IDIAP RESEARCH REPORT

ENHANCING STATE MAPPING-BASED CROSS-LINGUAL SPEAKER ADAPTATION USING PHONOLOGICAL KNOWLEDGE IN A DATA-DRIVEN MANNER

Hui Liang John Dines

Idiap-RR-08-2013

MARCH 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Enhancing State Mapping-Based Cross-Lingual Speaker Adaptation using Phonological Knowledge in a Data-Driven Manner

Hui Liang, and John Dines, *Member, IEEE*

***Abstract*—HMM state mapping with the Kullback-Leibler divergence as a distribution similarity measure is a simple and effective technique that enables cross-lingual speaker adaptation for speech synthesis. However, since this technique does not take any other potentially useful information into account for mapping construction, an approach involving phonological knowledge in a data-driven manner is proposed in order to produce better state mapping rules – state distributions from the input and output languages are clustered according to broad phonetic categories using a decision tree, and mapping rules are constructed only within each resultant leaf node. Apart from this, previous research shows that a regression class tree that follows the decision tree structure for state tying is detrimental to cross-lingual speaker adaptation. Thus it is also proposed to apply this new approach to regression class tree growth – state distributions from the output language are clustered according to broad phonetic categories using a decision tree, which is then directly used as a regression class tree for transform estimation. Experimental results show that the proposed approach can reduce mel-cepstral distortion consistently and produce state mapping rules and regression class trees that generalize to unseen test speakers. The impacts of the phonological/acoustic similarity between input and output languages upon the reliability of state mapping rules and upon the structure of regression class trees are also demonstrated and analyzed.**

***Index Terms*—data-driven enhancement, phonological constraints, minimum generation error, regression class tree, HMM state mapping, cross-lingual speaker adaptation**

I. Introduction

THE language barrier is a prominent hurdle to overcome in order to facilitate better communication among people across the globe. Real-time automated speech-to-speech (S2S) translation [1]–[3] is a technology which can provide a means to bridge the gap between languages and has the potential of largely reducing the cost of relying upon human interpreters. Therefore it has emerged as an important research topic. The typical architecture of an S2S translator consists of three modules: speech recognition, machine translation and speech synthesis. The output voice of the speech synthesis module usually comes from a professional speaker (e.g. the system presented in [4] and the Google Translate service), who has recorded a large amount of training data, so that high quality of output synthesised speech can be guaranteed. Such a speaker-specific solution is mature, but training data preparation for speaker-specific model training is inherently time-consuming and costly. As a result, the speech synthesis module lacks voice diversity. For the sake of voice diversity, research is being conducted on personalisation of S2S translation, i.e. how to make the output synthetic voice sound like a user's input voice despite the difference in language between the two. This research would enable translated speech to be produced with input voice characteristics of a user.

Owing to its statistical parametric nature, HMM-based speech synthesis [5] is a very flexible framework, in which, for example, voice characteristics, speaking styles and emotion of a speaker can be easily modified by adjusting parameters of HMM synthesis models. More specifically, HMM-based speech synthesis lends itself particularly well to personalized S2S translation since it includes a range of highly effective speaker adaptation algorithms that centre around the *average voice* synthesis paradigm [6], [7]. An average voice is an artificial voice trained on speech data collected from multiple real speakers (for example, by speaker adaptive training [8]), ideally modelling speaker-independent, phonetic and prosodic variations only. Before speech parameter generation, an average voice is adapted to a given target voice by speaker adaptation algorithms like CMLLR [9]. As only tens of adaptation utterances (i.e., a few minutes long in total) are needed from the target speaker for reproducing his/her voice, voice diversity in output synthesized speech can be easily achieved. Thus, the HMM-based speech synthesis framework and the average voice paradigm are the foundation of this work. In the context of personalized S2S translation, the term *cross-lingual speaker adaptation* is generally used to refer to adapting the voice characteristics of average voice synthesis models to those of given adaptation data in a different language from that of desired synthesis output. The respective languages of adaptation data and synthesis models are called *input language* (L_{in}, i.e. the language spoken by the target speaker) and *output language* (L_{out}, i.e. the language in which output speech is synthesized) hereafter.

The lack of correspondence between phonological representations of adaptation data in L_{in} and underlying state distributions of average voice synthesis models in L_{out} presents a difficulty to cross-lingual speaker adaptation, since there is no straightforward way of associating the adaptation data with these synthesis models. The state-of-the-art technique

Research leading to the results in this paper was funded by the 7th Framework Programme (FP7/2007-2013) of the European Union under the grant agreement 213845 (the EMIME project) and the Hasler Foundation (the CLAS3 project).

The authors are affiliated with Idiap Research Institute, CH-1920 Martigny, Switzerland (e-mail: {hliang, dines}@idiap.ch). H. Liang is also affiliated with École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.

that can construct this correspondence is HMM state mapping across languages, which is performed by training two sets of average voice synthesis models in L_{in} and L_{out} respectively and finding the closest matching states between the two model sets. Then adaptation data in L_{in} can be associated with average voice models in L_{out} via resulting state mapping rules. Since the HMM state mapping technique was introduced [10], the minimum Kullback-Leibler divergence (KLD) criterion has been typically employed to construct state mapping rules. This purely data-oriented criterion for finding the closest matching states across languages, though working acceptably well for cross-lingual speaker adaptation [11]–[13], may not always produce meaningful state mapping rules from the point of view of phonological features, especially when L_{in} is substantially phonologically distinct from L_{out}. Less meaningful state mapping rules are presumably more detrimental to the performance of cross-lingual speaker adaptation, thus needing to be identified and corrected.

Moreover, HMM state mapping has no explicit mechanism to eliminate any damaging impact of the mismatch[1] between the input and output languages, though it constructs correspondence between them. It was discovered that state mapping-based cross-lingual speaker adaptation could not benefit from the use of a regression class tree that followed the decision tree structure for state tying in the training stage of synthesis models [14]: given two substantially different languages like Mandarin and English, a global transform was a better choice [15]. Obviously, a global transform is not capable of fully capturing a speaker's voice characteristics and thus a new method of growing a regression class tree needs to be investigated.

In this paper, we firstly propose a refinement of the minimum KLD criterion by introducing phonological knowledge into HMM state mapping construction. The key idea of the proposed approach is to group average voice state distributions of two languages into phonologically consistent clusters, and then to construct mapping rules only *within* each of these clusters as per the minimum KLD criterion. This grouping is achieved by decision tree-based clustering [16]. Phonological constraints (i.e., questions for node splitting) are discovered node by node using a small set of development data in L_{out}, on which resulting state distribution clusters maximally provide reduction of mel-cepstral distortion (MCD). Secondly, instead of constructing a regression class tree by following the decision tree structure for state tying [14], we also apply phonological constraints to the regression class tree: questions for node splitting of a regression class tree (i.e., phonological constraints) are discovered node by node using a small set of development data in L_{out}, on which a resulting regression class tree maximally provides MCD reduction.

This paper is structured as follows: Cross-lingual speaker adaptation (CLSA) is reviewed in brief in Section II. Our proposed approach is elaborated in Section III. Speaker-dependent and speaker-independent experiments are presented in Sections IV and V respectively. Section VI provides conclusions and insights on future research directions.

[1]i.e. the fact that their acoustic spaces, phoneme inventories, prosodic patterns and so forth partially overlap

II. CLSA & HMM State Mapping

A target speaker's voice can be reproduced by training speaker-dependent HMM synthesis models on speech data from the speaker alone [17]. Unfortunately, building robust speaker-dependent models requires a large amount of training data from a target speaker. This requirement makes the speaker-dependent solution expensive, time-consuming and impractical for situations where diversity of synthetic voices is expected. Due to the statistical parametric nature of HMM-based speech synthesis [18], speaker adaptation techniques [9], [19], [20] and the average voice synthesis paradigm [6], [7] have been developed in order to address this problem. An average voice can be regarded as an artificial, more adaptable [21] voice trained on speech data collected from multiple real speakers (and probably involving shared-decision-tree-based context clustering [7, Ch. 4] and speaker adaptive training [8]). By means of speaker adaptation and the average voice synthesis paradigm, the voice characteristics of "source" synthesis models can be adapted to those of a target speaker, given only a small number of utterances in the target speaker's voice.

Unlike intra-lingual speaker adaptation, cross-lingual speaker adaptation adapts the voice characteristics of average voice synthesis models in L_{out} into those of a target speaker who can only provide adaptation data in L_{in}. As $L_{\text{in}} \neq L_{\text{out}}$, the inherent difficulty in cross-lingual speaker adaptation is how to extract speaker characteristics from L_{in} and apply them to L_{out} without having access to any direct relationships between phonological representations of adaptation data in L_{in} and underlying state distributions in L_{out}. Two types of techniques have been investigated so far: i) phoneme mapping [22]–[24], HMM state mapping [10]–[12] and speech feature frame mapping [25]; ii) bilingual modelling [10], [26]–[28] and speaker & language factorization [29]. Their common key point is to establish the missing relationships, either explicitly or implicitly.

This paper is focused on the investigation of HMM state mapping. It is built upon the assumption that languages have significant overlap in acoustic feature space and state mapping provides an appropriate level of granularity to capture this overlap while maintaining some correspondence between acoustic units (e.g., phonemes). HMM state mapping was introduced into cross-lingual speech synthesis by Qian *et al.* [10]. Establishing state mapping rules is carried out in a data-oriented manner, by finding the nearest state emission pdf (say, Y) of models in language L_A for each (say, X) of the state emission pdfs of models in language L_B according to a similarity measure of state emission pdfs (typically the Kullback-Leibler divergence). HMM state mapping works like a function $\mathcal{M}_{L_A \mapsto L_B}(X) = Y$, which captures the relationships between L_{in} and L_{out} at the sub-phonemic level. It is hoped that state mapping rules reflect correspondence between two different languages and are irrelevant to any specific speaker, so average voice synthesis models [6], [7], which are speaker-independent, are employed in the construction of state mapping rules.

Wu *et al.* proposed two manners to utilize HMM state

mapping rules [12]. The *data mapping* manner functions as follows: i) to apply state mapping rules between L_{in} and L_{out} to adaptation data such that the adaptation data in L_{in} is represented as a state sequence in L_{out}; ii) to carry out "intra-lingual" speaker adaptation on the side of L_{out}. As reported in [12], the data mapping manner provides good speaker similarity, but a slight foreign accent can be perceived and the speech quality is degraded.

As for the *transform mapping* manner proposed in [12], conventional intra-lingual speaker adaptation on the side of L_{in} is performed first. Then resultant speaker-specific transforms for L_{in} are associated with state distributions of synthesis models in L_{out} through state mapping rules between L_{out} and L_{in}. So the average voice synthesis model in L_{out} can be adapted with the transforms for L_{in}. As reported in [12], the transform mapping manner provides good speech quality, but speaker similarity is degraded.

The lack of an L_{in} accent in the synthesized speech in L_{out} is sometimes considered detrimental to speaker similarity, for in general L_{in} is the native language of a target speaker and his/her L_{out} accent is more or less affected by L_{in}. This does not matter in the case of cross-lingual speaker adaptation, as speaker similarity should be actually judged by comparing synthesized speech in L_{out} with original speech in L_{in}. Using original speech in L_{out}, whose accent is fairly likely to be affected by L_{in}, as reference speech is only for making it manageable to evaluate speaker similarity during research on cross-lingual speaker adaptation [30].

III. Enhancing State Mapping-Based CLSA using Phonological Knowledge in a Data-Driven Manner

In this paper HMM state mapping is presented from the *data mapping* perspective since previous analysis [13], [15] has shown a preference for data mapping, though the proposed approach may equally generalize to transform mapping as well. We also focus on adaptation of spectrum, which is the dominant component of speaker identity [31]. Since the spectral feature is mel-cepstrum, mel-cepstral distortion (MCD) is employed as the objective measure of spectrum adaptation performance.

A. Optimality of Purely KLD-Based State Mapping Construction

It is natural to question the optimality of the minimum KLD criterion for state mapping construction, since it is purely data-oriented without taking any other potentially useful knowledge into consideration. To test its optimality, a cross-lingual speaker adaptation experiment in the data mapping manner as in [15] was conducted: adapting US English average voice models trained on WSJ-SI84 with 100 Mandarin adaptation utterances in speaker MMh's voice recorded in a soundproof and anechoic chamber (see Sec. IV-A1). A slight difference in this experiment was that this time HMM state mapping rules defined by the k-th best match in L_{out} were used for each state in L_{in}, instead of always selecting the best match satisfying the minimum KLD criterion (i.e., $k \equiv 1$).

TABLE I
Results obtained under the k-th best match criterion for cross-lingual speaker adaptation in the data mapping manner

k	MCD (dB)	k	MCD (dB)
1	7.67	10	7.76
2	7.64	20	7.98
3	7.64	30	8.16
4	7.64	40	8.38
5	7.80	50	8.48

We evaluated for ten values of k in turn and calculated corresponding MCD. Results in Table I show that while MCD does generally increase with increasing k, this is only apparent for k>5. This phenomenon suggests that while KLD is an effective measure of model distribution similarity, there may exist additional latent factors that can be combined with it to achieve more effective state mapping rules.

B. Attempt to Introduce Simple Phonological Knowledge into State Mapping Construction

Having demonstrated that the minimum KLD criterion may not be optimal for constructing HMM state mapping rules, it was hypothesized that the most significant missing factor was the potential lack of phonological consistency in the constructed mapping rules. For example, a state representing vowels could be mapped to a state representing consonants when minimum KLD is the only criterion. Obviously this kind of mapping rule does not make much sense. Hence, such undesirable mapping rules may be avoided by taking advantage of the knowledge of underlying phoneme categories.

Taking the case of k=1 in Table I (i.e., the data mapping baseline), state distributions of the average voice synthesis models in English and Mandarin were categorized according to seven broad phoneme categories (silence, vowel, plosive, fricative, affricate, approximant and nasal) and then state mapping rules were constructed under the minimum KLD criterion *within* each of the seven categories. A state was assigned to a phoneme category, providing that one of the central phone contexts to which the state had been tied belonged to the category. Thus, it was possible for a state to be a member of more than one phoneme category.

The US English average voice models were then adapted using 100 Mandarin adaptation utterances in speaker MMh's voice and the new set of 2975 state mapping rules in total of mel-cepstral features. Then mel-cepstral distortion was calculated and is presented in Table II.

TABLE II
Objective evaluation results of data mapping systems using different methods of state mapping construction

Method of state mapping construction	MCD (dB)
minimum KLD criterion only	7.67
phonological knowledge-guided	7.48

The introduction of phonological knowledge into state mapping construction had 1342 out of the 2975 state mapping rules corrected. Table III shows the details of the 1342 "incorrect" state mapping rules that resulted from the use of only the minimum KLD criterion.

TABLE III
DETAILS OF "INCORRECT" STATE MAPPING RULES THAT RESULTED FROM ONLY THE MINIMUM KLD CRITERION

State category		# of	State category		# of
Man.	Eng.	map.	Man.	Eng.	map.
S	P	534	V	Ap+F	4
V	Ap	160	Af+N+P	Ap	3
N	V	86	Ap	F	3
Af	F	59	V	F+S	3
N	Ap	33	Af	N	2
P	F	31	Af+F	P	2
Af	P	28	Ap	F+P	2
P	S	26	Ap+N	V	2
S	N	23	Ap+S	V	2
S	F+P	21	F+P	S	2
S	Af+F+P	20	S	Af+P	2
F	S	18	V	Ap+N+P	2
Ap	S	17	V	F+P	2
Ap	V	15	Af	Ap	1
S	V	15	Af+F	S	1
F	V	14	Af+N+P	F	1
F	P	13	Af+P	F	1
P	Ap	12	Af+P	S	1
V	F	11	Ap+N	P	1
Ap	P	10	Ap+N	S	1
V	P	10	Ap+P+S	F	1
Af+N	F+P	9	Ap+S	F+P	1
S	F	9	Ap+S	P	1
N	P	8	F	Af+P+V	1
Ap	N	7	F+N	Ap	1
F	Ap	7	F+N	V	1
S	Ap	7	P	S+V	1
V	N	7	P	V	1
V	S	7	P+S	Ap+F	1
Af	F+P	6	S	Af+P+V	1
Af	F+S	6	S	Ap+F	1
P	F+S	6	S	Ap+F+N+P+V	1
Af	S	5	S	Ap+N+P	1
Af+N	P	5	S+V	Ap	1
F+N	S	5	S+V	P	1
N	S	5	V	Af+F+P	1
Af+N	F+S	4	V	Af+F+P+S	1
P	F+S+V	4	V	Af+P	1
P	N	4	V	Ap+F+S	1

S=silence, V=vowel, P=plosive, F=fricative, N=nasal
Af=affricate, Ap=approximant

Table II clearly shows that phonological knowledge can help to improve state mapping rules constructed under the minimum KLD criterion. This finding indicates that phonologically less meaningful mapping rules are harmful in practice and should be eliminated. Therefore, the investigation of further means to exploit phonological knowledge was pursued as detailed in the remainder of this paper.

C. Phonological Knowledge-Guided State Mapping Construction

In Sec. III-B, a naive grouping of average voice state distributions was applied based on phonologically consistent clusters, such that state mapping rules were constructed under the minimum KLD criterion, but *within* each of these clusters. Hence an HMM state in L_{in} could only be mapped to its phonologically consistent counterpart in L_{out} and vice versa. Previous evidence is noted that usually purely knowledge-based approaches are not as effective, for instance, the manual phoneme mapping construction between Mandarin and English presented in [24]. Preferably, a method of introducing phonological knowledge should be developed in a data-driven manner. As a result, decision tree-based state clustering is employed in this work in a similar fashion to that in synthesis model training. Well-trained HMM state distributions of average voice synthesis models in L_{in} and L_{out} are grouped using a decision tree such that each leaf node of the tree is a phonologically consistent cluster. Optimization of this tree is performed such that the MCD of development data in L_{out} is minimized.

1) Question Design: Out of a huge number of phonetic and prosodic contexts used in HMM-based speech synthesis, the most important ones for spectrum modelling are assumed to be the triphone part – left phoneme ("L-"), central phoneme ("C-") and right phoneme ("R-"). Consequently, the triphone contexts are considered an essential factor for grouping average voice state distributions of L_{in} and L_{out}. In particular, we use the seven broad phoneme categories based on articulation manners that are commonly shared across languages: silence, vowel, plosive, fricative, affricate, approximant and nasal. Thus for triphone contexts, there are a total of 21 questions used in the decision tree-based state clustering/grouping.

A state distribution belongs to a particular category if any context-dependent model to which the state is tied belongs to this category. Therefore, a state may be associated with multiple questions. For example, a state distribution is associated with both questions "C_affricate" and "C_plosive" if it is tied to context-dependent phones `*-ch+*`, `*-k+*` and `*-p+*`.

2) Question Selection Criterion: The maximum likelihood criterion has been employed in decision tree-based clustering during synthesis model training for selecting the best question to split a node [16]. Nonetheless, the goal of speech synthesis is to generate speech as close as natural speech, which is only achieved indirectly through the maximum likelihood criterion.

The minimum generation error (MGE) criterion was proposed [32] to more directly target the goal of speech synthesis. "Generation error" refers to the distortion of generated speech parameters from corresponding natural speech parameters, which can be defined as an objective metric (e.g. mel-cepstral distortion). The MGE criterion has been applied to model parameter training [32] as well as decision tree-based state clustering [33], and was found to outperform the maximum likelihood criterion. According to this criterion, the question selected to split a decision tree node should be the one that minimizes a predefined measure of distortion over a particular set of speech data (the training data of synthesis models or a new set of development data) – this idea is used in the proposed approach to grow decision trees for state mapping construction.

Mel-cepstral distortion is chosen to measure generation error and is minimized on development data in L_{out} based on adaptation of synthesis models using data in L_{in}. Therefore a bilingual corpus (in L_{in} and L_{out}) is required in the proposed approach. The bilingual corpus does not need to be large as it is not used for model training like in [10].

Such a bilingual corpus is indispensable when the focus of research is on adapting only speaker characteristics in the context of cross-lingual speaker adaptation. Without a bilingual corpus, the difference in speaker between synthesis

models and adaptation data would be always handled together with that in language by the same adaptation transforms, which are actually supposed to capture only characteristics of a target speaker's voice.

3) Procedure for Enhancing HMM State Mapping Constructions: Bilingual data (in L_{in} and L_{out}) from a certain number of speakers is collected such that adaptation data in L_{in} is used to estimate adaptation transforms and development data in L_{out} is used for optimization according to the MGE criterion. A separate set of test data is retained, which has no intersection with training, adaptation or development data. The overall procedure can be summarized as follows:

1) For each of the N emitting states of an HMM, form one root node by pooling all average voice state distributions of L_{in} and L_{out} that correspond to this emitting state.
2) Find the next non-terminal leaf node X across the N decision trees in the manner of breadth-first search.

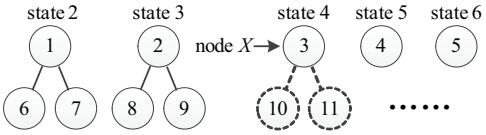


Fig. 1. Illustration of step 1 and step 2 in the case of N=5. The numbers within the nodes indicate the order in which the nodes are split.

3) Find the best split for leaf node X under the MGE criterion. If either of the following conditions is true, X is considered a terminal leaf node. Otherwise X is split using the selected question.
   a) One or both child nodes contain state distributions from only one language;
   b) The best split produces an MCD reduction less than threshold $\varepsilon_{\Delta\text{MCD}}$ ($\varepsilon_{\Delta\text{MCD}} > 0$).
4) Go back to Step 2 or stop when all leaf nodes are terminal leaves.

In order to find the best split for a node X in Step 3 above, average voice state distributions belonging to X are categorized according to every question and the improvement is found by:

1) Recalculating state mapping rules between the input and output languages based on each of the possible node splits;
2) Performing cross-lingual speaker adaptation in the normal data mapping manner using these newly formed mapping rules in X's child nodes;
3) Calculating MCD on held-out development data. The question producing the greatest reduction is selected.

This procedure is visualised in Fig. 2, where node 3 (X=3) in Fig. 1 is taken as an example.

As [32] and [33] report, MGE is a remarkably time-consuming optimization criterion, especially when it is used for decision tree-based clustering. Fortunately, as there are merely 21 questions altogether in the proposed approach, the computational cost is still manageable. Note that the proposed approach degenerates into the conventional state mapping construction if none of the N root nodes are split (i.e., no phonologically consistent clusters are created).

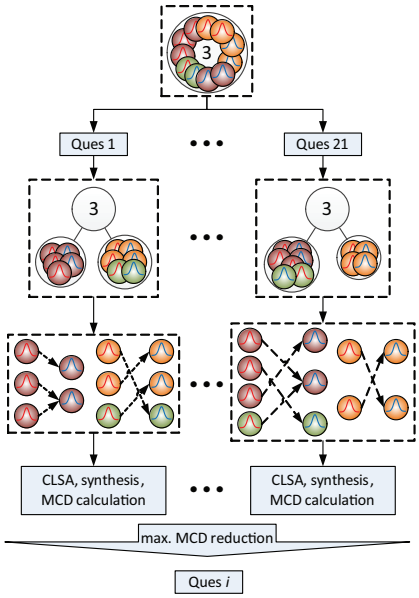


Fig. 2. Procedure of finding the best question to split a node under the MGE criterion for state mapping construction. The blue/red distributions indicate those belonging to $L_{\text{out}}/L_{\text{in}}$ and the background colours indicate that the state distributions belong to different triphone categories.

D. Phonological Knowledge-Guided Regression Class Tree Construction

In previous experiments [15], it was demonstrated that regression class trees derived using the usual approaches based on either state tying [14] or Euclidean clustering [34, Ch. 9] did not lead to effective cross-lingual speaker adaptation. Thus it is proposed to apply the approach elaborated in Sec. III-C to regression class tree growth. The same question set, question selection criterion and principle of growing a tree can be applied. HMM state mapping rules are fixed while a regression class tree is generated by the proposed approach. The overall procedure can be summarized as follows:

1) Form the root node of a regression class tree by pooling all the average voice state distributions of L_{out}.
2) Find the next non-terminal leaf node Y in the regression class tree in the manner of breadth-first search.
3) Find the best split for non-terminal leaf node Y under the MGE criterion:
   a) Split Y according to each of the *valid* questions ("valid" means that a question does not produce a child containing no state distributions);
   b) Perform cross-lingual speaker adaptation with the current regression class tree structure;
   c) Calculate MCD on held-out development data.

   The question producing the greatest MCD reduction exceeding threshold $\varepsilon_{\Delta\text{MCD}}$ ($\varepsilon_{\Delta\text{MCD}} > 0$) is selected for splitting Y. Otherwise Y is considered a terminal leaf node.
4) Go back to Step 2 or stop growing the regression class tree when all leaf nodes are terminal leaves.

This key idea of the above process is visualised in Fig. 3, where Y=3 is taken as an example.

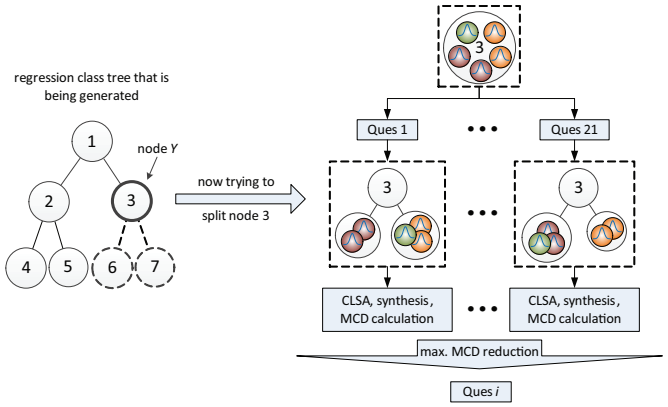


Fig. 3. Key idea of the process of finding the best question to split a node of a regression class tree under the MGE criterion. The blue distributions indicate those belonging to L_{out} and the background colours indicate that the state distributions belong to different triphone categories. The numbers within the nodes indicate the order in which the nodes are split.

Note that the above approach degenerates into cross-lingual speaker adaptation based on a single global transform if no split that reduces MCD on the root node is produced. In such cases, the ability to transfer speaker-specific information between the particular pair of input and output languages via the state mapping technique is limited, as we would expect for two very disparate languages.

Due to the use of the MGE criterion, the proposed approach needs much longer time to generate a regression class tree than conventional ones (e.g. [14] and [34, Ch. 9]). However, this process needs to be carried out only once and a resultant regression class tree applies to any target speaker. Thus the computational cost is still acceptable.

IV. Speaker-Dependent Experiments

A. Experimental Setup

We trained two sets of average voice, single Gaussian-per-state synthesis models on the corpora SPEECON (12.3 hours in Mandarin as L_{in}) and WSJ-SI84 (13.7 hours in US English as L_{out}) respectively using the HTS-2007 system [35] for speaker-dependent[2] experiments. The HMM topology used was five-state and left-to-right with no skip. Speech features were 39th-order STRAIGHT [37] mel-cepstra plus one dimension of energy, $\log F_0$, five-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. All the speaker-dependent cross-lingual adaptation experiments were performed on these two sets of average voice models, using the CSMAPLR [38] algorithm for speaker adaptation and global variances calculated on adaptation data for synthesis.

1) Speakers and Speech Data: Three male (MMh, MM3 and MM6) and two female (MF2 and MF7) speakers were selected from a bilingual corpus recorded in a soundproof, anechoic chamber [39] for speaker-dependent experiments. The five speakers read exactly the same prompts in both Mandarin and English. MF2 is a truly bilingual speaker of Mandarin and English, and the remaining four are native Mandarin speakers. MMh, MF7 and MM3 have reasonably natural English accents[3] but MM6's English is strongly Mandarin-accented. Therefore, only MF2, MMh, MF7 and MM3 were considered training speakers of enhanced state mapping rules.

Adaptation data of each of the five speakers consisted of 100 Mandarin utterances (files 0026∼0125). Development data of each of the four training speakers consisted of 100 English utterances (files 0026∼0125). Test data of each of the five speakers consisted of 25 English utterances (files 0001∼0025).

2) Systems for Comparison: Four groups of experiments were conducted. Within each group, state mapping rules of mel-cepstra between Mandarin and English were derived from *one* of the four training speakers by the proposed approach while those of $\log F_0$, band aperiodicity and duration were still constructed under only the minimum KLD criterion. Then all these mapping rules were used for cross-lingual adaptation of the US English average voice towards each of the four remaining speakers. $\varepsilon_{\Delta\text{MCD}}$ was set to 0.0005dB. The baseline system merely involved the minimum KLD criterion in construction of state mapping rules of all the streams of the state emission pdfs.

Only global transform-based adaptation was investigated in these speaker-dependent experiments. Investigation of regression class-based adaptation is provided in Sec. V.

B. Objective Evaluation

Original recordings of the test data of the five speakers were aligned using the English average voice models and speech samples for objective evaluation were synthesized using the resulting durations. Results of objective evaluation of the four groups of cross-lingual speaker adaptation experiments are presented in Fig. 4 and Table IV. These MCD measurements were calculated on the entire test data set of the five speakers.

TABLE IV
MCD reduction (ΔMCD) in dB produced by the proposed approach, i.e., the difference between the leftmost and rightmost values on each curve in Figure 4

	TrnSpkr			
Data set	MF2	MMh	MM3	MF7
MF2_dev	0.36			
MF2_test	0.39	0.21	0.26	0.23
MMh_dev		0.29		
MMh_test	0.20	0.26	0.16	0.17
MM3_dev			0.21	
MM3_test	0.14	0.14	0.21	0.11
MF7_dev				0.23
MF7_test	0.16	0.16	0.13	0.25
MM6_test	0.05	0.06	0.02	0.09

It can be seen from Fig. 4 that enhanced mapping rules constructed on the development data of a single bilingual speaker consistently provide improvement on his/her own test data. When applying such mapping rules to other target speakers, it is observed that the MCD curves of these target speakers

[2]"Speaker-dependent" in this section means HMM state mapping rules are enhanced on the basis of development data from a single speaker. These speaker-dependent experiments were originally presented in [36].

[3]"Natural" refers to English accents that people speaking English as their first language have and that are not affected by the phonemes and prosody of other languages.

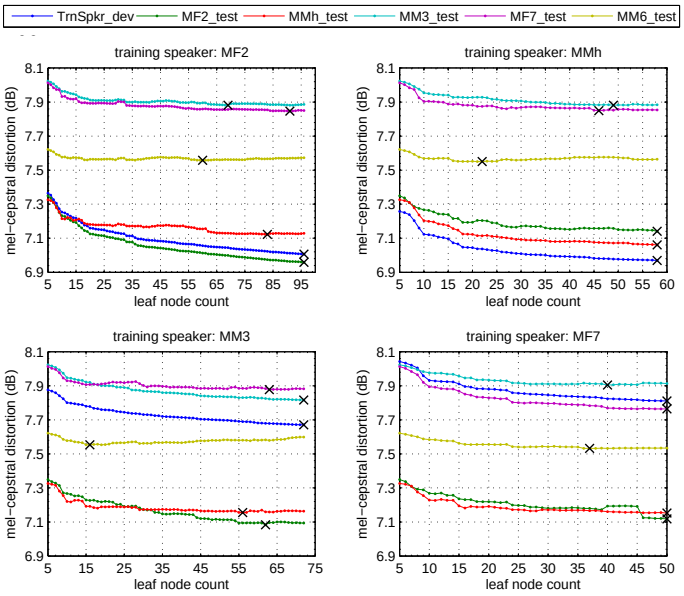


Fig. 4. MCD in relation to the leaf node count during decision tree generation (Crosses indicate minimums on the curves. "TrnSpkr_dev" refers to the development data of respective training speakers. "_test" refers to test data. The six points on the vertical axis in each sub-figure come from the baseline.)

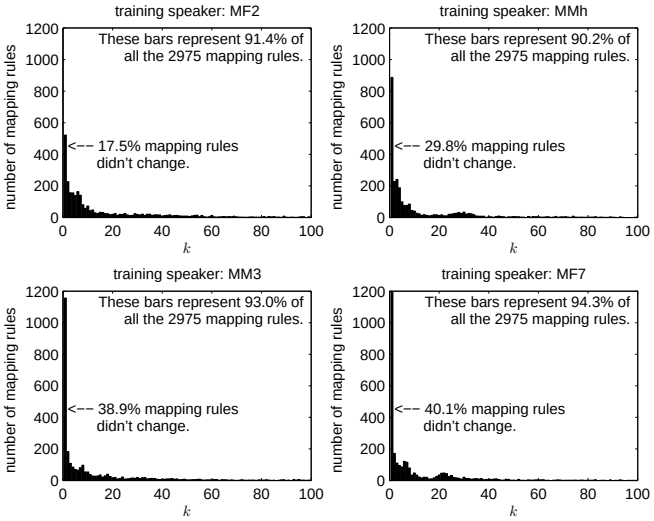


Fig. 5. Histogram of KLD rank (k) using the proposed approach

still have a nearly monotonically decreasing tendency. In other words, mapping rules constructed from a single speaker still maintained a degree of speaker independence. The exception is MM6, who received the least MCD reduction among all the speakers. This result may come from the fact that MM6 has the most pronounced foreign accent when speaking English. State-of-the-art cross-lingual speaker adaptation techniques are not effective at transferring accent information so that the average voice synthesis models in natural US English retain their US accent even after adaptation. The MCD measurements on his English test data thus inherently give lower reductions due to the disagreement in accent between the natural and synthesized utterances.

C. Impact of Phonological Knowledge on Mapping Rules

A total of 2975 mapping rules of mel-cepstra were constructed, one for each of the state distributions in the set of Mandarin average voice models. Fig. 5 shows how k varies under the data-driven use of phonological constraints (see the definition of k in Sec. III-A).

Two common traits are observed across the four histograms in Fig. 5. Firstly, the bars corresponding to k=1 are significantly taller than any others and mapping rules are concentrated in the range of k<20. Thus, the minimum KLD criterion continues to play a dominant role and KLD remains as a good measure of phonological similarity of context-dependent model distributions from two different languages. Secondly, a significant proportion (with a minimum of 59.9%) of state mapping rules were selected with k>1 after phonological constraints were introduced. Therefore, it is also evident that the minimum KLD criterion on its own may not be sufficient, as suggested by the initial analysis in Sec. III-A. It is also interesting to note from both Table IV and Fig. 5 that the proposed approach has the most impact when the truly bilingual speaker MF2 was the training speaker, in terms of the number of changed mapping rules, MCD reduction and providing the best generalization to other speakers (except MM6, as discussed previously).

D. Questions Used for Root Node Splitting

One means to analyze the generalization of the proposed approach is to take into account questions which are close to the root nodes of the decision trees. Table V shows the questions associated with the root node of each decision tree for each of the training speakers.

TABLE V
ROOT NODE QUESTIONS FOR THE EMITTING STATES (2∼6) IN AN HMM

	MF2	MMh	MM3	MF7
2	L-nasal	L-nasal	L-nasal	L-nasal
3	C-nasal	C-nasal	C-vowel	C-nasal
4	C-nasal	C-nasal	C-affricate	C-affricate
5	R-fricative	C-affricate	C-nasal	C-affricate
6	L-silence	L-plosive	L-plosive	L-silence

It is interesting to see that most questions chosen by the proposed method are shared across speakers, thereby confirming that phonological constraints plays a remarkably speaker-independent role in enhancing state mapping rules.

E. Subjective Evaluation

Formal subjective evaluation was performed in the form of AB and ABX listening tests for naturalness and speaker similarity, respectively. All of the speech samples were selected from the experiment group corresponding to the top-left sub-figure in Fig. 4, since MF2 seems to provide the best generalisation to other speakers. Using the baseline and the proposed approach, five sentences from the 25 used in the objective evaluation were synthesized for each of the five speakers. As a result, each listener was presented with 50 pairs

of utterances in total: 5 (pairs of sentences)×5 (speakers)×2 (tests). Note that unadapted duration models of the English average voice were used and that original reference speech in the speaker similarity test was in English. Formal subjective evaluation results are shown in Fig. 6.

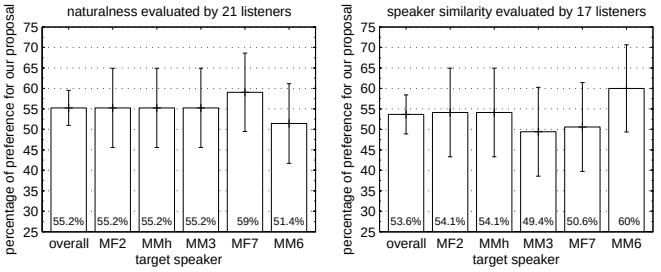


Fig. 6. Subjective evaluation results produced on the basis of MF2-dependent state mapping rules (Whiskers indicate 95% confidence intervals.)

Fig. 6 suggests that the proposed approach produced statistically significant improvement in naturalness with an overall preference score of 55.2%, while speaker similarity was not greatly impacted. To be more specific, it was observed that speech was produced with less "muffled" characteristics by the proposed approach.

V. Speaker-Independent Experiments

The effectiveness and generalization across speakers of the proposed approach to state mapping construction have been demonstrated in Sec. IV. It has been also confirmed that while KLD is a good objective function for determining state mappings, the minimum KLD criterion on its own may produce phonologically inconsistent associations between states, thereby leading to sub-optimal results. In this section we examine enhanced state mapping rules on speech data of multiple bilingual speakers and the use of a regression class tree in the speaker adaptation process.

A. Experimental Setup

We trained three sets of average voice, single Gaussian-per-state synthesis models on the corpora GlobalPhone (13.4 hours in Mandarin as L_{in}), PHONDAT1 (9.6 hours in German as L_{in}) and WSJCAM0 (18.9 hours in UK English as L_{out}) respectively using the HTS-2010 system [40] for speaker-independent[4] experiments. The use of WSJCAM0 in addition to WSJ-SI84 was for testing the proposed approach on more corpora. Mandarin from the Sino-Tibetan language family and German from the West Germanic language family were chosen as input languages because they are "far from" and "close to" English respectively, which is also a West Germanic language. This should give us some insights into the extent to which the dissimilarity of L_{in} and L_{out} can affect the performance of cross-lingual speaker adaptation. Table VI indicates the similarity of Mandarin/German to English from another angle.

[4]"Speaker-independent" in this section means HMM state mapping rules and regression class trees are enhanced on the basis of development data from multiple speakers.

TABLE VI
Statistics of KLD over the entire state mapping sets of the two language pairs when only the minimum KLD criterion was applied to mapping construction

Language pair	KLD mean	KLD median
German & English	23.7	18.2
Mandarin & English	75.8	19.0

The HMM topology used was five-state and left-to-right with no skip. Speech features were 39th-order STRAIGHT [37] mel-cepstra plus one dimension of energy, $\log F_0$, 21-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. All of the speaker-independent cross-lingual speaker adaptation experiments were performed using the CSMAPLR [38] algorithm, transforms being estimated from one iteration. Global variances for synthesis were calculated on adaptation data.

Ten Mandarin-English speakers (Chinese) [39] and ten German-English (Germans) [41] speakers were used in speaker-independent experiments. They all have reasonably natural English accents (see the accent rating in [39], [41]) and were grouped as shown in Table VII. The groupings were used for cross validation since the number of available bilingual training speakers was limited.

TABLE VII
Grouping of training speakers in speaker-independent experiments (For each language pair, each time four speaker groups were used as the training partition and the two leftover speakers were test speakers.)

Group ID	1	2	3	4	5
male Germans	GM1	GM2	GM3	GM6	GM7
female Germans	GF1	GF2	GF4	GF6	GF7
Group ID	6	7	8	9	0
male Chinese	MMh	MM3	MM4	MM5	MM7
female Chinese	MF1	MF2	MF4	MF5	MF7

Adaptation data of each of the 20 speakers consisted of 100 Mandarin or German utterances (files 0026∼0125). Development data consisted of 100 English utterances (files 0026∼0125) and test data consisted of 25 English utterances (files 0001∼0025).

B. Systems for Analysis of the Proposed Approach

Experiments were conducted in the form of 5-fold cross validation with gender balance maintained. There were always four male and four female speakers (i.e., four speaker groups in Table VII) in the training partition and one male and one female speakers (i.e., the leftover speaker group) in the test partition.

In each experiment, enhanced state mapping rules of mel-cepstra between English and German/Mandarin were derived from the training partition by the proposed approach, while those of $\log F_0$, band aperiodicity and duration were still constructed under the minimum KLD criterion. These mapping rules were used for cross-lingual adaptation of the UK English average voice towards each of the test speakers.

Likewise, the proposed approach to growing a regression class tree for mel-cepstra was applied to the training partition

of each experiment. Global transforms were employed for $\log F_0$, band aperiodicity and duration. The resulting regression class tree and global transforms were used for cross-lingual adaptation of the UK English average voice towards each of the test speakers.

Four settings (pro_glo, pro_dec, kld_pro and pro_pro as described in Table VIII) were evaluated in the speaker-independent experiments. $\varepsilon_{\Delta\text{MCD}}$ was set to 0.0005dB.

TABLE VIII
SETTINGS OF SPEAKER-INDEPENDENT EXPERIMENTS

	State mapping construction	Regression class tree growth
kld_glo	minimum KLD criterion	global transform
pro_glo	proposed approach	
kld_dec	minimum KLD criterion	decision tree structure
pro_dec	proposed approach	
kld_glo	minimum KLD criterion	global transform
kld_pro		proposed approach
pro_glo	proposed approach	global transform
pro_pro		proposed approach

C. Objective Evaluation

Original recordings of development and test data of the 20 speakers were aligned using the UK English average voice models and speech samples for objective evaluation were synthesized using resulting durations. Results of objective evaluation on the development data set are presented in Tables IX and X.

TABLE IX
MCD (DB) ON THE DEVELOPMENT DATA OF THE TRAINING PARTITION & THE PERCENTAGE OF MAPPING RULES THAT REMAINED UNCHANGED

Lang.	L_{in} = German, L_{out} = British English					
Groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	Avg.
kld_glo	6.04	6.13	6.08	6.07	6.08	6.08
pro_glo	5.93	6.04	5.98	6.00	5.99	5.99
diff.	**0.11**	**0.09**	**0.10**	**0.07**	**0.09**	**0.09**
	50.2%	56.8%	45.5%	49.3%	52.1%	50.8%
kld_dec	5.93	6.04	6.00	5.99	6.00	5.99
pro_dec	5.82	5.94	5.88	5.91	5.92	5.89
diff.	**0.11**	**0.09**	**0.12**	**0.09**	**0.08**	**0.10**
	54.4%	47.6%	45.5%	54.2%	60.0%	52.3%
Lang.	L_{in} = Mandarin, L_{out} = British English					
Groups	6-7-8-9	6-7-8-0	6-7-9-0	6-8-9-0	7-8-9-0	Avg.
kld_glo	7.07	7.09	7.04	7.06	7.08	7.07
pro_glo	6.96	6.97	6.91	6.93	6.97	6.95
diff.	**0.11**	**0.12**	**0.13**	**0.13**	**0.10**	**0.12**
	39.4%	25.6%	29.3%	35.7%	22.8%	30.6%
kld_dec	7.19	7.22	7.17	7.19	7.23	7.20
pro_dec	7.06	7.08	6.99	7.02	7.10	7.05
diff.	**0.13**	**0.14**	**0.18**	**0.17**	**0.13**	**0.15**
	41.7%	46.1%	41.7%	47.5%	42.4%	43.9%

Table IX shows that in comparison with mapping rules between Mandarin and English, a significantly larger proportion of state mapping rules between German and English remained unchanged after the proposed approach was applied, which suggests that the state mapping rules between German and English constructed under the minimum KLD criterion were more reliable than those between Mandarin and English. This is also reflected in the fact that MCD reduction concerning Mandarin and English was greater than that concerning German and English. These phenomena demonstrate that the phonological similarity of the input and output languages impacts on the effectiveness of the minimum KLD criterion in creating links between the two languages.

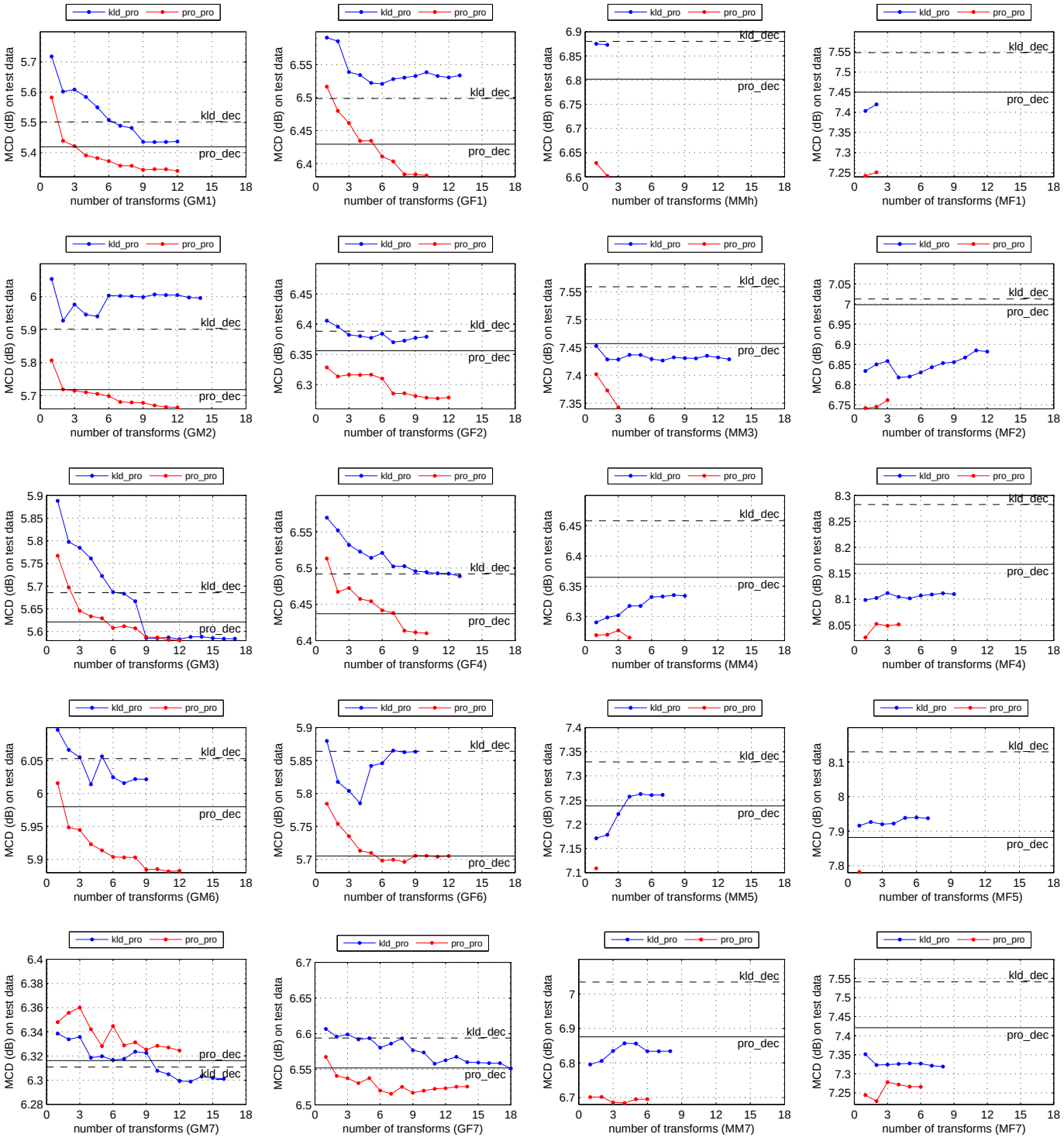
TABLE X
MCD (DB) ON THE DEVELOPMENT DATA OF THE TRAINING PARTITION & THE NUMBER OF REGRESSION CLASS TREE LEAVES

Lang.	L_{in} = German, L_{out} = British English					
Groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	Avg.
kld_glo	6.04	6.13	6.08	6.07	6.08	6.08
kld_pro	5.87	6.00	5.94	5.93	5.95	5.94
diff.	**0.17**	**0.13**	**0.14**	**0.14**	**0.14**	**0.14**
	19	9	18	14	14	14.8
pro_glo	5.93	6.04	5.98	6.00	5.99	5.99
pro_pro	5.79	5.92	5.86	5.86	5.87	5.86
diff.	**0.15**	**0.12**	**0.13**	**0.13**	**0.12**	**0.13**
	14	12	12	12	12	12.4
Lang.	L_{in} = Mandarin, L_{out} = British English					
Groups	6-7-8-9	6-7-8-0	6-7-9-0	6-8-9-0	7-8-9-0	Avg.
kld_glo	7.07	7.09	7.04	7.06	7.08	7.07
kld_pro	7.05	7.07	7.01	7.03	7.07	7.05
diff.	**0.02**	**0.02**	**0.03**	**0.03**	**0.01**	**0.02**
	8	7	9	13	2	7.8
pro_glo	6.96	6.97	6.91	6.93	6.97	6.95
pro_pro	6.95	6.97	6.91	6.91	6.97	6.94
diff.	**0.01**	**0.00**	**0.01**	**0.02**	**0.01**	**0.01**
	6	1	4	3	2	3.2

Table X shows that the proposed approach could reduce MCD by enhancing the regression class tree structure, especially for the language pair of German and English. When the language pair was Mandarin and English, the proposed approach could only produce negligible MCD reductions and very small regression class trees. These results suggest that the proposed approach also can be used to control the appropriate number of transforms, depending on the phonological similarity of two languages. They also strengthen the finding in [15] that a global transform is sufficient when the input and output languages are substantially phonologically distinct: In this circumstance, it would be enough to apply the proposed approach to state mapping construction only and to use a global transform in adaptation.

In Fig. 7, objective results on the test data of the two test speakers of each fold of the cross-validation experiments are presented for a comparative analysis.

The two columns on the right side in Fig. 7 confirm that the best solution in the case of Mandarin and English was achieved by only applying the proposed approach to state mapping construction and using a global transform in adaptation. This is understandable. Firstly, one purpose of using a regression class tree in speaker adaptation is to capture speaker information in adaptation data at an increasingly finer grained level by dividing and clustering model distributions according to their proximity in the model space into different regression classes and then estimating respective transforms for these classes. Secondly, adaptation algorithms like CMLLR blindly handle all kinds of mismatch (in terms of speaker, language, recording environment, etc) between synthesis models and adaptation data with a single set of transforms. Thus as the number of adaptation transforms increase, more Mandarin-specific information that had no relation to speaker identity is inadvertently captured from adaptation data. Given the substantial difference



(a) male German/English speakers (b) female German/English speakers (c) male Mandarin/English speakers (d) female Mandarin/English speakers

Fig. 7. MCD measurements in relation to the number of transforms in various conditions (The four columns correspond to male Germans, female Germans, male Chinese and female Chinese, respectively. The leftmost point on each red curve indicates the result of pro_glo and the leftmost point on each blue curve indicates the result of kld_glo.)

between Mandarin and English, it is not surprising that the quality of synthesized English is degraded immediately after the number of adaptation transforms grows.

As for German and English, the two columns on the left side in Fig. 7 show that the proposed approach can be applied to state mapping construction first and then to regression class tree growth, producing a further MCD reduction in most cases. The regression class trees in the case of German and English were larger and produced greater MCD reductions, compared with those in the case of Mandarin and English. This demonstrates that owing to the phonological and acoustic similarity of German to English, adaptation algorithms are better able to utilize greater quantities of adaptation data given an appropriate regression class tree. These two columns also show: (1) the MCD scores produced by applying the proposed approach to both state mapping construction and regression class tree growth (pro_pro, the red curves) are more likely to decrease further than those produced by applying the proposed approach to regression class tree growth only (kld_pro, the blue curves); (2) when using enhanced state mapping rules, enhanced regression class trees generated by the proposed approach (pro_pro, the red curves) eventually produced MCD scores smaller than those the regression class tree following the decision tree structure of the UK English average voice models produced (pro_dec, the solid black horizontal lines), except for the speaker GM7. Thus it is concluded that the best and most robust approach for German and English should be the combination of state mapping enhancement and regression class tree enhancement by the proposed approach.

D. Iterative Enhancement

The proposed approach can be applied to state mapping enhancement and regression class tree enhancement iteratively in an alternating fashion. Namely, using the regression class tree obtained in the i-th iteration, state mapping rules can be enhanced again and then this regression class tree from the i-th iteration can continue to grow in the (i+1)-th iteration.

There are two methods of enhancing state mappings in the (i+1)-th iteration based on the regression class tree from the i-th iteration:

1) Construct state mapping rules from scratch. This method is denoted by "M-0" hereafter.
2) Construct state mapping rules by extending the decision tree that has produced enhanced mapping rules in the i-th iteration. This method is denoted by "M-ext" hereafter.

In the case of Mandarin-to-English adaptation, this is unlikely to have any impact due to the small size of the regression class trees obtained in the first iteration. However, results of the German-to-English adaptation suggest some potential. Hence both M-0 and M-ext were tested in the second iteration for the language pair of German and English. MCD measurements after the second iteration of state mapping enhancement are listed in Table XI.

Then enhanced state mapping rules obtained in the second iteration were used to continue to grow regression class trees obtained in the first iteration. MCD measurements after the second iteration of regression class tree growth are listed in Table XII.

TABLE XI
MCD (DB) ON THE DEVELOPMENT DATA OF THE TRAINING PARTITION & THE PERCENTAGE OF MAPPING RULES THAT REMAINED UNCHANGED AFTER STATE MAPPING ENHANCEMENT IN THE SECOND ITERATION

Lang. Groups	L_{in} = German, L_{out} = UK English					
	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	Avg.
baseline†	5.79	5.92	5.86	5.86	5.87	5.86
M-0	5.77	5.91	5.85	5.85	5.85	5.85
	64.7%	73.6%	65.8%	64.2%	56.6%	65.0%
M-ext	5.77	5.89	5.85	5.85	5.84	5.84
	91.2%	86.9%	91.7%	84.7%	79.4%	86.8%

† The baseline results are the outcome of pro_pro (i.e., from the first iteration).

TABLE XII
MCD (DB) ON THE DEVELOPMENT DATA OF THE TRAINING PARTITION & THE NUMBER OF REGRESSION CLASS TREE LEAVES AFTER REGRESSION CLASS TREE GROWTH IN THE SECOND ITERATION

Lang. Groups	L_{in} = German, L_{out} = UK English					
	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	Avg.
baseline†	14	12	12	12	12	12.4
using M-0	5.77	5.91	5.85	5.85	5.85	5.85
	16	13	12	14	14	13.8
using M-ext	5.77	5.89	5.85	5.85	5.84	5.84
	16	14	12	12	13	13.4

† The baseline results are the outcome of pro_pro (i.e., from the first iteration).

It is observed that the further improvements given by state mapping enhancement and regression class tree enhancement in the second iteration are negligible, no matter whether M-0 or M-ext was employed. Consequently, it can be confirmed that a single iteration of state mapping construction and regression class tree growth by the proposed approach is sufficient for German and English.

E. Subjective Evaluation

Naturalness and speaker similarity of speech which was synthesized by the proposed approach being applied to both state mapping construction and regression class tree growth (i.e., system pro_pro) were assessed in the form of AB and ABX tests respectively. The three systems to be compared against were a conventional intra-lingual speaker adaptation system, kld_glo (i.e. the starting point which the proposed approach was applied to) and kld_dec (i.e. the conventional, data mapping-based CLSA system as in [12]). Each listener was presented with 60 utterance pairs in total: 3 (pairs)$\times$10 (test speaker groups)$\times$2 (tests). The sentence of each pair was randomly selected from the 25 test sentences. All the natural and synthesized stimuli were in English and duration models of the UK English average voice were used in the synthesis of all these stimuli. Subjective evaluation results can be found in Fig. 8.

Firstly, it is noted that the proposed approach mainly improved naturalness of synthesized speech in the speaker-independent experiments, as observed in the previous speaker-dependent experiments in Sec. IV. According to the speaker discrimination experiments in [30], we hypothesize that a

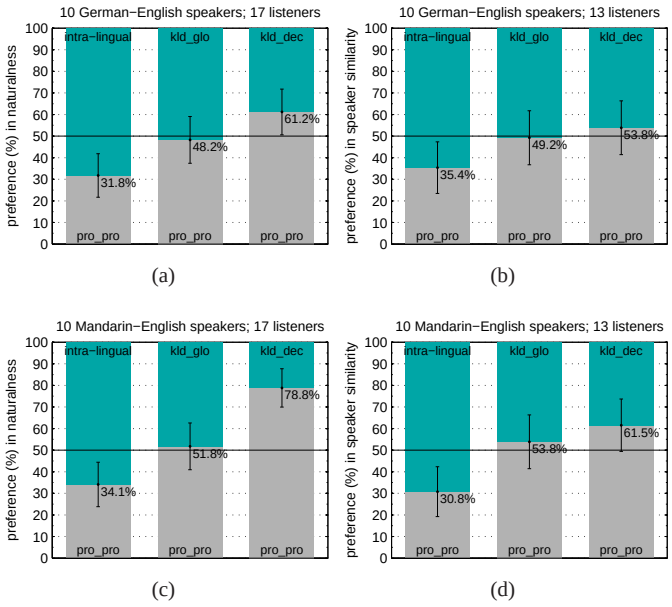


Fig. 8. Results of subjective evaluations on the proposed approach (Whiskers indicate 95% confidence intervals.)

limiting factor in these experiments is the quality of speech generated by cross-lingual speaker adaptation, which hinders listeners' judgment of speaker identity.

Secondly, it is observed that applying the proposed approach to both state mapping construction and regression class tree growth (i.e. pro_pro) produced a significantly better system than the conventional CLSA system (i.e. kld_dec). The proposed approach can automatically generate a suitable regression class tree structure for cross-lingual speaker adaptation so that L_{in}-specific information from adaptation data can be suppressed as much as possible. The contrast between Figures 8(a) and 8(c) appears to suggest that the proposed approach is more effective for a pair of languages which are more phonologically dissimilar.

Lastly, Fig. 8 shows that intra-lingual speaker adaptation still outperformed cross-lingual speaker adaptation, which suggests that the language mismatch problem has not yet been resolved although the proposed approach alleviated some of the negative effects.

VI. Conclusion

An approach that enhanced state mapping-based cross-lingual speaker adaptation using phonological knowledge in a data-driven manner was proposed in this paper. It was applied to HMM state mapping construction such that phonologically inconsistent state mapping rules can be avoided. It was also applied to regression class tree growth such that the appropriate size of a regression class tree and phonologically consistent transform grouping can be achieved automatically.

The proposed approach was firstly applied in a speaker-dependent setting. It was found that enhanced mapping rules constructed by the proposed approach still maintained a degree of speaker independence, even when trained on speech data of a single speaker. While KLD remains a good measure of phonological similarity of context-dependent models from two different languages, the minimum KLD criterion on its own may not be sufficient. It is also apparent that training speakers' proficiency in their non-native languages is important. A high level of proficiency can potentially produce better state mapping rules, in other words, a greater MCD reduction.

The effectiveness and generality of the proposed approach was then demonstrated on two language pairs (German & English, Mandarin & English) in a speaker-independent setting. It was further found that the less phonologically similar the input and output languages were, the less effective the minimum KLD criterion was for creating links between the two languages. The phonological/acoustic similarity of the input language to the output language also has a significant impact on the size of a regression class tree that can be grown by the proposed approach. It continues to be observed that a large regression class tree is of much less use in the current state mapping-based cross-lingual speaker adaptation framework.

The iterative enhancement under the MGE criterion shows rapid convergence. This appears to suggest that there is limited room to improve the simple HMM state mapping technique with the K-L divergence as a measure of state distribution similarity. An explicit step to separate language information from speaker characteristics in adaptation transforms is necessary (e.g. [29]).

References

[1] L. Levin, A. Lavie, M. Woszczyna, D. Gates, M. Gavalda, D. Koll, and A. Waibel, "the Janus-III Translation System: Speech-to-speech translation in multiple domains", *Machine Translation*, vol. 15, pp. 3–25, 2000.

[2] B. Zhou, Y. Gao, J. Sorensen, D. Déchelotte, and M. Picheny, "A hand-held speech-to-speech translation system", in *Proc. of ASRU*, Nov. 2003, pp. 664–669.

[3] M. Kurimo, W. Byrne, J. Dines, P. N. Garner, M. Gibson, Y. Guan, T. Hirsimäki, R. Karhila, S. King, H. Liang, K. Oura, L. Saheer, M. Shannon, S. Shiota, J. Tian, K. Tokuda, M. Wester, Y.-J. Wu, and J. Yamagishi, "Personalising speech-to-speech translation in the EMIME project", in *Proc. of the ACL 2010 System Demonstrations*, Jul. 2010, pp. 48–53.

[4] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs", in *Proc. of NAACL-HLT*, Jun. 2012, pp. 437–445.

[5] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis", *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[6] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", *IEICE Transactions on Information and Systems*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[7] J. Yamagishi, "Average-voice-based speech synthesis", Ph.D. dissertation, Tokyo Institute of Technology, Mar. 2006.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training", in *Proc. of ICSLP*, Oct. 1996, pp. 1137–1140.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[10] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1231–1239, Aug. 2009.

[11] Y.-N. Chen, Y. Jiao, Y. Qian, and F. K. Soong, "State mapping for cross-language speaker adaptation in TTS", in *Proc. of ICASSP*, Apr. 2009, pp. 4273–4276.

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

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

[14] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, “Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis”, in *Proc. of ICASSP*, May 2004, pp. 5–8.

[15] H. Liang and J. Dines, “An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation”, in *Proc. of Interspeech*, Sep. 2010, pp. 622–625.

[16] S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modelling”, in *Proc. of the Workshop on Human Language Technology*, 1994, pp. 307–312.

[17] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English”, in *Proc. of IEEE Workshop on Speech Synthesis*, Sep. 2002, pp. 227–230.

[18] A. W. Black, H. Zen, and K. Tokuda, “Statistical parameteric speech synthesis”, in *Proc. of ICASSP*, Apr. 2007, pp. 1229–1232.

[19] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Speaker adaptation for HMM-based speech synthesis system using MLLR”, in *Proc. of ESCA/COCOSDA Workshop on Speech Synthesis*, Nov. 1998, pp. 273–276.

[20] ——, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR”, in *Proc. of ICASSP*, May 2001, pp. 805–808.

[21] J. Yamagishi, O. Watts, S. King, and B. Usabaev, “Roles of the average voice in speaker-adaptive HMM-based speech synthesis”, in *Proc. of Interspeech*, Sep. 2010, pp. 418–421.

[22] M. Moberg, K. Pärssinen, and J. Iso-Sipilä, “Cross-lingual phoneme mapping for multilingual synthesis systems”, in *Proc. of Interspeech*, Oct. 2004, pp. 1029–1032.

[23] J. Latorre, K. Iwano, and S. Furui, “New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer”, *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, Oct. 2006.

[24] Y.-J. Wu, S. King, and K. Tokuda, “Cross-lingual speaker adaptation for HMM-based speech synthesis”, in *Proc. of ISCSLP*, Dec. 2008, pp. 1–4.

[25] Y. Qian, J. Xu, and F. K. Soong, “A frame mapping based HMM approach to cross-lingual voice transformation”, in *Proc. of ICASSP*, May 2011, pp. 5120–5123.

[26] J. Köhler, “Multilingual phone models for vocabulary-independent speech recognition tasks”, *Speech Communication*, vol. 35, pp. 21–30, 2001.

[27] T. Schultz and A. Waibel, “Language-independent and language-adaptive acoustic modeling for speech recognition”, *Speech Communication*, vol. 35, pp. 31–51, 2001.

[28] J. Latorre, K. Iwano, and S. Furui, “Polyglot synthesis using a mixture of monolingual corpora”, in *Proc. of ICASSP*, Mar. 2005, pp. 1–4.

[29] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulović, and J. Latorre, “Statistical parametric speech synthesis based on speaker and language factorization”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 6, pp. 1713–1724, Aug. 2012.

[30] M. Wester and H. Liang, “Cross-lingual speaker discrimination using natural and synthetic speech”, in *Proc. of Interspeech*, Aug. 2011, pp. 2481–2484.

[31] O. Türk and L. M. Arslan, “Subjective evaluations for perception of speaker identity through acoustic feature transplantations”, in *Proc. of Eurospeech*, Sep. 2003, pp. 2093–2096.

[32] Y.-J. Wu and R.-H. Wang, “Minimum generation error training for HMM-based speech synthesis”, in *Proc. of ICASSP*, May 2006, pp. 89–92.

[33] Y.-J. Wu, W. Guo, and R.-H. Wang, “Minimum generation error criterion for tree-based clustering of context-dependent HMMs”, in *Proc. of Interspeech*, Sep. 2006, pp. 2046–2049.

[34] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *the HTK Book*, Mar. 2009.

[35] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “Robust speaker-adaptive HMM-based text-to-speech synthesis”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[36] H. Liang and J. Dines, “Phonological knowledge guided HMM state mapping for cross-lingual speaker adaptation”, in *Proc. of Interspeech*, Aug. 2011, pp. 1825–1828.

[37] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “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*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[38] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[39] M. Wester and H. Liang, “The EMIME Mandarin bilingual database”, University of Edinburgh, Tech. Rep. EDI-INF-RR-1396, Feb. 2011.

[40] J. Yamagishi and O. Watts, “The CSTR/EMIME HTS system for Blizzard Challenge 2010”, in *Proc. of the Blizzard Challenge*, Sep. 2010.

[41] M. Wester, “The EMIME bilingual database”, University of Edinburgh, Tech. Rep. EDI-INF-RR-1388, 2010.