McCree, A. and Barnwell III, T. (1995). A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Trans. on Speech Audio Process.*, **3**(4), 242–250.

Moore, D., Dines, J., Magimai.-Doss, M., Vepa, J., Cheng, O., and Hain, T. (2006). Juicer: A weighted finite-state transducer speech decoder. In *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms MLMI'06*. IDIAP-RR 06-21.

Ney, H. (1999). Speech translation: coupling of recognition and translation. In *Proc. ICASSP-99*, pages 517–520, Phoenix, Arizona.

Noth, E., Batliner, A., Kiessling, A., Kompe, R., and Niemann, H. (2000). Verbmobil: the use of prosody in the linguistic components of a speech understanding system. *Speech and Audio Processing, IEEE Transactions on*, **8**(5), 519 –532.

Paul, D. B. and Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362, Harriman, New York.

Qian, Y., Lang, H., and Soong, F. (2009). A cross-language state sharing and mapping approach to bilingual (Mandarin – English) TTS. *IEEE Trans. Speech, Audio & Language Process.*, **17**(6), 1231–1239.

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

Sjölander, K. and Beskow, J. (2000). Wavesurfer — an open source speech tool. In *Proc. ICSLP 2000*, pages 464–467.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998). Web-based educational tools for speech technology. In *Proc. ICSLP 1998*.

Toda, T. and Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 816–824.

Tokuda, K., Kobayashi, T., Masuko, T., and Imai, S. (1994). Mel-generalized cepstral analysis — a unified approach to speech spectral estimation. In *Proc. ICSLP-94*, pages 1043–1046, Yokohama, Japan.

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

Tsujimura, N. (2006). *An Introduction to Japanese Linguistics (Blackwell Textbooks in Linguistics)*. Blackwell Publishing Limited.

Tsuzaki, M., Tokuda, K., Kawai, H., and Ni, J. (2011). Estimation of perceptual spaces for speaker identities based on the cross-lingual discrimination task. In *INTERSPEECH 2011*, pages 157–160, Florence, Italy.

Wester, M. (2010). Cross-lingual talker discrimination. In *Proc. Interspeech 2010*, Tokyo, Japan.

Wester, M. and Karhila, R. (2011). Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In *Proc. ICASSP 2011*, pages 5372 – 5375, Prague, Czech Republic.

Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y.-J., Saheer, L., King, S., Oura, K., Garner, P. N., Byrne, W., Guan, Y., Hirsimäki, T., Karhila, R., Kurimo, M., Shannon, M., Shiota, S., Tian, J., Tokuda, K., and Yamagishi, J. (2010). Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan.

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: A review. In *Proceedings of the ISCA workshop on adaptation methods for speech recognition*, pages 11–19.

Wu, Y.-J. and Tokuda, K. (2009). State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proc. Interspeech 2009*, pages 528–531, Brighton, U.K.

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

Yamagishi, J., Lincoln, M., King, S., Dines, J., Gibson, M., Tian, J., and Guan, Y. (2009b). 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.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Guan, Y., Oura, K., Tokuda, K., Karhila, R., and Kurimo, M. (2010). Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora. *IEEE Audio, Speech, & Language Processing*, **18**(5), 984–1004.

Yoneyama, K. (2004). A cross-linguistic study of diphthongs in spoken word processsing in Japanese and English. In *Proc. Interspeech 2004*, Jeju Island, Korea.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. EUROSPEECH-99*, pages 2374–2350, Budapest, Hungary.

Young, S. J., Odell, J. J., and Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modeling. In *Proc. ARPA Human Language Technology Workshop*, pages 307–312, Plainsboro, NJ.

Yu, Z., Wu, Y.-J., Zen, H., Nankaku, Y., and Tokuda, K. (2008). Analysis of stream-dependent tying structure for HMM-based speech synthesis. In *Proc. ICSP 2008*, pages 655–658.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Inf. & Syst.*, **E90-D**(1), 325–333.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2007b). A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. & Syst.*, **E90-D**(5), 825–834.