IDIAP RESEARCH REPORT

DATA-DRIVEN ENHANCEMENT OF STATE MAPPING-BASED CROSS-LINGUAL SPEAKER ADAPTATION

Hui Liang

Idiap-RR-38-2012

DECEMBER 2012

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

DATA-DRIVEN ENHANCEMENT OF STATE MAPPING-BASED CROSS-LINGUAL SPEAKER ADAPTATION

對基於狀態映射的跨語言說話人自適應之數據驅動式改進

Thèse n. 5576 (2012)
présentée le 30 octobre 2012
à la Faculté Sciences et Techniques de l'Ingénieur
laboratoire de l'IDIAP
programme doctoral en Génie Électrique
École Polytechnique Fédérale de Lausanne (EPFL)

pour l'obtention du grade de Docteur ès Sciences
par

梁晖 **(Hui LIANG)**

acceptée sur proposition du jury :

Prof. Jean-Philippe Thiran, président du jury
Prof. Hervé Bourlard, directeur de thèse
Dr. John Dines, co-directeur de thèse
Prof. 徳田恵一 (Keiichi Tokuda), rapporteur
Prof. Tanja Schultz, rapporteur
Dr. Jean-Marc Vesin, rapporteur

洛桑聯邦理工學院 (EPFL), Lausanne, 2012

Version Information

This is the final version submitted to EPFL on 19th November 2012.

http://publications.idiap.ch/downloads/papers/2012/Liang_THESIS_2012.pdf

Compiled with LaTeX at 15:27 (UTC+1), 18th November, 2012

故天將降大任於是人也
必先苦其心志
勞其筋骨
餓其體膚
空乏其身
行拂亂其所為
所以動心忍性
曾益其所不能
人恆過　然後能改
困於心　衡於慮　而後作
徵於色　發於聲　而後喻
入則無法家拂士
出則無敵國外患者
國恆亡
然後知生於憂患而死於安樂也

孟子

致所有幫助過我、在乎我或愛我的人
To whoever helped, cares about or loves me …

梁/LIANG, 晖/Hui (in my Mainland Chinese passport)
梁文暉 (in my genealogy)

Translation of the short article on the previous page

When God [1] is going to place great responsibility upon a man, He always frustrates his spirit and will, exhausts his muscles and bones, exposes him to starvation and poverty, harasses him with troubles and setbacks beforehand, in order to stimulate his mind, to toughen his nature and to enhance his abilities.

A man does not reform if nothing happens to him. A man is motivated to reform only when a certain situation that makes him perplexed and thus start to think happens. Usually a man does not realize his problems until he is told implicitly or explicitly.

A country will definitely collapse if it has neither internal upstanding ministers and counsellors nor external enemies and troublemakers.

Adversity makes men, and prosperity makes monsters.

— Mencius, 372 BC – 289 BC

(Translated by the author of the thesis)

1. the ruler of the universe in the traditional Chinese culture

Résumé

Cette thèse est motivée par l'objectif de développer des systèmes de traduction parole à parole personnalisés et est axée autour d'un de ces composants fondamentaux – l'adaptation interlinguale de locuteur dans le cadre de la synthèse vocale à partir d'une entrée textuelle. Un système de traduction parole à parole personnalisé permet de traduire un signal d'entrée parlé pour une personne donnée en un signal de sortie parlé tout en maintenant l'identité liée à la voix.

Avant de traiter les questions techniques, cette thèse aborde la perception par l'homme de l'identité du locuteur. Des tests d'écoute ont été menés afin de déterminer si les gens sont en mesure de différencier des individus lorsqu'ils s'expriment dans plusieurs langues. Les résultats de ces expériences montrèrent que cette tâche est réalisable. Toutefois, il était difficile pour les auditeurs testés de différencier les locuteurs lorsque, à la fois, la langue et le type de parole varient (enregistrements initiaux ou échantillons synthétisés).

Le problème fondamental dans l'adaptation interlinguale de locuteur est de déterminer comment appliquer les techniques d'adaptation de locuteur lorsque les données d'adaptation sont dans une langue différente de celle employée pour générer les modèles de synthèse. Une grande partie de cette thèse est consacrée à l'analyse et à l'amélioration de l'adaptation interlinguale de locuteur reposant sur les correspondances d'états MMC. Tout d'abord, les conséquences d'une adaptation interlinguale non supervisée sont examinées, compte tenu du lien direct avec le scénario d'application d'une traduction parole à parole personnalisée. La comparaison des systèmes supervisés et non-supervisés montre que la performance de l'adaptation interlinguale non-supervisée est comparable à la méthode supervisée, même si le taux d'erreur phonème des systèmes non-supervisés est d'environ 75%.

Ensuite, les conséquences de la disparité de langue entre les modèles de synthèse et les données d'adaptation sont examinées. Il a été constaté que cette disparité transmet des informations indésirables de la langue, des données d'adaptation vers les modèles de synthèse, limitant l'efficacité des transformations des classes de régression, de l'utilisation d'une quantité plus importante de données d'adaptation, ainsi que de l'estimation itérative des transformations d'adaptation.

Troisièmement, en vue de résoudre les problèmes causés par la disparité de langue, un

cadre d'adaptation axé sur les données et utilisant des connaissances phonologiques est proposé. L'idée fondamentale est de grouper les états MMC en fonction des connaissances phonologiques et en se basant sur les données, pour ensuite associer chaque état avec un homologue phonologiquement cohérent dans une langue différente. Ce cadre est également utilisé lors de la construction d'un arbre de régression pour l'estimation des transformations. Il ressort que le cadre proposé atténue l'impact négatif de la disparité de langue, et conduit à une solide amélioration par rapport aux précédentes méthodes de l'état de l'art.

Enfin, un cadre de transformation hiérarchique à deux couches est proposé, où une couche vise à capturer les caractéristiques de la voix d'un locuteur cible, et l'autre couche compense la disparité de langue. Une étude initiale a été menée afin de déterminer une méthode permettant de construire cette structure hiérarchique de transformations. Bien que les résultats préliminaires soient prometteurs, des investigations plus approfondies restent nécessaire pour confirmer la validité de cette approche.

Mots-clés disparité de langue, correspondance d'états MMC, amélioration axée sur les données, hiérarchie d'adaptation à deux couches, adaptation interlinguale de locuteur, traduction de parole à parole, synthèse vocale en utilisant des MMCs

(Translated by Laurent El Shafey as per the English version)

Abstract

The thesis work was motivated by the goal of developing personalized speech-to-speech translation and focused on one of its key component techniques – cross-lingual speaker adaptation for text-to-speech synthesis. A personalized speech-to-speech translator enables a person's spoken input to be translated into spoken output in another language while maintaining his/her voice identity.

Before addressing any technical issues, work in this thesis set out to understand human perception of speaker identity. Listening tests were conducted in order to determine whether people could differentiate between speakers when they spoke different languages. The results demonstrated that differentiating between speakers across languages was an achievable task. However, it was difficult for listeners to differentiate between speakers across both languages and speech types (original recordings versus synthesized samples).

The underlying challenge in cross-lingual speaker adaptation is how to apply speaker adaptation techniques when the language of adaptation data is different from that of synthesis models. The main body of the thesis work was devoted to the analysis and improvement of HMM state mapping-based cross-lingual speaker adaptation. Firstly, the effect of unsupervised cross-lingual adaptation was investigated, as it relates to the application scenario of personalized speech-to-speech translation. The comparison of paired supervised and unsupervised systems shows that the performance of unsupervised cross-lingual speaker adaptation is comparable to that of the supervised fashion, even if the average phoneme error rate of the unsupervised systems is around 75%.

Then the effect of the language mismatch between synthesis models and adaptation data was investigated. The mismatch is found to transfer undesirable language information from adaptation data to synthesis models, thereby limiting the effectiveness of generating multiple regression class-specific transforms, using larger quantities of adaptation data and estimating adaptation transforms iteratively.

Thirdly, in order to tackle the problems caused by the language mismatch, a data-driven adaptation framework using phonological knowledge is proposed. Its basic idea is to group HMM states according to phonological knowledge in a data-driven manner and then to map each state to a phonologically consistent counterpart in a different language. This framework

is also applied to regression class tree construction for transform estimation. It is found that the proposed framework alleviates the negative effect of the language mismatch and gives consistent improvement compared to previous state-of-the-art approaches.

Finally, a two-layer hierarchical transformation framework is developed, where one layer captures speaker characteristics and the other compensates for the language mismatch. The most appropriate means to construct the hierarchical arrangement of transforms was investigated in an initial study. While early results show some promise, further in-depth investigation is needed to confirm the validity of this hierarchy.

Keywords language mismatch, HMM state mapping, data-driven enhancement, two-layer adaptation hierarchy, cross-lingual speaker adaptation, personalized speech-to-speech translation, HMM-based speech synthesis

摘要

本文的研究工作因開發個人化的語音到語音翻譯這一目標而開始，重點放在其一項關鍵技術——用於文語轉換的跨語言說話人自適應。個人化的語音到語音翻譯器能夠在將一個人的講話翻譯成另一種語言的同時，在合成語音中保留這個人的音色。

在著手處理技術問題之前，本文的研究工作首先著眼於理解人對音色的感知，希望借聽辨實驗確定普通人是否能夠在若干說話人講不同語言的時候將他們區分開來。聽辨實驗結果顯示，在跨語言的時候區分不同的說話人是可以辦到的。然而，要在跨語言且跨語音類型（原始錄音和合成語音相比較）的情況下區別不同的說話人就比較困難了。

跨語言說話人自適應裡的核心挑戰是如何在自適應數據的語言和語音合成模型的語言不同的時候應用說話人自適應技術。本論文工作的主體部分放在分析和改進基於隱馬模型狀態映射的跨語言說話人自適應。首先，鑒于無監督跨語言說話人自適應和個人化的語音到語音翻譯的應用場合相關，本文對它的效果加以研究。對成對的有監督及無監督的系統的對比顯示，無監督跨語言說話人自適應的性能與有監督時的性能相當，即使這些無監督的系統的平均音位識別錯誤率達75%左右。

在此之後，本文研究了語音合成模型和自適應數據之間語言不匹配所帶來的影響。本文發現，這種不匹配把並不需要的語言信息從自適應數據轉移到了語音合成模型裡，故而限制了對如下幾個方面的有效利用：生成多個針對特定回歸類別的轉換矩陣、使用數量較大的自適應數據、迭代式估計自適應轉換矩陣。

第三，為了解決由語言不匹配引發的問題，本文提出一個使用語音知識的數據驅動的自適應框架。它的基本想法是根據語音知識將隱馬模型的狀態以數據驅動的方式進行分組，然後將每個狀態映射到另一種語言中一個語音類別一致的狀態上。這個框架也被用於建立估計轉換矩陣所需的回歸類樹。本文發現，這個框架可以減弱語言不匹配的影響，即使供測試的說話人不同，系統性能也有一致的提高。

最後，本文研究了一個雙層的變換框架——一層用於捕獲說話人的特徵，另一層用於補償語言不匹配。本文對最恰當的建立這個雙層變換框架的方法進行了研究。初步的結果顯示這個雙層框架有研究價值，但它還需進一步深入研究以證實確實有效。

關鍵詞 語言不匹配，隱馬模型狀態映射，數據驅動型改進，雙層自適應層次結構，跨語言說話人自適應，個人化的語音到語音翻譯，基於隱馬模型的語音合成

(Translated by the author of the thesis as per the English version)

Acknowledgements / 致謝

I am so delighted that I was admitted by *École Polytechnique Fédérale de Lausanne* and could pursue a PhD degree in the past four years at the Idiap Research Institute directed by Prof. Hervé Bourlard. I am really grateful that my thesis director, Prof. Hervé Bourlard, provided me with a great research environment – sufficient funding, powerful computing facilities and an atmosphere of openness and mutual assistance. I want to thank the EMIME project ("Effective Multilingual Interaction in Mobile Environment" in the Seventh Framework Programme of the European Union; 31 months), the Idiap Research Institute (7 months) and the Hasler Foundation (10.5 months) very much for sponsoring my research work. I am also very grateful for the time (nearly one month) and energy that the thesis jury members (Prof. Hervé Bourlard, Prof. Keiichi Tokuda, Prof. Tanja Schultz, Dr Jean-Marc Vesin and Prof. Jean-Philippe Thiran) spent in reviewing this thesis, as well as the constructive comments and suggestions they kindly provided.

非常高興，我能被瑞士洛桑聯邦理工學院錄取，並在Hervé Bourlard教授治下之達爾莫爾感知人工智能研究所裡度過四年多博士生的學習生活。感謝我的論文主管Hervé Bourlard教授在這四年裡為我提供了一個理想的研究場所——充足的項目資金、强大的工作硬件系統、良好互助的研究氛圍。感謝歐盟第七次框架計劃的「移動環境下高效多語言交互」項目（三十一個月）、達爾莫爾感知人工智能研究所（七個月）和Hasler基金會（十個半月）對我研究工作的資助。同時，非常感謝諸位論文評審（Hervé Bourlard教授、德田惠一教授、Tanja Schultz教授、Jean-Marc Vesin博士和Jean-Philippe Thiran教授）在審閱本論文上所花的近一個月的時間和精力，以及他們對這篇論文的寶貴的反饋建議。

I am deeply indebted to my supervisor, Mr John Dines! Undoubtedly the PhD thesis you are reading wouldn't exist without John's help. Honestly, I have no idea how I can express my gratitude completely and precisely by language. Perhaps the only way is to bow by 90 degrees, as we normally do in East Asia. In the past four years, John offered me encouragement, support, guidance and advice. I cannot forget how relaxed I felt after each weekly meeting with him, as John could always ease my stress caused by experimental results and strengthen my confidence. John always tried to broaden my horizons, to help me think about a problem from other perspectives and to remind me to pay attention to other relevant research topics. John seemed to be never tired of revising my paper/thesis drafts (the content as well as the

English wording). Sometimes John sacrificed his spare time to revise my papers and reports, for instance, on his way to a holiday resort or at weekends (so I hope Mrs Dines didn't feel unhappy ^_^). John never minded helping and advising me on issues irrelevant to my PhD work, for example, job hunting. In summary, John did countless things for me in the past four years. I really appreciate all of them!

非常非常感謝我的導師John Dines先生！毫無疑問，沒有John的幫助，你正在閱讀的這篇博士論文就不會存在。坦白地說，我不知道應當如何用語言精準地表達對John的感謝，或許唯有東亞人的傳統禮儀——一個九十度的鞠躬才能完整地傳遞出我的感激。四年來，John對我有鼓勵、有支持，有引導、有建議。我無法忘記每週和他開會之後心裡的暢快感覺，因為John總能緩解實驗結果讓我產生的焦慮，為我增强信心！John一向不忘拓展我的眼界，幫助我從不同角度看問題，提醒我關注和我相關的一些其它研究方向。但凡寫論文，John總是不厭其煩地多次修改我的草稿（内容以及英語表述）。John有時會犧牲自己的空餘時間幫我修改論文和報告，比如他休假旅行的途中，或是週末（但愿Dines太太沒覺得不高興^_^）。John也從不介意在和研究工作無關的事情上給我建議與幫助，比如找工作。總之，在這四年裡John為我操了很多心。我很感激他為我做的一切！

Mr Philip Garner is unofficially my co-supervisor. Because of the difference between our research directions, Phil didn't supervise my research work. Nonetheless, Phil encouraged me from time to time. In the past four years, especially when John was busy, Phil was always willing to help me or to offer advice, no matter whether I encountered a major or minor issue. Chats with Phil were always enjoyable because of his sense of humour. Thank you, Phil!

Philip Garner先生算是我的副導師。因為研究方向之差異，Phil並不指導我的學術研究，不過他對我的工作也時有鼓勵。四年來，尤其是在John忙碌的時候，不論大事小事，Phil總是非常樂意地幫我解決，或是提供建議。Phil很幽默，和他交談總是非常愉快。謝謝你，Phil！

Idiap is located in the French-speaking part of Switzerland. So there have been always only a handful of native speakers of English out of the 100 or so Idiapers. I was so lucky, as John is Australian and Phil is British. Having being "flanked" by them for four years, I feel that my English has been improved significantly: not only did the four basic skills (listening, speaking, reading and writing) get improved, but also I learned appropriate and euphemistic English as well as colloquial expressions from them. Quite often people who met me for the first time asked me three questions: "Where do you come from?", "How long have you lived in England?", "How did you learn English (or the British accent)?" To be honest, even now, I still feel nervous somehow when speaking English with John or Phil. However, I am absolutely sure that I am a thousand times more confident than I was four years ago, when I have to speak English with a non-Chinese. I presume that the English knowledge I learned from Phil and John was more than anything else that I learned in the past four years. Thanks a lot for the help that they kindly offered and that they might be unconscious of in the four years!

因為研究所坐落在瑞士的法語區，所以整個研究所的百來號人裡，一直只有屈指可數的幾位以英語為母語。我非常幸運，因為John是澳大利亞人，Phil是英國人。在他們兩位長達四年的「夾擊」之下，我感覺我的英文水平有了突飛猛進的提高，無論是英文的聽、說、讀、寫基本功，還是英語裡的禮貌用辭、委婉表達和日常口頭語。初次見面的人經常會問我三個問題：「你從哪裡來？」「你在英國住過多久？」「你是怎麼學的英文（或英式口音）？」老實說，在John和Phil面前，即使是現在，我講英文的時候仍然會覺得不知名的緊張；但在面對其他外國人、需要講英文的時候，我心裡非常清楚，我比四年前自信了一萬倍。我都懷疑，我從Phil和John那裡學到的英文，比我這四年裡學到的其它任何東西都多。非常感謝兩位四年來有心以及無意中的幫助！

I would like to thank my Indian colleague, Ms Lakshmi Saheer as well, with whom I spent four years in studying and working together in the same small research group. We helped each other while working for the same project and travelling for symposiums. I really appreciate the harmonious workmate relationship, which was another source of happiness during the four years.

我還要感謝我的印度同事Lakshmi Saheer女士。我們在同一個小研究組裡度過了四年時光，在學習和研究中互相幫助，在會議旅行中互相照應。這種融洽的同事關係讓我的四年時光過得相當愉快。

I want to thank Dr Junichi Yamagishi based in the Centre for Speech Technology Research (CSTR), the University of Edinburgh and Dr Michael Pucher based in the Telecommunications Research Center Vienna (FTW). They kindly offered me speech synthesis training scripts and German questions for decision tree-based clustering respectively, which played a critical role in my research. I also need to thank Dr Mirjam Wester based in CSTR for the collaborative work on bilingual (Mandarin and English) speech data recording and human perception of speaker identity across languages (and across speech types). This work forms an important part of my thesis.

感謝愛丁堡大學語音技術研究中心的山岸順一博士為我提供語音合成訓練腳本，以及維也納電信研究中心的Michael Pucher博士為我提供用於德語的決策樹聚類的語音問題。它們在我的研究工作中扮演了非常重要的角色。我還要感謝愛丁堡大學語音技術研究中心的Mirjam Wester博士與我共同完成雙語（中文普通話和英語）語音的錄製工作和跨語言（以及跨語音類型）條件下說話人音色辨識的研究工作。這項工作是我論文中的重要一環。

Switzerland was absolutely a strange country when I just arrived four years ago. Luckily, Jie LUO, my colleague at Idiap, picked me up at the Zürich airport. Weifeng LI and his wife kindly offered suggestions for me to settle in Martigny. Thank you so much for the convenience you brought me in the very beginning!!

回憶起四年前，我初到陌生的瑞士，羅傑到蘇黎世機場為我接機，李衛鋒夫婦為我最初在Martigny的生活安頓提供建議。他們的熱情幫助給了我極大的方便。很感謝這幾

位熱心的同事！！

I also want to thank a few friends whom I got to know via the Internet, as they brought happiness into my monotonous life abroad. First of all, Yannic Evers, a friend from Germany. We have been getting along quite well though he is ten years younger than I am. I feel honoured to be viewed as his tutor in the Chinese language and he kindly promised me that he would try his best to help me whenever I wanted to study German. He gave me a German name, Erik Löwe, as a gift in return. I like it a lot. Then, a Swiss friend, Daniel Rusidovski. We had cosy chats and travelled together a couple of times. The birthday greeting from him was a complete surprise.

我還想感謝幾位通過互聯網認識的朋友。他們給我單調的國外生活帶來了許多歡樂。首先，我的德國朋友艾雍奕。雖然他比我年輕將近十歲，可是我們仍然很談得來。我很榮幸能被他視為個人漢語老師，並得到他的許諾——如果我想學習德文，他將全力幫助。他花時間幫我想到的德文名「Erik Löwe」，我相當喜歡。除了小奕，還有瑞士朋友Daniel Rusidovski。我們有過愉快的聊天、幾次一起出行。他發給我的生日祝福，完全出乎我的意料。

I am really happy knowing some local Christian friends, especially the Taylors (Audrey & Robert) from Britain and several young missionaries (Jim Law, Austin Larsen, Armand Overstreet, Vincent Dieduksman, Carter J. Pilling, Daniel James, etc). When being together with them, I felt unaffected friendship and concern, and saw virtues such as charity, integrity and tolerance which are shared by the East Asian and Christian cultures.

我很高興能認識一些瑞士當地信仰基督教的朋友們，尤其是講英文的泰勒老兩口（奧黛麗和羅伯特）和幾位年輕的傳教士（Jim Law、Austin Larsen、Armand Overstreet、Vincent Dieduksman、Carter J. Pilling、Daniel James等）。和他們在一起時，我感受到了一份質樸的友誼和關懷，看到了諸如仁愛、正直、寬容等在東亞和基督教文化里共存的美德。

I would like to thank Laurent El Shafey for translating the English abstract of this thesis into French as well as giving me a lovely, childlike nickname “Oui-Oui” ;-). I need to thank Ms Sylvie Millius, Ms Nadine Rousseau and the system maintenance crew for helping me to handle miscellaneous administrative and technical problems.

感謝研究所的Laurent El Shafey為我翻譯出了法文版的論文摘要，以及他給我的可愛的頑童暱稱「Oui-Oui」;-)。感謝研究所的兩位行政管理（Sylvie Millius和Nadine Rousseau）和所有系統管理維護人員在平日為我解決種種繁雜小事。

I have been always missing Prof. Frank Soong and my then mentor Ms Yao QIAN based in Microsoft Research Asia. I was very happy each time I met them in an academic conference. They opened a door for me by giving me an opportunity of internship, which has been far-reaching to my career or even my life.

我也一直很想念身在微軟亞洲研究院的宋詠平老師和錢瑤女士。每次在學術會議上碰到他們，我都覺得非常開心。沒有當年他們給我的那個影響深遠的實習機會，也就不會有今天的我。他們為我打開了一扇門。

This long list below enumerates others who once helped me or brought me happiness in the past four years (names are listed in random order and some could be missing): Yang Sun, Mathew Magimai Doss, David Imseng, Dinesh Babu Jayagopi, Tatiana Tommasi, Jagannadan Varadarajan, Nicolae Suditu, Edgar Roman-Rangel, Joan-Isaac Biel, Dairazalia Sanchez-Cortes, Marco Fornoni, Trinh-Minh-Tri Do, Aniruddha Adiga, Serena Soldo, Ramya Rasipuram, Leonidas Lefakis, Marc Ferras, Deepu Vijayasenan, Eileen Yi-Lee Lew, Radu-Andrei Negoescu, Venkatesh Bala Subburaman, Sree Hari Krishnan Parthasarathi, Sree Harsha Yella, Petr Motlicek, Bogdan Raducanu, Giulia Garau, Alfred Dielmann, Raphael Ullmann, Niklas Johansson, Hugo Penedones, Riwal Lefort, Thomas Meyer, Miloš Cerňak, Benjamin Picart, Paul Gay, Gokul Chittaranjan, Roy Geoffrey Wallace, Senera Soldo, Zoltan Tüske, Laurent El Shafey, Mary Knox, Sriram Prasath Elango, Jesus Martínez-Gómez, Samuel Kim, Constantin-Cosmin Atanasaoei, Afsaneh Asaei, David Alexander Gregory, Joshua Rundell, Florian Peter Sladky, Sebastian Kempe, Vincent Michellod.

接下來這個長長的名單，記錄著這四年裡其他曾經幫助過我或者給我帶來過快樂的人（人名以隨機順序列出，或有遺漏）：孫陽、Mathew Magimai Doss、David Imseng、Dinesh Babu Jayagopi、Tatiana Tommasi、Jagannadan Varadarajan、Nicolae Suditu、Edgar Roman-Rangel、Joan-Isaac Biel、Dairazalia Sanchez-Cortes、Marco Fornoni、杜明智、Aniruddha Adiga、Serena Soldo、Ramya Rasipuram、Leonidas Lefakis、Marc Ferras、Deepu Vijayasenan、劉怡利、Radu-Andrei Negoescu、Venkatesh Bala Subburaman、Sree Hari Krishnan Parthasarathi、Sree Harsha Yella、Petr Motlicek、Bogdan Raducanu、Giulia Garau、Alfred Dielmann、Raphael Ullmann、Niklas Johansson、Hugo Penedones、Riwal Lefort、Thomas Meyer、Miloš Cerňak、Benjamin Picart、Paul Gay、Gokul Chittaranjan、Roy Geoffrey Wallace、Senera Soldo、Zoltan Tüske、Laurent El Shafey、Mary Knox、Sriram Prasath Elango、Jesus Martínez-Gómez、金Samuel、Constantin-Cosmin Atanasaoei、Afsaneh Asaei、David Alexander Gregory、Joshua Rundell、傅培安、康紹嚴、Vincent Michellod。

Finally, I am so thankful for the constant loving concern over time from my parents and the whole family!

最後，感謝我的父母和所有家人對我長久以來不變的關懷與牽掛！

As an East Asian guy who has lived in Europe for over four years, I sometimes felt that East Asia was the farthest place in the world from Europe, geographically, culturally, ethnically and linguistically. It is not an easy task for me to write the emotive acknowledgements properly and appropriately in English. There could be misuses of English verbs, adjectives, adverbs, etc in the above text. In order to express my gratitude unambiguously, I also wrote the acknowledgements in my mother tongue, the Chinese language.

作為一個在歐洲生活了四年多的東亞人，我有時覺得，東亞是這個世界上離歐洲最為遙遠的地方，無論是從地理、文化、人種還是語言的角度來看。用英文得體且恰當地書寫極富情感的致謝部分，對我來說，並非易事。你也許已經在前文中發現了不少英文用詞錯誤。故而我在本頁以母語中文再次書寫了致謝一文，以便精準地記錄下我的感激之情。

梁　晖
2012年11月16日

To non-Chinese speakers:
"Hui Liang" is not my name but a poor phonetic transcription of my name.

Contents

List of Figures

List of Tables

Glossary

Technical acronyms in this thesis are exhaustively listed below in alphabetical order.

BNDAP	band aperiodicity
CAT	cluster adaptive training
CLSA	cross-lingual speaker adaptation
CMLLR	constrained maximum likelihood linear regression
CSMAPLR	constrained structural maximum *a posteriori* linear regression
(D)MOS	(differential) mean opinion score
EM	expectation-maximization
GV	global variance
H(S)MM	hidden (semi-)Markov model
IPA	International Phonetic Alphabet
K-L, **KLD**	Kullback-Leibler, Kullback-Leibler divergence
MCD	mel-cepstral distortion
MCEP	mel-cepstrum
MGE	minimum generation error
MMC	modèle de Markov caché
MSD	multi-space distribution
pdf	probability density function
RMSE	root-mean-square error
STRAIGHT	speech transformation and representation using adaptive interpolation of weighted spectrum

1 Introduction

1.1 Motivations

Language is a powerful tool of communication. Human beings enjoy the freedom to easily communicate with one another due to the use of sophisticated languages. Regrettably, we also suffer a great deal from the fact that there exist in the world a huge number of languages which are often mutually unintelligible. The language barrier is a prominent hurdle to overcome in order to facilitate better communication among people across the globe. Efforts to clear this hurdle have been attempted long before the rise of technological solutions. For example, quite a few auxiliary languages were invented and supposed to play the role of a lingua franca, such as Esperanto, Ido, Interlingua, Lojban, etc. However, even Esperanto, the best-known among all these auxiliary languages [Byram, 2004, page 464], remains of little importance more than 120 years after its debut. On the other hand, although English, a natural language, is functioning as the de facto lingua franca of today for historical reasons, this does not mean at all that one can travel around the world without difficulty in communicating with locals.

Learning a foreign language undoubtedly requires a lot of time and energy. It would be highly desirable that technology can lend a hand in freeing people of the obstacle posed by the language barrier. Real-time automated speech-to-speech translation [Levin et al., 2000, Zhou et al., 2003], 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, has emerged as an important research topic. Researchers working on this topic have been following the straightforward architecture, which consists of three consecutive modules – automatic speech recognition, machine translation and text-to-speech synthesis, to build automated speech-to-speech translators (as shown in Figure 1.1).

The output voice identity of the speech synthesis module in Figure 1.1 usually comes from a professional speaker (e.g., the system presented in [Bangalore et al., 2012] and the Google Translate service) who has recorded a large amount of training data, so that high quality of output synthesized speech can be guaranteed. This is a mature, but time-consuming and costly solution. It is not realistic to collect a variety of voices and, as a result, the speech

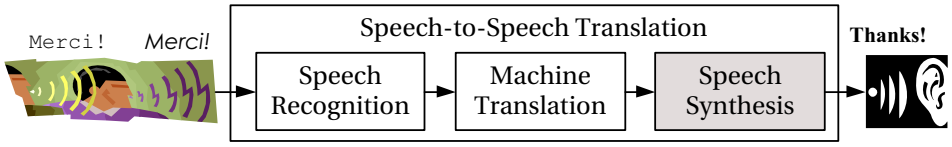


Figure 1.1 – Typical architecture of an automated speech-to-speech translator

synthesis module lacks voice diversity. As Figure 1.1 shows, two different speakers speak to the translator but the same synthetic voice is heard. Having the same output voice may impede communication when several people use speech-to-speech translators at the same time. For the sake of voice diversity, research is being conducted on personalization of automated speech-to-speech translation, namely, to discover how to make the output synthetic voice sound like a user's input voice despite the difference in language between the two. An exemplar is the project called *Effective Multilingual Interaction in Mobile Environments*[1] (EMIME) [Kurimo et al., 2010], which was mainly aimed at building a mobile device with personalized speech-to-speech translation embedded such that one would be able to "speak" any foreign language easily. Figure 1.2 visualizes the general idea.

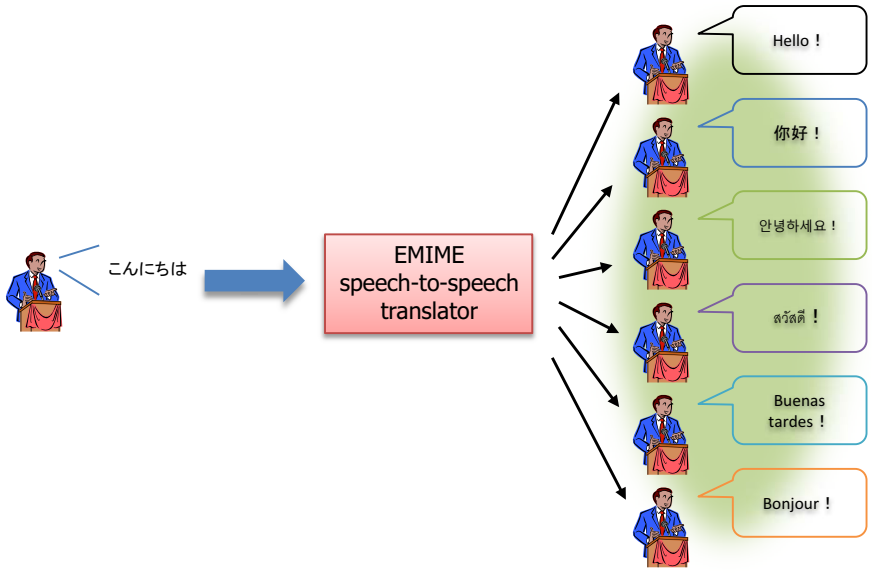


Figure 1.2 – Personalization of automated speech-to-speech translation (e.g., the EMIME project)

Personalization of speech synthesis in recent research relies on speaker adaptation, a technology which can produce synthesized speech in a particular speaker's voice using merely tens of adaptation utterances collected from this speaker. In the context of personalized speech-to-speech translation, the key speaker adaptation technology is generally called *cross-lingual speaker adaptation* (CLSA) [Wu et al., 2008, Chen et al., 2009, Wu et al., 2009, Gibson et al.,

1. http://www.emime.org

2010, Oura et al., 2010, Peng et al., 2010] for text-to-speech synthesis. In other words, the focus of research is how to adapt a speech synthesis module trained on speech data in a desired language with a certain number of adaptation utterances in a different language collected from a target speaker. So far cross-lingual speaker adaptation for speech synthesis is a fairly new research topic that has not yet been investigated in depth.

There exist two dominant solutions to text-to-speech synthesis: unit selection (concatenative synthesis) [Hunt and Black, 1996] and HMM-based speech synthesis (statistical parametric synthesis) [Zen et al., 2009]. Unit selection produces new utterances by concatenating natural speech segments selected from a large pre-recorded corpus, trying to minimize a weighted summation of target costs (i.e., how well a candidate speech segment from the corpus matches the required one) and concatenation costs (i.e., how well two adjacent candidate speech segments combine). Personalization of unit selection relies on applying voice conversion techniques [Kain and Macon, 1998, Sündermann et al., 2006] to these natural candidate speech segments. However, voice conversion techniques have limited ability to capture the full range of speaker variability [Watts et al., 2009] and are detrimental to the high quality of natural speech. Furthermore, the large pre-recorded corpus normally contains only a handful of speakers due to the high cost of collection of a great deal of speech data. The difference between the voice characteristics in the pre-recorded corpus and those of a target speaker may be considerable and as a result cause additional difficulty in voice conversion. In summary, unit selection is not a good choice when voice diversity is demanded in output synthesized speech.

By contrast, owing to its statistical parametric nature, HMM-based speech synthesis 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 speech-to-speech translation since it includes a range of highly effective speaker adaptation algorithms that centre around the so-called *average voice* synthesis paradigm [Yamagishi and Kobayashi, 2007, Yamagishi, 2006]. An average voice is an artificial voice trained by speaker adaptive training [Anastasakos et al., 1996] on speech data collected from tens or even hundreds of real speakers, ideally modelling speaker-independent phonetic and prosodic variations only. Since an average voice is obtained by averaging out speaker characteristics of many real speakers, it would not differ remarkably from the voice in adaptation data in most cases [Yamagishi et al., 2010a]. Although it is preferred to collect a lot of training data from each real speaker, tens of utterances per speaker are acceptable in practice for training an average voice synthesizer [Yamagishi et al., 2010a]. These are two main advantages of the average voice synthesis paradigm. Before speech parameter generation, an average voice is adapted towards a given target speaker by means of speaker adaptation algorithms like CMLLR [Gales, 1998]. As only tens of adaptation utterances are needed from the target speaker, voice diversity in output synthesized speech can be easily achieved. Consequently, the HMM-based speech synthesis framework and the average voice synthesis paradigm are the foundation of the thesis work.

1.2 Scope of the Thesis

As mentioned above, the thesis work was motivated by personalization of speech-to-speech translation. Prior to addressing any technical difficulties in developing personalized speech-to-speech translation, it is firstly necessary to understand human perception of speaker identity, i.e., to determine whether or not people can distinguish between speakers across languages and also speech types (natural versus synthesized). Listening tests were conducted to help answer this question.

A key component technique of personalization of speech-to-speech translation is cross-lingual speaker adaptation for speech synthesis. It is a fairly new topic and previous relevant research is limited. After comparing state-of-the-art approaches, HMM state mapping [Wu et al., 2009] is selected to enable cross-lingual speaker adaptation for the thesis work. In order to discover major difficulties in state mapping-based cross-lingual speaker adaptation, the unsupervised adaptation approach and the impact of the language mismatch between synthesis models and adaptation data are investigated. "Language mismatch" refers to the fact that the acoustic space, phoneme inventory, prosodic patterns, articulatory features and so forth of a language partially overlap those of another language.

Then the two following approaches to improving cross-lingual speaker adaptation are focused upon. They both require a bilingual corpus containing many speakers.

1. Typically the minimum Kullback-Leibler divergence [Kullback and Leibler, 1951] criterion is employed to determine state mapping relations. In order to enhance this simple criterion, a jointly data-driven and phonological knowledge-guided approach is proposed. This approach is adjusted and also applied to the generation of regression class trees with a more appropriate structure for transform estimation.
2. A two-layer transformation framework is investigated, where the two layers capture language information and speaker characteristics respectively. The goal is to factorize language information out of speaker characteristics so that the language mismatch will not have any impact on synthesis quality. Initial experiments towards the establishment of such a hierarchy and training the two layers of transforms are presented.

Though speech-to-speech translation involves speech recognition, machine translation and speech synthesis, the main focus of the thesis is only on speech synthesis. The other two components are minimally touched.

1.3 Contributions to the State of the Art

The main contributions in the following chapters to the state of the art of cross-lingual speaker adaptation for speech synthesis can be summarized as follows:

1. The ability of people to distinguish between speakers across different languages is inves-

tigated and confirmed. Speech quality is found to play a significant role in distinguishing between speakers.

2. The possibility of using unsupervised cross-lingual speaker adaptation for personalized speech-to-speech translation is examined. Unsupervised cross-lingual speaker adaptation is found to be comparable to supervised cross-lingual speaker adaptation.

3. The mismatch between the input and output languages is found to be a major detrimental factor in cross-lingual speaker adaptation. It hampers the effectiveness of regression class tree-based adaptation, thereby limiting the ability of adaptation algorithms to benefit from larger quantities of adaptation data. It also hampers the effectiveness of iterative estimation of adaptation transforms.

4. Jointly data-driven and phonological knowledge-guided enhancement under the minimum generation error criterion is proposed and applied to both state mapping construction and regression class tree growth. It alleviates the negative effect of the mismatch between the input and output languages and gives consistent improvement compared to previous state-of-the-art approaches.

5. A linear transformation-based two-layer hierarchy is developed, where one layer captures speaker characteristics and the other compensates for the mismatch between the input and output languages. The basic structure and training methodology of this hierarchy have been determined based on the limited number of available bilingual speakers.

1.4 Outline of the Thesis

This thesis is composed of 7 chapters. Chapter 2 gives an overview of hidden Markov models, speaker adaptation and the HMM-based speech synthesis framework (including its training, synthesis and evaluation stages).

In Chapter 3, multilingual speech processing, the state of the art of cross-lingual speaker adaptation for text-to-speech synthesis, required speech resources and the challenges of evaluating cross-lingual speaker adaptation systems are discussed. The ability of people to distinguish between speakers across languages is investigated in Chapter 3.

In Chapter 4, several paired supervised and unsupervised cross-lingual speaker adaptation systems are compared, in order to examine the possibility of using unsupervised cross-lingual adaptation in the context of personalized speech-to-speech translation. Then the focus of Chapter 4 moves on to the investigation of the impact of the language difference between adaptation data and average voice synthesis models. An intra-lingual speaker adaptation system and four kinds of HMM state mapping-based cross-lingual speaker adaptation systems are compared. Various thresholds are used in the comparison to adjust the number of regression class-specific adaptation transforms. The iterative fashion of transform estimation in the context of cross-lingual speaker adaptation is also examined.

In Chapter 5, a jointly data-driven and phonological knowledge-guided approach is proposed for the purpose of enhancing HMM state mapping construction and regression class tree growth. Firstly, the purely data-oriented minimum K-L divergence criterion is improved by introducing phonological constraints into the procedure of HMM state mapping construction. Then phonological knowledge is applied to guiding regression class tree construction. The effectiveness and generalization across speakers of the proposed approach are evaluated in this chapter. Finally, the possibility of iterative enhancement in an alternating fashion is examined.

In Chapter 6, a hierarchical transformation framework is designed, in which there exist two layers of linear transforms capturing target speaker characteristics and language information respectively. How this hierarchy should be constructed and trained is investigated through several adaptation experiments.

The thesis concludes with Chapter 7, where the contributions and limitations of the findings in the thesis are presented. Possible future work is also discussed.

2 Statistical Parametric Speech Synthesis

The very first speech synthesizers developed by Christian Gottlieb Kratzenstein and Wolfgang von Kempelen individually more than 200 years ago were mechanical apparatus that mimicked human organs of articulation (e.g., vocal tract, vocal folds and so forth). They were able to produce simple sounds like /aː/, /eː/, /iː/, /oː/, /uː/, etc [Schroeder, 1993]. In addition to such mechanical synthesis techniques, researchers also developed electrical synthesis techniques such as articulatory synthesis, source-filter synthesis, concatenative synthesis [Klatt, 1987] and statistical parametric synthesis [Zen et al., 2009].

Nowadays the two dominant speech synthesis techniques are concatenative synthesis and statistical parametric synthesis. One of the reasons is that storing a vast quantity of speech recordings is no longer a problem. Concatenative speech synthesis is a straightforward technique, which produces an artificial utterance by concatenating natural speech segments that are selected from a pre-recorded corpus as per a certain criterion. Artificial speech of high quality and with good naturalness can be achieved through this technique, because the costly pre-recorded corpus is normally very large, covering sufficient variation in the production of speech.

Nevertheless, the inflexibility of concatenative speech synthesis becomes a formidable obstacle when voice diversity is required. In the last two decades, the statistical parametric HMM-based framework and its peripheral speaker adaptation technology, which were originally devised for automatic speech recognition, were introduced into speech synthesis and have received a great deal of attention from the speech synthesis research community. The HMM-based speech synthesis framework provides an elegant and principled solution to handle voice diversity. This chapter presents a brief overview of the fundamentals of statistical parametric speech synthesis and speaker adaptation, which form the foundations of this thesis work.

2.1 Hidden Markov Models

The hidden Markov model was proposed by Leonard E. Baum and his colleagues in the 1960s [Baum and Petrie, 1966, Baum et al., 1970] and was introduced into speech recognition research in the 1980s [Bahl et al., 1983, Poritz and Richter, 1986, Lippmann et al., 1987, Lee et al., 1988, Rabiner et al., 1989, Lee, 1989]. This introduction led to a major advance in the research on speech processing and had a profound impact on it.

2.1.1 Fundamentals

A hidden Markov model is a finite state machine that generates a sequence of discrete-time observations. At a particular time t, it changes its state from $q_{t-1} = i$ into $q_t = j$ according to state transition probabilities $\{a_{i,j}\}$, and then generates an observation $\boldsymbol{o}_t$ according to the output probability distribution of state j ("observation" refers to feature representations of speech signals). The modifier "hidden" refers to the fact that i and j are unknown. That is, the state that generates $\boldsymbol{o}_t$ cannot be directly observed, though $\boldsymbol{o}_t$ is known.

An HMM $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi})$ consisting of N emitting states can be specified by the three factors $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\Pi}$: state transition probabilities $\boldsymbol{A} = \{a_{i,j} \,|\, i = 1,2,\cdots,N, j = 1,2,\cdots,N\}$ (from state i to state j), output probability distributions $\boldsymbol{B} = \{b_j(\boldsymbol{o}_t) \,|\, j = 1,2,\cdots,N\}$ and initial state probabilities $\boldsymbol{\Pi} = \{\pi_i \,|\, i = 1,2,\cdots,N\}$. Depending on the values of π_i and $a_{i,j}$ (i.e., zero or non-zero), the topology of λ may be ergodic or left-to-right without skips.

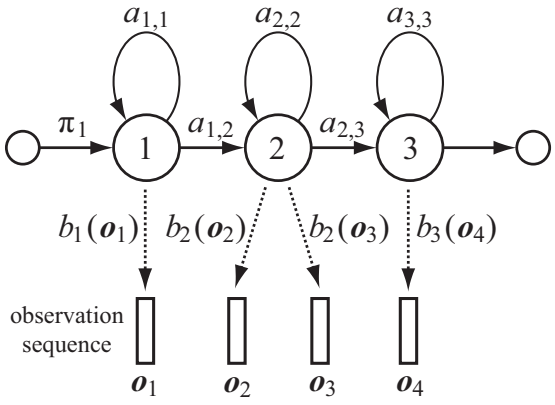


Figure 2.1 – 3-state left-to-right HMM without skips

Figure 2.1 presents an illustration of a 3-state left-to-right HMM with no skips (i.e., it is not possible to move from state 1 to state 3 directly). This kind of HMM topology particularly suits speech signal modelling: (1) Speech signals are a temporal series, meaning that normally there should not be a skip over one or more following states or reversion to a previous state; (2) Speech signals can be approximately considered stable in a very short period (e.g., 5ms), which corresponds to $\{a_{i,i} \,|\, i = 1,2,\cdots,N\}$; (3) Speech signals vary over time, which corresponds to $\{a_{i,i+1} \,|\, i = 1,2,\cdots,N-1\}$; (4) Speech signals themselves can be described by

$\{b_j(\boldsymbol{o}_t) \,|\, j = 1,2,\cdots,N\}$.

The output probability distributions $\{b_j(\boldsymbol{o}_t) \,|\, j = 1,2,\cdots,N\}$ may be either discrete or continuous. When used for speech signal modelling, $\{b_j(\boldsymbol{o}_t) \,|\, j = 1,2,\cdots,N\}$ are continuous and usually composed of a mixture of multivariate Gaussian distributions as follows:

$$b_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} w_{jm} \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right), \tag{2.1}$$

where M is the number of Gaussian mixture components in state j; w_{jm}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$ are the weight, mean vector and covariance matrix of the m-th Gaussian mixture component in state j respectively. The weights $\{w_{jm} \,|\, m = 1,2,\cdots,M\}$ must satisfy the constraints

$$\sum_{m=1}^{M} w_{jm} = 1, \qquad j = 1,2,\cdots,N \tag{2.2}$$

$$w_{jm} \geqslant 0, \qquad j = 1,2,\cdots,N, \quad m = 1,2,\cdots,M \tag{2.3}$$

such that

$$\int_{\boldsymbol{o}_t} b_j(\boldsymbol{o}_t)\,\mathrm{d}\boldsymbol{o}_t = 1, \qquad j = 1,2,\cdots,N. \tag{2.4}$$

In case that the observation vector $\boldsymbol{o}_t$ can be divided into S independent streams (e.g., spectral and excitation features are modelled by different streams in HMM-based speech synthesis), $\{b_j(\boldsymbol{o}_t) \,|\, j = 1,2,\cdots,N\}$ can be reformulated as follows:

$$\boldsymbol{o}_t = \left[\boldsymbol{o}_{t,1}^\top \quad \boldsymbol{o}_{t,2}^\top \quad \cdots \quad \boldsymbol{o}_{t,S}^\top\right]^\top,$$

$$b_j(\boldsymbol{o}_t) = \prod_{s=1}^{S} b_{js}(\boldsymbol{o}_{t,s}) \tag{2.5}$$

$$= \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} w_{jsm} \mathcal{N}\left(\boldsymbol{o}_{t,s}; \boldsymbol{\mu}_{jsm}, \boldsymbol{\Sigma}_{jsm}\right)\right]^{\gamma_s}, \tag{2.6}$$

where M_s is the number of Gaussian mixture components in stream s; w_{jsm}, $\boldsymbol{\mu}_{jsm}$ and $\boldsymbol{\Sigma}_{jsm}$ are the weight, mean vector and covariance matrix of the m-th Gaussian mixture component in stream s of state j respectively; γ_s is the weight of stream s.

2.1.2 Three Fundamental Problems

There are three fundamental problems with respect to HMMs: (1) how to calculate the probability of a particular observation sequence; (2) how to find the optimal state sequence that generates a given observation sequence; (3) how to optimize HMM parameters given an observation sequence. This subsection touches on the three problems in brief.

Calculation of the Probability of a Particular Observation Sequence

If the HMM state sequence that generates an observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T)$ is known to be $\boldsymbol{Q} = (q_1, q_2, \cdots, q_T)$, calculating the probability of $\boldsymbol{O}$ being generated by λ is a trivial problem. The formula is simply

$$p(\boldsymbol{O}, \boldsymbol{Q}|\lambda) = \pi_{q_1} b_{q_1}(\boldsymbol{o}_1) \cdot \prod_{t=2}^{T} a_{q_{t-1},q_t} b_{q_t}(\boldsymbol{o}_t). \tag{2.7}$$

But given the "hidden" nature of HMMs, $\boldsymbol{Q}$ is actually an invisible sequence. All the possible $\boldsymbol{Q}$ should be taken into consideration. As a result, the probability of $\boldsymbol{O}$ being generated by HMMs λ should be calculated by

$$p(\boldsymbol{O}|\lambda) = \sum_{\text{all } \boldsymbol{Q}} p(\boldsymbol{O}, \boldsymbol{Q}|\lambda) \tag{2.8}$$

$$= \sum_{\text{all } \boldsymbol{Q}} \left[\pi_{q_1} b_{q_1}(\boldsymbol{o}_1) \cdot \prod_{t=2}^{T} a_{q_{t-1},q_t} b_{q_t}(\boldsymbol{o}_t) \right]. \tag{2.9}$$

It is not possible to calculate $p(\boldsymbol{O}|\lambda)$ directly, as "all $\boldsymbol{Q}$" corresponds to N^T permutations of states and $\boldsymbol{Q}$ is a very long sequence in practice (i.e., T is a large number). Hence the efficient forward-backward algorithm is employed to solve this problem. Forward and backward probabilities are defined as follows:

$$\text{Forward: } \alpha_t(j) = p(\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_t, q_t = j|\lambda), \tag{2.10}$$

$$\text{Backward: } \beta_t(j) = p(\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, \cdots, \boldsymbol{o}_T | q_t = j, \lambda). \tag{2.11}$$

They are initialized by

$$\alpha_1(j) = \pi_j b_j(\boldsymbol{o}_1), \qquad j = 1, 2, \cdots, N, \tag{2.12}$$

$$\beta_T(j) = 1, \qquad j = 1, 2, \cdots, N \tag{2.13}$$

and calculated recursively using

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) \cdot a_{i,j} \right] b_j(\boldsymbol{o}_t), \qquad j = 1, 2, \cdots, N, \quad t = 2, 3, \cdots, T, \tag{2.14}$$

$$\beta_t(j) = \sum_{k=1}^{N} a_{j,k} b_k(\boldsymbol{o}_{t+1}) \beta_{t+1}(k), \qquad j = 1, 2, \cdots, N, \quad t = 1, 2, \cdots, T-1. \tag{2.15}$$

Then the probability of generating an observation sequence $p(\boldsymbol{O}|\lambda)$ can be obtained by

$$p(\boldsymbol{O}|\lambda) = \sum_{j=1}^{N} \alpha_t(j) \beta_t(j), \qquad t = 1, 2, \cdots, T. \tag{2.16}$$

Determination of the Optimal State Sequence that Generates a Given Observation Sequence

Among all the N^T permutations of states, there exists a sequence $\boldsymbol{Q}^* = \left(q_1^*, q_2^*, \cdots, q_T^*\right)$ which maximizes $p(\boldsymbol{O}, \boldsymbol{Q}|\lambda)$. This optimal state sequence is useful for decoding, initializing HMM parameters at the training stage, etc. The Viterbi algorithm can efficiently find the optimal state sequence $\boldsymbol{Q}^*$ given an observation sequence $\boldsymbol{O}$ and HMM parameters λ.

Suppose $\delta_t\left(j\right)$ denotes the probability of the optimal state sequence until time t and ending at state j, i.e.,

$$\delta_t\left(j\right) = \max_{q_1, q_2, \cdots, q_{t-1}} p\left(q_1, q_2, \cdots, q_{t-1}, q_t = j, \boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_t, \middle| \lambda\right), \tag{2.17}$$

which is initialized by

$$\delta_1\left(j\right) = \pi_j b_j(\boldsymbol{o}_1), \qquad j = 1, 2, \cdots, N \tag{2.18}$$

and calculated recursively using

$$\delta_t\left(j\right) = \left[\max_{i \in \{1,2,\cdots,N\}} \delta_{t-1}(i) \cdot a_{i,j}\right] b_j(\boldsymbol{o}_t), \qquad j = 1, 2, \cdots, N, \quad t = 2, 3, \cdots, T. \tag{2.19}$$

Having obtained the N values of $\delta_T\left(q_T\right)$, we can determine q_T^* by choosing the maximal $\delta_T\left(q_T\right)$ and then trace back to q_1^* along the path which corresponds to $\delta_T\left(q_T^*\right)$ in order to find the entire optimal state sequence $\boldsymbol{Q}^*$ that generates $\boldsymbol{O}$.

Optimization of HMM Parameters Given an Observation Sequence

HMM parameters are typically estimated under the maximum likelihood criterion, i.e., to estimate λ which maximizes $p(\boldsymbol{O}|\lambda)$ given an observation sequence $\boldsymbol{O}$. Unfortunately there is no closed solution to this problem. The Baum-Welch algorithm [Baum, 1972], which is a special case of the expectation-maximization (EM) algorithm [Dempster et al., 1977], is generally employed for estimation of λ. Given initial values calculated by flat-start or using annotations, it functions in an iterative manner to update λ until $p(\boldsymbol{O}|\lambda)$ converges. The EM fashion guarantees that $p(\boldsymbol{O}|\lambda)$ increases as the estimation process of λ is repeated, but it is very likely that $p(\boldsymbol{O}|\lambda)$ converges at a local maximum rather than at a global maximum.

In the E-step of each iteration, the auxiliary function $\mathcal{Q}(\tilde{\lambda}, \lambda)$ of λ to be estimated given $\tilde{\lambda}$ from the previous iteration is computed by

$$\mathcal{Q}(\tilde{\lambda}, \lambda) = \sum_{\text{all } \boldsymbol{Q}} p\left(\boldsymbol{Q}|\boldsymbol{O}, \tilde{\lambda}\right) \log p\left(\boldsymbol{O}, \boldsymbol{Q}|\lambda\right). \tag{2.20}$$

In the subsequent M-step, $\mathcal{Q}(\tilde{\lambda}, \lambda)$ is maximized with respect to λ, since it can be proved that this is equivalent to maximizing $p(\boldsymbol{O}|\lambda)$ with respect to λ. As a result of the maximization, λ is

updated with the following equations [Bilmes, 1998]:

$$\pi_i = \gamma_1(i), \qquad i = 1,2,\cdots,N \tag{2.21}$$

$$a_{i,j} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)}, \qquad i = 1,2,\cdots,N, \quad j = 1,2,\cdots,N \tag{2.22}$$

$$w_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)}{\sum_{t=1}^{T} \gamma_t(i)}, \qquad i = 1,2,\cdots,N, \quad m = 1,2,\cdots,M \tag{2.23}$$

$$\boldsymbol{\mu}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m) \cdot \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(i,m)}, \qquad i = 1,2,\cdots,N, \quad m = 1,2,\cdots,M \tag{2.24}$$

$$\boldsymbol{\Sigma}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m) \cdot \left(\boldsymbol{o}_t - \boldsymbol{\mu}_{im}\right)\left(\boldsymbol{o}_t - \boldsymbol{\mu}_{im}\right)^{\top}}{\sum_{t=1}^{T} \gamma_t(i,m)}, \qquad i = 1,2,\cdots,N, \quad m = 1,2,\cdots,M \tag{2.25}$$

where $\gamma_t(i)$ is the probability of being in state i at time t, $\gamma_t(i,m)$ is the probability of being in the m-th sub-state distribution of state i at time t and $\xi_t(i,j)$ is the probability of being in state i at time t and state j at time (t+1). They are computed according to the following equations:

$$\begin{aligned}\gamma_t(i) &= \frac{p\left(\boldsymbol{O}, q_t = i \middle| \lambda\right)}{\sum_{i=1}^{N} p\left(\boldsymbol{O}, q_t = i \middle| \lambda\right)}, \\ &= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}, \qquad i = 1,2,\cdots,N, \quad t = 1,2,\cdots,T;\end{aligned} \tag{2.26}$$

$$\begin{aligned}\gamma_t(i,m) &= \frac{p\left(\boldsymbol{O}, q_t = i, s_t = m \middle| \lambda\right)}{\sum_{l=1}^{N}\sum_{n=1}^{M} p\left(\boldsymbol{O}, q_t = l, s_t = n \middle| \lambda\right)}, \\ &= \gamma_t(i) \cdot \frac{w_{im}\mathcal{N}\left(\boldsymbol{o}_t \middle| \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}\right)}{\sum_{k=1}^{M} w_{ik}\mathcal{N}\left(\boldsymbol{o}_t \middle| \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right)}, \\ &\qquad\qquad i = 1,2,\cdots,N, \quad t = 1,2,\cdots,T, \quad m = 1,2,\cdots,M;\end{aligned} \tag{2.27}$$

$$\begin{aligned}\xi_t(i,j) &= \frac{p\left(\boldsymbol{O}, q_t = i, q_{t+1} = j \middle| \lambda\right)}{\sum_{l=1}^{N}\sum_{n=1}^{N} p\left(\boldsymbol{O}, q_t = l, q_{t+1} = n \middle| \lambda\right)}, \\ &= \frac{\alpha_t(i)a_{i,j}b_j(\boldsymbol{o}_{t+1})\beta_{t+1}(j)}{\sum_{l=1}^{N}\sum_{n=1}^{N} \alpha_t(l)a_{l,n}b_n(\boldsymbol{o}_{t+1})\beta_{t+1}(n)}, \\ &\qquad\qquad i = 1,2,\cdots,N, \quad j = 1,2,\cdots,N, \quad t = 1,2,\cdots,T.\end{aligned} \tag{2.28}$$

2.1.3 Context-Dependent Modelling

In general, a single language contains tens of phonemes. Technically an HMM can be easily trained for each of the phonemes to build a speech processing system, but in practice, such a system performs poorly, because acoustic realizations of phonemes in natural speech vary widely depending on their contexts and simple phoneme models cannot capture the substantial variations. A straightforward solution is to model context-dependent phones rather than isolated phonemes. For instance, triphones are widely employed as a modelling unit. A triphone model `L-C+R` describes how the core phoneme `C` is articulated when it is preceded by a phoneme `L` and succeeded by a phoneme `R`, thereby being affected by coarticulation.

In order to capture all the specific acoustic variations of phonemes properly, it is necessary to train every context-dependent model robustly. However, the number of context-dependent models increases exponentially with the size of the contextual window. It is not possible to guarantee that each context-dependent model can be trained over sufficient speech data. A compromise solution to this data sparsity problem is to share training data across similar context-dependent models such that model parameters receive adequate data for their robust estimation. An additional problem to be solved is how to estimate models for contexts that have not been observed at all in training data.

The most common technique for sharing training data across context-dependent models is decision tree-based clustering [Young et al., 1994]. Its basic idea is depicted in Figure 2.2. A set of phonologically derived questions needs to be prepared beforehand (see this figure for examples). These questions can divide context-dependent models into different clusters in each of which the context-dependent models are close to one another. Meanwhile, these clusters generalize to unobserved contexts in training data.

A decision tree is grown in a top-down manner as follows. Initially, all the context-dependent model distributions (denoted by $\mathbb{S}$ as a set) derived from training data observations $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_T)$ are pooled to form a root node. The log likelihood $L(\mathbb{S})$ of training data $\boldsymbol{O}$ being generated by $\mathbb{S}$ is calculated on the assumption that all of the context-dependent model distributions in that node are merged to form a shared mean vector $\boldsymbol{\mu}(\mathbb{S})$ and a shared covariance matrix $\boldsymbol{\Sigma}(\mathbb{S})$. A reasonable approximation of $L(\mathbb{S})$ is given by

$$L(\mathbb{S}) \approx \sum_{t=1}^{T} \sum_{S \in \mathbb{S}} \log\Big(p\big(\boldsymbol{o}_t \,\big|\, \boldsymbol{\mu}(\mathbb{S}), \boldsymbol{\Sigma}(\mathbb{S})\big)\Big) \gamma_S(\boldsymbol{o}_t), \tag{2.29}$$

where $\gamma_S(\boldsymbol{o}_t)$ is the posterior probability of $\boldsymbol{o}_t$ being generated by model distribution S. This node is then split into two, $\mathbb{S}_{\text{yes}}(\mathfrak{q})$ and $\mathbb{S}_{\text{no}}(\mathfrak{q})$, by finding the question $\mathfrak{q}$ which partitions the context-dependent model distributions in the parent node so as to give the maximum increase in log likelihood, i.e., to maximize $\Delta L_{\mathfrak{q}}$ defined by

$$\Delta L_{\mathfrak{q}} = L\big(\mathbb{S}_{\text{yes}}(\mathfrak{q})\big) + L\big(\mathbb{S}_{\text{no}}(\mathfrak{q})\big) - L(\mathbb{S}) \tag{2.30}$$

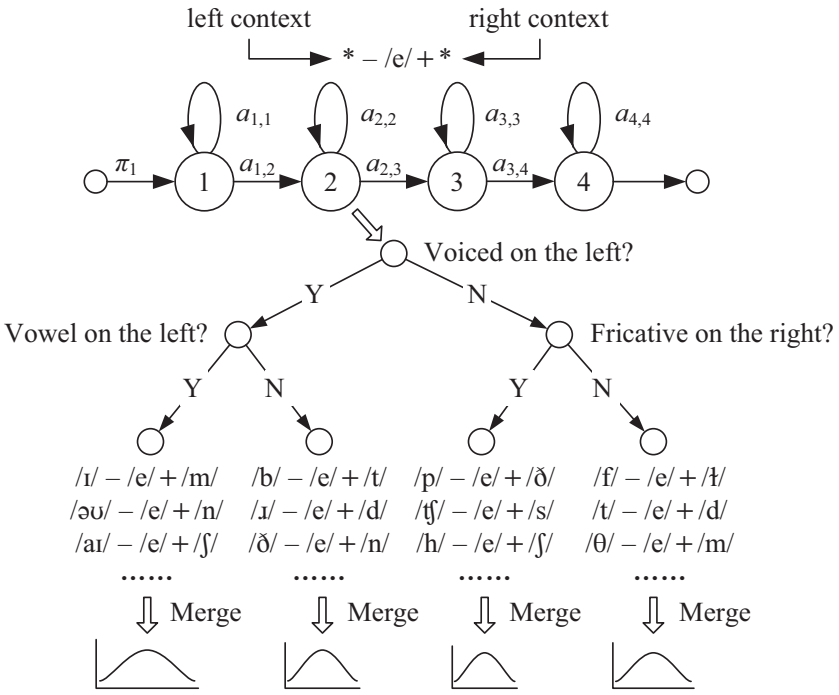


Figure 2.2 – Example of decision tree-based clustering of triphone models

with respect to q. This process is then repeated by splitting the node which yields the greatest increase in log likelihood until the increase falls below a predefined threshold. To ensure that all terminal nodes have sufficient training data associated with them, a minimum occupation count is applied.

Alternatively, the above maximum likelihood criterion can be replaced with minimum description length (MDL) [Shinoda and Watanabe, 2000]. It can be shown that the MDL criterion is equivalent to maximum likelihood with a stopping criterion on the basis of a likelihood threshold that is calculated with respect to model complexity. As a result, the minimum occupation count does not need to be applied. Finally, training data is shared within each leaf node to estimate the tied context-dependent model distribution.

2.2 Speaker Adaptation

In order to build a personalized speech processing system, we can collect speech data from the target speaker and then train a set of *speaker-dependent* HMMs on his/her data alone. Unfortunately, a set of robust speaker-dependent HMMs requires a large amount of training data from the target speaker, typically hundreds or thousands of utterances. This requirement makes the speaker-dependent solution expensive, time-consuming and impractical for situations where diversity of target speakers is expected. Due to the statistical parametric nature of the HMM-based speech processing framework, speaker adaptation techniques [Gales, 1998] have been developed in order to address this problem. By means of speaker adaptation, the

voice characteristics of "source" HMMs can be adapted to those of a target speaker, given only tens of adaptation utterances in the target speaker's voice. In fact the "source" HMMs can be any well-trained models but generally they are trained on speech data collected from multiple speakers in order not to be biased towards any particular type of speaker (i.e., to be speaker-independent).

2.2.1 Maximum Likelihood Linear Transformation

Model-space linear transformation is a simple, powerful and widely used approach to speaker adaptation. It can be used to estimate speaker-specific linear transforms that capture the differences between "source" speaker-independent models and given adaptation data, and to apply them to Gaussian mixture components of the speaker-independent models in order to adapt voice characteristics towards those of the given adaptation data. Such speaker-specific linear transforms are typically estimated under the maximum likelihood criterion, i.e., the combination of original distributions of the speaker-independent models and these linear transforms should maximize the likelihood of the given adaptation data being generated by this combination.

In the simplest case, a single set of *global* transforms $\left(\hat{\boldsymbol{A}}'_s, \hat{\boldsymbol{b}}'_s, \hat{\boldsymbol{H}}'_s\right)$ is applied to every Gaussian mixture component of "source" speaker-independent models for adaptation towards a target speaker s's voice as follows:

$$\boldsymbol{\mu}_{s,m} = \hat{\boldsymbol{A}}'_s \boldsymbol{\mu}_m + \hat{\boldsymbol{b}}'_s, \tag{2.31}$$

$$\boldsymbol{\Sigma}_{s,m} = \hat{\boldsymbol{H}}'_s \boldsymbol{\Sigma}_m \hat{\boldsymbol{H}}'^{\top}_s, \tag{2.32}$$

where $\left\{\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \middle| m = 1, 2, \cdots, M\right\}$ are mean vectors and covariance matrices of the M Gaussian mixture components of the speaker-independent models. $\left(\hat{\boldsymbol{A}}'_s, \hat{\boldsymbol{b}}'_s, \hat{\boldsymbol{H}}'_s\right)$ is the result of the following expression when transform estimation is carried out under the maximum likelihood criterion:

$$\begin{aligned}\left(\hat{\boldsymbol{A}}'_s, \hat{\boldsymbol{b}}'_s, \hat{\boldsymbol{H}}'_s\right) &= \underset{(\boldsymbol{A}'_s, \boldsymbol{b}'_s, \boldsymbol{H}'_s)}{\arg\max}\, p\left(\boldsymbol{O}_s \middle| \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right), \\ \boldsymbol{O}_s &= \left(\boldsymbol{o}_{s,1}, \boldsymbol{o}_{s,2}, \cdots, \boldsymbol{o}_{s,T}\right), \\ \boldsymbol{\mu}_s &= \left(\boldsymbol{\mu}_{s,1}, \boldsymbol{\mu}_{s,2}, \cdots, \boldsymbol{\mu}_{s,M}\right), \\ \boldsymbol{\Sigma}_s &= \left(\boldsymbol{\Sigma}_{s,1}, \boldsymbol{\Sigma}_{s,2}, \cdots, \boldsymbol{\Sigma}_{s,M}\right),\end{aligned} \tag{2.33}$$

where $\left\{\boldsymbol{o}_{s,t} \middle| t = 1, 2, \cdots, T\right\}$ are observations in speaker s's voice from time 1 to time T and $\left\{\boldsymbol{\mu}_{s,m}, \boldsymbol{\Sigma}_{s,m} \middle| m = 1, 2, \cdots, M\right\}$ are mean vectors and covariance matrices of the M Gaussian mixture components that have been adapted to speaker s's voice.

2.2.2 Regression Class

A single set of global transforms obtained according to Eq. (2.33) cannot fully capture the voice characteristics of a target speaker and furthermore cannot make good use of adaptation data when there is a moderately large amount (for example, 100~200 utterances). Basically, the more adaptation data there is, the greater number of transforms should be trained. Hence, one solution is to divide the M speaker-independent Gaussian distributions into groups, the number of which depends on the amount of adaptation data such that a more finely grained transform can be robustly estimated for each group of Gaussian distributions.

A regression class tree is usually involved for the purpose of *automatically* adjusting the number of finely grained transforms according to the amount of adaptation data and each leaf node is a regression class (see Figure 2.3). A regression class tree is traversed in the top-down manner during transform estimation and the search starts at the root node. Transforms are generated only for the nodes which (i) have sufficient data and (ii) are either leaf nodes or have any children without sufficient data. For example, a shared transform is generated when neither node 3 nor node 4 has enough data but they as a whole do; the transform for node 2 is generated using data and distributions from both nodes 1 and 2 when node 2 does not have enough data but node 1 does. This mechanism has an advantage that as many, robust, finely grained transforms as possible can be estimated on available adaptation data. Whether adaptation data is sufficient for a node is determined by a threshold [1] on the number of adaptation data frames associated with the node.

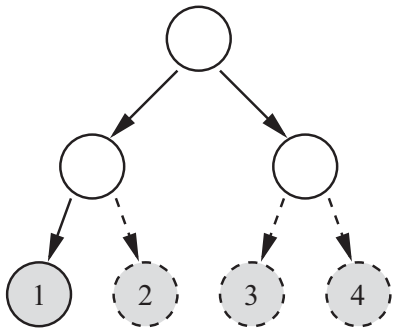


Figure 2.3 – Regression class tree

There are two main methods of generating a regression class tree. One is to pool all the Gaussian distributions of "source" speaker-independent models at a single node and to keep splitting all the leaf nodes according to a distribution similarity measure (e.g., the Euclidean distance between mean vectors [Young et al., 2009, Chapter 9]) and a stopping threshold on the measure. The other one is to connect the root nodes of the decision trees obtained in the training stage of "source" speaker-independent models to form a regression class tree, which is especially beneficial to adaptation of pitch [Yamagishi et al., 2004]. This is by default the method of generating a regression class tree in this thesis, unless a different method is

1. For instance, it is `HADAPT:SPLITTHRESH` in HTK.

mentioned explicitly.

2.2.3 Constrained Maximum Likelihood Linear Regression

One example of model-space maximum likelihood linear transformation is constrained maximum likelihood linear regression (CMLLR) [Gales, 1998], where "constrained" means transformation matrices are jointly applied to mean and covariance parameters (i.e., $\boldsymbol{H}'_{s,i} \equiv \boldsymbol{A}'_{s,i}$, $i = 1,2,\cdots,I$). Interestingly, this constraint allows CMLLR to be regarded and implemented as feature-space transformation, i.e., transformation matrices can be applied to speech features instead of speaker-independent model parameters. Implementing CMLLR as feature-space transformation provides the additional benefit of full covariance modelling, where full covariance statistics are captured in CMLLR transforms, thus requiring fewer parameters for estimation.

From the perspective of feature-space transformation, CMLLR estimates a set of linear transforms for speech features of given adaptation data such that the likelihood of the adaptation data is maximized. An above-mentioned regression class tree may be involved in the course of estimation. To be specific, CMLLR produces I linear transforms

$$\hat{\boldsymbol{W}}_s = \left\{ \hat{\boldsymbol{W}}_{s,i} \middle| \hat{\boldsymbol{W}}_{s,i} = \begin{bmatrix} \hat{\boldsymbol{b}}_{s,i} & \hat{\boldsymbol{A}}_{s,i} \end{bmatrix}, i = 1,2,\cdots,I \right\},$$

where $\hat{\boldsymbol{A}}_{s,i}$ is a square matrix and $\hat{\boldsymbol{b}}_{s,i}$ is a column vector so as to capture speaker s's voice characteristics from T observation frames of his/her adaptation data $\left\{ \boldsymbol{o}_{s,t} \middle| t = 1,2,\cdots,T \right\}$. $\hat{\boldsymbol{W}}_s$ is the result of the following expression:

$$\begin{aligned}
\hat{\boldsymbol{W}}_s &= \arg\max_{\boldsymbol{W}_s} \sum_{\text{all } \boldsymbol{Q}_s} p\left(\bar{\boldsymbol{O}}_s, \boldsymbol{Q}_s \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \qquad (2.34)\\
\bar{\boldsymbol{O}}_s &= \left(\bar{\boldsymbol{o}}_{s,1}, \bar{\boldsymbol{o}}_{s,2}, \cdots, \bar{\boldsymbol{o}}_{s,T}\right),\\
\bar{\boldsymbol{o}}_{s,t} &= \boldsymbol{A}_{s,\mathscr{X}_1(t)} \boldsymbol{o}_{s,t} + \boldsymbol{b}_{s,\mathscr{X}_1(t)}, \qquad t = 1,2,\cdots,T, \quad \mathscr{X}_1(t) \in \{1,2,\cdots,I\},\\
\boldsymbol{\mu} &= \left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_M\right),\\
\boldsymbol{\Sigma} &= \left(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_M\right),
\end{aligned}$$

where $\left\{ \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \middle| m = 1,2,\cdots,M \right\}$ are mean vectors and covariance matrices of the M Gaussian distributions of "source" speaker-independent models, $\boldsymbol{Q}_s$ is a possible state sequence corresponding to $\bar{\boldsymbol{O}}_s$, and $\mathscr{X}_1(t)$ represents mapping relations from a feature frame $\boldsymbol{o}_{s,t}$ to an adaptation transform $\hat{\boldsymbol{W}}_{s,i}$.

Going back to the perspective of model-space transformation, we can create speaker-adapted models by applying the inverse of CMLLR transforms $\hat{\boldsymbol{W}}_s$ to the "source" speaker-independent models:

$$\boldsymbol{\mu}_{s,m} = \hat{\boldsymbol{A}}^{-1}_{s,\mathscr{X}_2(m)} \left(\boldsymbol{\mu}_m - \hat{\boldsymbol{b}}_{s,\mathscr{X}_2(m)}\right), \qquad m = 1,2,\cdots,M, \quad \mathscr{X}_2(m) \in \{1,2,\cdots,I\}, \qquad (2.35)$$

$$\boldsymbol{\Sigma}_{s,m} = \hat{\boldsymbol{A}}_{s,\mathscr{X}_2(m)}^{-1} \boldsymbol{\Sigma}_m \left(\hat{\boldsymbol{A}}_{s,\mathscr{X}_2(m)}^{-1} \right)^\top, \qquad m = 1, 2, \cdots, M, \quad \mathscr{X}_2(m) \in \{1, 2, \cdots, I\}, \tag{2.36}$$

where $\mathscr{X}_2(m)$ defines mapping relations from a Gaussian distribution $\left(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right)$ to an adaptation transform $\hat{\boldsymbol{W}}_{s,i}$.

CSMAPLR

Constrained structural maximum *a posteriori* linear regression (CSMAPLR) [Nakano et al., 2006, Yamagishi et al., 2009a] is a speaker adaptation algorithm that improves the performance of speaker adaptation by CMLLR. $\hat{\boldsymbol{W}}_s$ is estimated as per Eq. (2.37) in CSMAPLR, i.e., under the structural maximum *a posteriori* criterion [Shinoda and Lee, 2001] instead of the maximum likelihood criterion (see Eq. (2.34) for the contrast):

$$\hat{\boldsymbol{W}}_s = \arg\max_{\boldsymbol{W}_s} \sum_{\text{all } \boldsymbol{Q}_s} p\left(\bar{\boldsymbol{O}}_s, \boldsymbol{Q}_s \middle| \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) p\left(\boldsymbol{W}_s\right), \tag{2.37}$$

where $p\left(\boldsymbol{W}_s\right)$ is the prior distribution of $\boldsymbol{W}_s$. The matrix Gaussian distribution is used for $p\left(\boldsymbol{W}_{s,i}\right)$ $(i = 1, 2, \cdots, I)$:

$$p\left(\boldsymbol{W}_{s,i}\right) \propto |\boldsymbol{\Omega}|^{-\frac{L+1}{2}} |\boldsymbol{\Psi}|^{-\frac{L}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left[\left(\boldsymbol{W}_{s,i} - \boldsymbol{H}_{s,i}\right)^\top \boldsymbol{\Omega}^{-1} \left(\boldsymbol{W}_{s,i} - \boldsymbol{H}_{s,i}\right) \boldsymbol{\Psi}^{-1} \right] \right\}, \tag{2.38}$$

where L is the dimensionality of speech features, $\text{tr}(\cdot)$ calculates the trace of a matrix, $\boldsymbol{\Omega} \in \mathbb{R}^{L \times L}$, $\boldsymbol{\Psi} \in \mathbb{R}^{(L+1)\times(L+1)}$ and $\boldsymbol{H}_{s,i} \in \mathbb{R}^{L \times (L+1)}$ are three hyperparameters of the prior distribution. In CSMAPLR, $\boldsymbol{\Psi}$ is fixed to an identity matrix and

$$\boldsymbol{\Omega} \equiv \tau \cdot \boldsymbol{I}_{L \times L} = \begin{bmatrix} \tau & & & \\ & \tau & & \\ & & \ddots & \\ & & & \tau \end{bmatrix}_{L \times L}, \qquad \tau > 0.$$

CSMAPLR requires a regression class tree in the course of transform estimation and $\boldsymbol{H}_{s,i}$ refers to the transform associated with a corresponding parent node. When estimating a transform for the root node of the regression class tree, this hyperparameter is set to $\begin{bmatrix} \boldsymbol{0}_{L\times 1} & \boldsymbol{I}_{L\times L} \end{bmatrix}$.

For the sake of simplicity, techniques that involves linear transformation-based speaker adaptation are reviewed/explained merely in terms of CMLLR in this thesis.

2.2.4 Speaker Adaptive Training

It is possible to train speaker-independent models on a speech corpus containing a lot of speakers by following the exact training procedure of speaker-dependent models. Due to the large number of contexts encountered in speech synthesis compared to the number of training speakers, it is highly likely that conventional decision tree clustering will lead to

overly specialized leaf nodes that do not provide good speaker-independent modelling. In the extreme case, each leaf node many come to represent a few contexts uttered by only a handful of training speakers. Though this problem can be solved by shared-decision-tree-based context clustering [Yamagishi, 2006, Chapter 4], speaker-independent models estimated in this fashion still capture both desired phonetic variations and unwanted variations among training speakers. The variations among training speakers often lead to distributions with overly large variances, as Figure 2.4 illustrates.

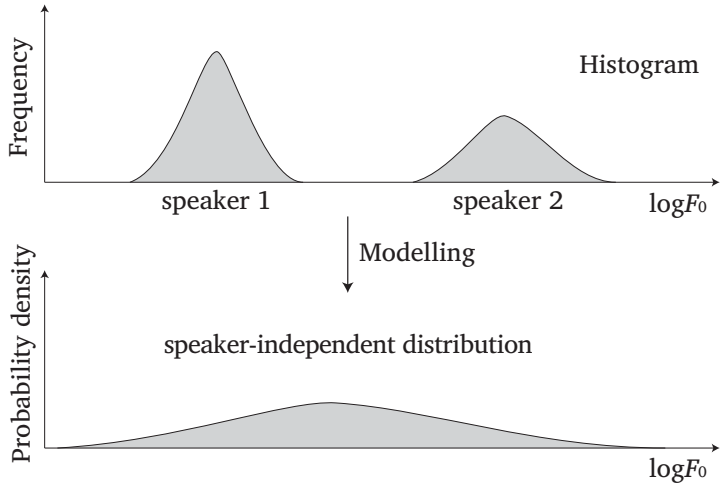


Figure 2.4 – Speaker-independent model distribution [Yamagishi, 2006, Chapter 5]

As a result, the speaker adaptive training (SAT) paradigm [Anastasakos et al., 1996] was proposed to separate unwanted variations among training speakers from phonetic variations. It jointly estimates a set of canonical SAT models that capture phonetic variations and training speaker-specific adaptation transforms that capture speaker variations. The following equations highlight the difference between speaker-independent model training in the speaker-dependent fashion and by speaker adaptive training:

$$\text{speaker-independent:} \qquad \arg\max_{\lambda} \prod_{s=1}^{S} p\left(\boldsymbol{O}_s|\lambda\right) \tag{2.39}$$

$$\text{speaker adaptive:} \qquad \arg\max_{(\lambda,\mathcal{G})} \prod_{s=1}^{S} p\left(\boldsymbol{O}_s|G_s(\lambda)\right) \tag{2.40}$$

where $\mathcal{G} = (G_1, G_2, \cdots, G_S)$ and $G_s(\cdot)$ denotes model transformation towards training speaker s. Obviously, it is possible to take advantage of multiple regression classes for the estimation of these speaker-specific adaptation transforms. In practice, speaker adaptive training is initialized with speaker-independent models and proceeds in an iterative manner: updating $\mathcal{G}$, then mean vectors followed by covariance matrices, and finally back to $\mathcal{G}$. After only a few iterations, the speaker-independent distribution in Figure 2.4 may converge to a point where it looks like the one in Figure 2.5.

CMLLR is particularly well suited to speaker adaptive training, as it can be implemented as

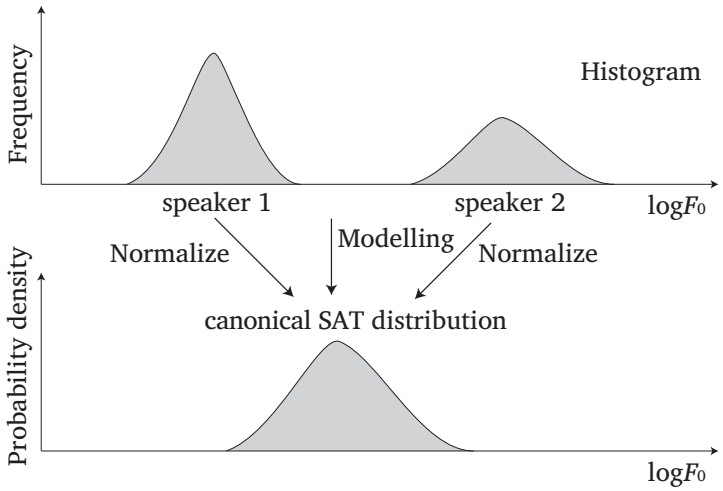


Figure 2.5 – Canonical SAT model distribution [Yamagishi, 2006, Chapter 5]

feature-space linear transformation so that normalized speech features, which ideally contain no variation among training speakers, can be easily obtained and used to update canonical model parameters [Young et al., 2009, Chapter 9].

2.3 HMM-Based Text-to-Speech Synthesis

HMM-based text-to-speech synthesis [Tokuda et al., 2002b] is a statistical parametric approach to speech synthesis. Parametrized speech, i.e., spectral features, excitation features and duration information, are modelled by context-dependent HMMs during the training stage [Yoshimura et al., 1999]. According to context-dependent labels derived from input plain text by a text analyzer, well-trained context-dependent HMMs are concatenated and speech parameters, which are finally converted into waveforms, are generated from the HMM sequence [2] [Tokuda et al., 1995a,b, 2000]. This parametric nature makes HMM-based speech synthesis a highly flexible solution – HMM parameters can be easily adjusted in order to achieve various speaker identities, speaking styles, etc. Figure 2.6 presents a flow chart of this process.

2.3.1 Basics

First of all, we discuss a few fundamental issues about the state-of-the-art HMM-based speech synthesis framework in this section.

2. Such an HMM sequence is viewed as a single and longer HMM in speech parameter generation.

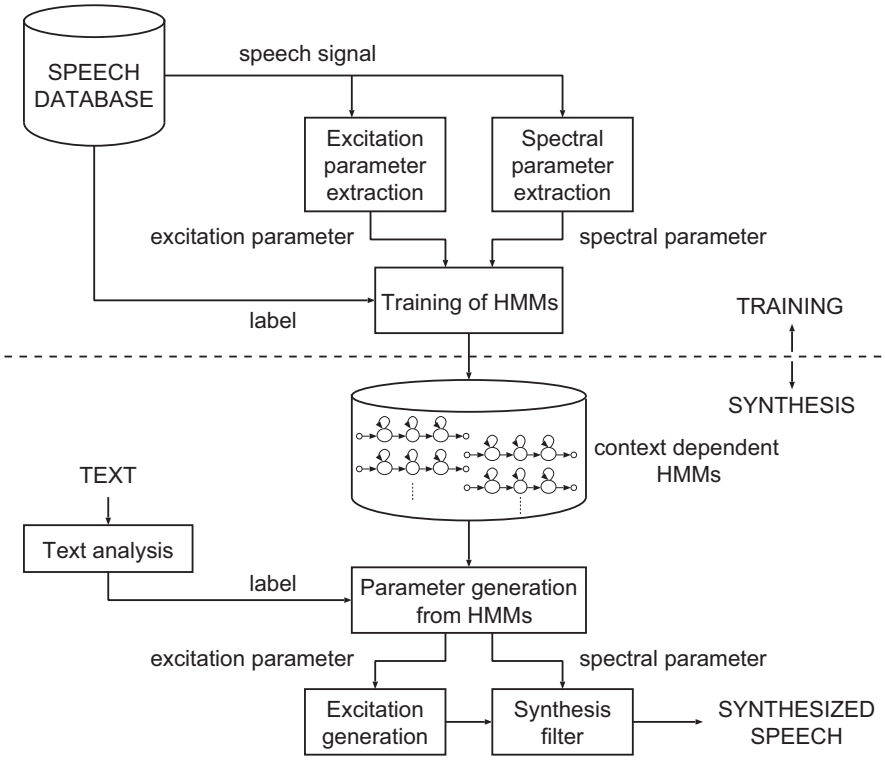


Figure 2.6 – Flow chart of HMM-based speech synthesis [Tokuda et al., 2002b]

HMM Topology

Typically, five-state (or three-state) left-to-right HMMs with no skip are employed for all the modelling units in HMM-based speech synthesis [Zen et al., 2009]. Research on model topology was conducted, for example, stochastic Markov graphs applied in [Eichner et al., 2000, 2001]. Though this was a flexible topology, it significantly increased computational complexity of speech parameter generation.

Acoustic Features

Spectral features, F_0 and band aperiodicity are modelled in different HMM streams for speech synthesis. A key requirement of the feature representation is that it should allow reconstruction of speech signals while having the requisite properties to be well modelled by HMMs. Commonly used spectral features include mel-(generalized) cepstrum [Tokuda et al., 1994], line spectrum pair [Soong and Juang, 1984], etc.

The STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) vocoding technique [Kawahara et al., 1999] is widely used for spectrum analysis and speech generation, as it generates more accurate smoothed spectrum and produces synthetic speech of high quality. STRAIGHT explicitly uses extracted F_0 information to conduct pitch-adaptive spectrum analysis combined with a surface reconstruction method

in the time-frequency region to remove periodic components from estimated spectrum. The estimated spectrum is then converted into spectral features like mel-cepstrum. An aperiodicity measure that represents the relative energy distribution of aperiodic components in the frequency domain is also extracted [Kawahara et al., 2001]. Band aperiodicity, which is employed to construct mixed excitation [3] for speech waveform generation, is comprised of the averages of the aperiodicity measurements over a certain number of frequency bands (e.g., five bands: 0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz and 6-8 kHz [Zen and Toda, 2005]). STRAIGHT employs an FFT-based process to generate waveforms.

F_0 Modelling

Because of the existence of voiced and unvoiced phonemes in languages, F_0 contours are intrinsically composed of segments with and without F_0 values. The multi-space distribution was proposed [Tokuda et al., 2002a] to model this kind of discontinuous speech feature. More specifically, F_0 is modelled by two spaces, one of which contains a normal, one-dimensional continuous Gaussian distribution (i.e., the "voiced space") while the other contains no distribution but a single sample point (i.e., the "unvoiced space"). The two spaces have their respective weights, indicating the probability of a frame being voiced or unvoiced.

A multi-space distribution is similar to but more general than a Gaussian mixture model, as it is allowed to contain various sorts of distributions in one model. For example, both discrete and continuous distributions are contained at the same time in the case of F_0 modelling. If each space contains a Gaussian distribution and the dimensionality of all the Gaussian distributions is a positive constant, the multi-space distribution degenerates into a Gaussian mixture model.

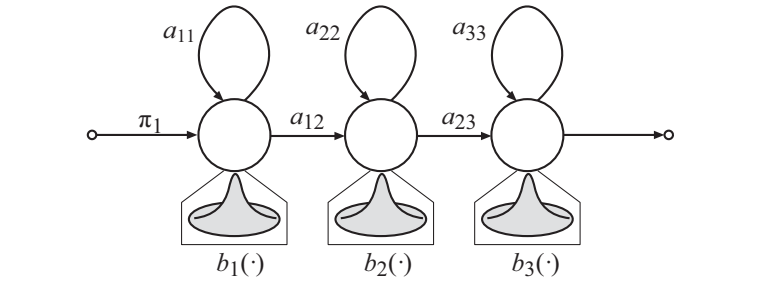
Duration Modelling

Duration is explicitly modelled in HMM-based speech synthesis using single Gaussian distributions [Yoshimura et al., 1998]. The dimensionality of a multivariate state duration distribution (when using only one stream) or the number of streams (when using a univariate state duration distribution per stream) is equal to the number of emitting states of an HMM, and the n-th dimension or stream corresponds to the n-th emitting state. Explicit duration modelling is straightforward since the length of phonemes needs to be determined at the synthesis stage, which is mainly for the purpose of simplifying and speeding up the process of speech parameter generation (see Section 2.3.3 for details). Furthermore, the speaking rate and duration patterns of synthesized speech can be easily adjusted by explicit duration modelling, which helps to achieve voice diversity.

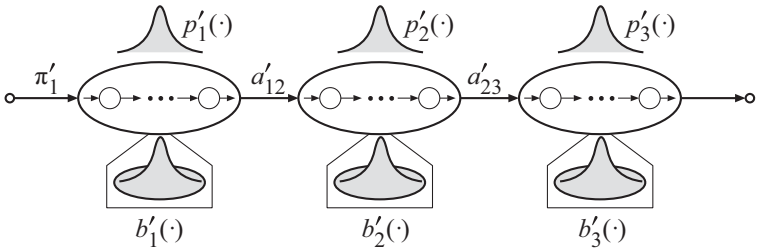
There exists an inconsistency in the conventional HMM-based speech synthesis framework

3. i.e. a sum of a pulse train with phase manipulation and white noise weighted by band aperiodicity in the frequency domain

– although speech parameters are generated using both HMMs and explicit state duration distributions in the synthesis stage, these HMMs and state duration distributions are not updated simultaneously in the training stage. The state duration distributions are actually estimated by using state occupancy probabilities obtained in the last iteration of embedded re-estimation of HMM parameters [Yoshimura et al., 1998, 1999] and then clustered by decision tree-based context clustering. In order to solve this inconsistency, the hidden semi-Markov model (HSMM) was introduced into speech synthesis [Zen et al., 2004]. The difference between an HMM and an HSMM is illustrated in Figure 2.7. State distributions of spectrum, pitch and duration can be estimated simultaneously in the training stage in HSMM-based speech synthesis, where state distributions of duration play the role of state transition matrices of HMMs. It was reported that the utilization of HSMMs could improve the naturalness of synthesized speech [Zen et al., 2004].



(a) hidden Markov model (HMM)



(b) hidden semi-Markov model (HSMM)

Figure 2.7 – Examples of an HMM and an HSMM (three emitting states, left to right, and without any skip) [Zen et al., 2004]. $p'_i(\cdot)$ indicates a state duration distribution.

The thesis work was not focused on the fundamentals of statistical parametric modelling for speech synthesis, so in this thesis *HMM-based speech synthesis* is considered a generic term that includes the HSMM-based framework.

Context-Dependent Synthesis Models

Tens of different contextual factors are employed in context-dependent labels for speech synthesis, including phonetic contexts around a base phoneme (e.g., its left and right neighbouring phonemes) and many prosodic contexts (e.g., stress, tone, part of speech, the position of a phoneme/syllable/word/phrase in the current syllable/word/phrase/utterance, the length

of the current syllable/word/phrase/utterance, etc) [Tokuda et al., 2002b]. This is because natural speech varies to a great extent. A phoneme can be uttered very differently in different situations, for example, stressed or not, at the beginning or the end of a sense group, at the end of an interrogative or a declarative sentence, etc. Therefore very long context-dependent labels denoting extremely specific phoneme variants are required to capture subtle acoustic variations for synthesizing natural-sounding speech. While not so critical for spectrum modelling, these contextual factors are essential for appropriate modelling of prosody (in particular, F_0 and duration).

It is apparent that HMM-based speech synthesis faces the problem of severe sparsity of training data due to the large number of contexts. This problem is resolved by decision tree-based clustering as described in Section 2.1.3.

2.3.2 Building Voice Models for HMM-Based Speech Synthesis

We can collect speech data from a particular speaker and build a speaker-dependent synthesizer by training HMMs on his speech data alone. This is not very difficult due to mature techniques: (1) HMM parameters can be estimated by the Baum-Welch algorithm, as discussed in Section 2.1.2; (2) very specific acoustic variations of phonemes can be captured by a huge amount of context-dependent models, as discussed in Section 2.3.1; and (3) the problem of the severe sparsity of training data can be handled by decision tree-based clustering, as discussed in Section 2.1.3. The main hurdle to training a set of speaker-dependent models is collection of plenty of speech data from a target speaker, which makes speaker-dependent modelling not always preferred and not even feasible when voice diversity is required. So we move on to discussing the average voice synthesis paradigm.

Average Voice Synthesis Models

Although in theory speaker adaptation techniques can be applied to synthesis models trained on speech data of any number of speakers, speaker adaptation performance is degraded when there is sharp distinction in terms of voice characteristics or phonetic/prosodic patterns between the "average" of training speakers and the target speaker [Yamagishi et al., 2010a]. So adapting speaker-dependent models is not appropriate in general. It is very likely that the voice and phonetic/prosodic patterns of target speakers do not match those of a set of speaker-dependent models.

In order to build synthesis models which suit as many target speakers as possible and thus to obtain better adaptation performance, the average voice synthesis paradigm was proposed in [Yamagishi, 2006, Yamagishi and Kobayashi, 2007]. An average voice can be regarded as an artificial voice trained on speech data collected from tens or hundreds of real speakers, by means of shared-decision-tree-based context clustering [Yamagishi, 2006, Chapter 4] and speaker adaptive training as discussed in Section 2.2.4. Shared-decision-tree-based context clustering

ensures average voice model distributions in every leaf node are derived from speech data of all the training speakers, i.e, to guarantee the speaker-independence of each synthesis model. Speaker adaptive training normalizes speech features of a diversity of training speakers such that ideally speaker-specific characteristics are extracted by adaptation transforms and average voice synthesis models capture only common phonetic and prosodic variations across training speakers.

Average voice synthesis models are more adaptable to various target speakers [Yamagishi et al., 2010b]. Firstly, they are not biased towards any type of target speaker. Secondly, they are trained over a huge quantity of normalized speech features, thereby covering much more phonetic and prosodic variations of a spoken language.

It has been demonstrated that average voice synthesis models can be trained on speech corpora designed for speech recognition like WSJ0 [Paul and Baker, 1992] and SPEECON [Iskra et al., 2002], so training data collection is not a major issue and we just need to collect a small amount of adaptation data from target speakers to achieve voice diversity by speaker adaptation [Yamagishi et al., 2010a].

2.3.3 Synthesis

The task of the synthesis stage, mathematically, means generating a speech feature sequence $\boldsymbol{O}^* = \left(\boldsymbol{o}_1^*, \boldsymbol{o}_2^*, \cdots, \boldsymbol{o}_T^*\right)$ from parameters λ of a particular HMM sequence on condition that $\boldsymbol{O}^*$ maximizes the probability $p(\boldsymbol{O}|\lambda)$:

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} p(\boldsymbol{O}|\lambda) \tag{2.41}$$

$$= \arg\max_{\boldsymbol{O}} \sum_{\text{all } \boldsymbol{Q}} p(\boldsymbol{O}|\boldsymbol{Q},\lambda)\, p(\boldsymbol{Q}|\lambda) \tag{2.42}$$

$$\approx \arg\max_{\boldsymbol{O}} p(\boldsymbol{O}|\boldsymbol{Q}^*,\lambda)\, p(\boldsymbol{Q}^*|\lambda) \tag{2.43}$$

There is no known closed solution to Eq. (2.41), though it can be solved by the EM algorithm [Tokuda et al., 2000]. In practice, as an approximation, this task is divided into two steps on the basis of Eq. (2.42): firstly, an optimal state sequence $\boldsymbol{Q}^*$ is determined by maximizing $p(\boldsymbol{Q}|\lambda)$ with respect to $\boldsymbol{Q}$; secondly, the optimal speech feature sequence $\boldsymbol{O}^*$ is generated by maximizing $p(\boldsymbol{O}|\boldsymbol{Q}^*,\lambda)$ with respect to $\boldsymbol{O}$.

If speech feature vectors were independent and identically distributed, we could model only "static" features (i.e., those extracted from speech waveforms directly) and then simply concatenating state mean vectors in λ would produce the desired $\boldsymbol{O}^*$ for $p(\boldsymbol{O}|\boldsymbol{Q}^*,\lambda)$. Obviously, this results in a sudden change of speech features at every state boundary and there would be audible discontinuities. Thus dynamic features are included in speech feature vectors of training data for producing smooth speech feature trajectories [Tokuda et al., 1995a,b]. Given the explicit relationship between static and dynamic features, the parameter generation

algorithm proposed in [Tokuda et al., 1995a,b] enables inference of observations that involves both static and dynamic statistics including covariance matrices. The following describe how to obtain the desired $\boldsymbol{O}^*$ for $p(\boldsymbol{O}|\boldsymbol{Q}^*, \lambda)$ given dynamic features.

Let $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{c}_1^\top & \boldsymbol{c}_2^\top & \cdots & \boldsymbol{c}_T^\top \end{bmatrix}^\top$ be the "static part" of speech features $\boldsymbol{O} = \begin{bmatrix} \boldsymbol{o}_1^\top & \boldsymbol{o}_2^\top & \cdots & \boldsymbol{o}_T^\top \end{bmatrix}^\top$, i.e.,

$$\boldsymbol{o}_t = \begin{bmatrix} \boldsymbol{c}_t^\top & \Delta \boldsymbol{c}_t^\top & \Delta^2 \boldsymbol{c}_t^\top \end{bmatrix}^\top, \qquad t = 1, 2, \cdots, T. \tag{2.44}$$

The first- and second-order dynamic features $\Delta \boldsymbol{c}_t$ and $\Delta^2 \boldsymbol{c}_t$ there are defined by

$$\Delta \boldsymbol{c}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) \boldsymbol{c}_{t+\tau}, \qquad t = 1, 2, \cdots, T, \tag{2.45}$$

$$\Delta^2 \boldsymbol{c}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) \boldsymbol{c}_{t+\tau}, \qquad t = 1, 2, \cdots, T. \tag{2.46}$$

Suppose D is the dimensionality of $\boldsymbol{c}_t$. The relation between $\boldsymbol{O}$ and $\boldsymbol{C}$ can be expressed by

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C}, \tag{2.47}$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1^\top \\ \boldsymbol{W}_2^\top \\ \vdots \\ \boldsymbol{W}_T^\top \end{bmatrix}, \tag{2.48}$$

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{W}_t^{(0)} & \boldsymbol{W}_t^{(1)} & \boldsymbol{W}_t^{(2)} \end{bmatrix}, \qquad t = 1, 2, \cdots, T, \tag{2.49}$$

$$\begin{aligned} \boldsymbol{W}_t^{(n)} = \Big[& \underbrace{\boldsymbol{0}_{D\times D}}_{1\text{st}} \quad \cdots \quad \boldsymbol{0}_{D\times D} \quad \underbrace{w^{(n)}\left(-L_-^{(n)}\right) \cdot \boldsymbol{I}_{D\times D}}_{(t-L_-^{(n)})\text{-th}} \quad \cdots \quad \underbrace{w^{(n)}(0) \cdot \boldsymbol{I}_{D\times D}}_{t\text{-th}} \quad \cdots \\ & \underbrace{w^{(n)}\left(L_+^{(n)}\right) \cdot \boldsymbol{I}_{D\times D}}_{(t+L_+^{(n)})\text{-th}} \quad \boldsymbol{0}_{D\times D} \quad \cdots \quad \underbrace{\boldsymbol{0}_{D\times D}}_{T-\text{th}} \Big]^\top, \qquad n = 0, 1, 2, \end{aligned} \tag{2.50}$$

where $\boldsymbol{I}_{D\times D}$ and $\boldsymbol{0}_{D\times D}$ are D-by-D identity and zero matrices, respectively. By solving

$$\frac{\partial p(\boldsymbol{W}\boldsymbol{C}|\boldsymbol{Q}^*, \lambda)}{\partial \boldsymbol{C}} = \boldsymbol{0}, \tag{2.51}$$

we can obtain the desired static speech feature sequence $\boldsymbol{C}$ for waveform generation:

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}, \tag{2.52}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{q_1^*}^\top & \boldsymbol{\mu}_{q_2^*}^\top & \cdots & \boldsymbol{\mu}_{q_T^*}^\top \end{bmatrix}^\top, \tag{2.53}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{q_1^*} & & & \\ & \boldsymbol{\Sigma}_{q_2^*} & & \\ & & \ddots & \\ & & & \boldsymbol{\Sigma}_{q_T^*} \end{bmatrix}, \tag{2.54}$$

where $\boldsymbol{\mu}_{q_t^*}$ and $\boldsymbol{\Sigma}_{q_t^*}$ represent the mean vector and covariance matrix of state q_t^* respectively. Finally, the STRAIGHT vocoder can be employed to convert the static speech feature sequence $\boldsymbol{C}$ into a waveform.

A major problem with the above-mentioned algorithm is that due to the statistical processing, generated trajectories of speech features are often excessively smoothed and thus lead to muffled synthesized speech. In order to alleviate this over-smoothing effect, a new algorithm that also takes the output probability of global variances (GV) of generated trajectories into consideration was proposed [Toda and Tokuda, 2005]. The variance vector calculated over all the *static* speech feature frames of an utterance is defined as the global variance of this utterance.

In the training/adaptation stage, global variances of training/adaptation utterances are modelled by a single Gaussian distribution with parameters λ_{GV} (λ_{GV} and λ are estimated independently). Considering both λ_{GV} and λ in the synthesis stage, the optimal static speech feature sequence $\boldsymbol{C}^*$ is generated by maximizing the product $p(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda)^{\omega} \cdot p\left(\mathbf{gv}(\boldsymbol{C})|\lambda_{\mathrm{GV}}\right)$ with respect to $\boldsymbol{C}$ instead (ω is a constant weight). It was reported that the utilization of global variances significantly improved the naturalness of synthesized speech [Toda and Tokuda, 2005]. In addition, it is possible to model global variances in a context-dependent fashion [Oura et al., 2009].

2.3.4 Subjective Evaluation

The output of speech recognition is plain text or phoneme transcriptions. A computer can efficiently and precisely assess a speech recognizer by calculating the word or phoneme error rate of this textual output using reference plain text or phoneme transcriptions. By contrast, as the output of speech synthesis is sound, dependable assessment of a speech synthesizer has to rely on people's perception of this acoustic output.

The dependency leads to a couple of problems. Firstly, judgements from a listener (even if he is an expert) could be unintentionally biased, due to his hearing, the quality of earphones or headphones he uses, the extent of quietness of the environment and even his mood when he listens to the acoustic output [Black and Tokuda, 2005]. Secondly, judgements on merely a few acoustic output samples are not representative of the overall performance of the speech synthesizer that generates these acoustic samples. As a result, the most reliable approach to evaluating a speech synthesizer is to obtain judgements from a sizable group of people listening to a large number of acoustic samples generated by the speech synthesizer and then scoring them according to certain criteria. After that, the performance of the speech

synthesizer is typically presented by the average and a confidence interval of all these scores or by a box plot.

The crucial objective of speech synthesis is to generate speech which sounds as natural as if it is uttered by a real person. This not only determines how acceptable/pleasant a speech synthesizer is to human ears, but also impacts upon the intelligibility of synthesized speech. Naturalness of synthesized speech is thus considered one of the key measures of the performance of a speech synthesizer. On top of that, it may be desirable that the voice characteristics of a particular person can be fully reproduced in synthesized speech, thereby bringing voice diversity for speech synthesizers in order to make them more favourable to customers or helping to reconstruct the lost voice of a disabled person. Hence, speaker similarity between a synthetic voice and the original target voice functions as another important measure of the performance of a speech synthesizer.

Naturalness

Naturalness of synthesized speech is typically evaluated in the form of an AB test or an MOS (mean opinion score) test. In an AB test, listening test subjects are presented with pairs of speech samples first, where one is generated by the speech synthesizer being evaluated and the other is generated by a baseline. Then they choose one sample which they think sounds more natural from each pair. In an MOS test, a subject listens to only a single speech sample generated by either the speech synthesizer being evaluated or a baseline. Then he is required to score the sample on a 5-point scale ranging from 1 to 5, where 1 means “completely unnatural” and 5 means “completely natural” [Fraser and King, 2007].

A problem with AB tests is that it only permits pairwise comparison. Thus when comparing multiple synthesis systems, it requires many tests to be run.

Speaker Similarity

Likewise, speaker similarity between a synthetic voice and an original target voice is typically evaluated in two similar forms: an ABX test or a DMOS (differential mean opinion score) test. In an ABX test, a listener is presented with an original recording in a target voice and then a pair of speech samples generated by the speech synthesizer being evaluated and a baseline respectively. After that, he is required to choose from the pair one speech sample which he thinks has a closer voice identity to the target voice. In a DMOS test, a listener is presented with an original recording in a target voice and then a speech sample generated by either the speech synthesizer being evaluated or a baseline. After that, he is required to score the synthesized sample on a 5-point scale ranging from 1 to 5, where 1 means “sounds like a totally different person” and 5 means “sounds like exactly the same person” [Fraser and King, 2007].

A potential problem with speaker similarity evaluation is that listeners may not be always immune to speech quality in an ABX test. They may subconsciously choose the sample with

better naturalness in a pair, especially when the two synthetic voices sound almost equally similar to the reference speaker.

In addition, an ABX test only permits pairwise comparison too. It has the same problem as an AB test when there are multiple systems to compare.

Intelligibility

Intelligibility is usually evaluated by inviting listeners to transcribe semantically unpredictable sentences and then calculating the word error rate of their transcriptions. Semantically unpredictable sentences can prevent listeners from guessing a few missing words based on semantic context in order to ensure the only factor influencing the result is the intelligibility of the speech itself.

Intelligibility is not evaluated in the thesis work as it is not an objective of the research nor is it significantly impacted. The HMM-based speech synthesis framework has been shown to provide good intelligibility [Hashimoto et al., 2011].

2.3.5 Objective Evaluation

Subjective evaluation of a speech synthesizer that relies on human perception, as described above, requires a lot of effort and is considerably time-consuming and costly. Furthermore, human perception is not always sufficiently sensitive, meaning that it is very difficult for listeners to make trustworthy judgements when improvement and degradation are subtle. Using merely subjective evaluation could hinder or even obstruct the progress of research. Therefore, several objective metrics are also employed in synthesis evaluation. Objective metrics can accelerate research, indicate small changes resulting from the utilization of new algorithms or experimental settings and reveal promising research directions. Since in general they only correlate with human perception loosely [Gray Jr. and Markel, 1976, Barnwell III, 1980, Yamagishi et al., 2010a], objective metrics should be employed with care when drawing conclusions based on them alone.

Basically, an objective metric is a “distance” between a synthesized utterance and its corresponding original recording. Owing to the parametric nature of HMM-based speech synthesis, objective metrics can be calculated easily over the speech features of the synthesized utterance and the original recording. For the sake of convenience and meaningful comparison, the synthesized utterance to be assessed is normally generated using time-aligned durations from the original recording (except when evaluating duration prediction) – In this way, frame-by-frame calculation of objective metrics can be easily conducted and it can be assumed that two aligned frames are produced by the same context-dependent phone.

The source-filter model is employed in the HMM-based speech synthesis framework, so the parametric output of an HMM-based speech synthesizer contains spectral and excitation

feature trajectories. Spectral distortion, the voicing error rate, RMSE and correlation coefficient of F_0 are typically used in order to measure the "distance" between generated spectra and F_0 contours and those of a corresponding original recording.

Mel-Cepstral Distortion

Mel-cepstral distortion (MCD) [Kubichek, 1993] can be viewed as approximate logarithmic spectral distance. As mel-cepstral distortion decreases, voice quality could be found to be perceptually better [Toda et al., 2004]. Mel-cepstral distortion is used in this thesis because mel-cepstrum (MCEP) was the only spectral feature considered in the entire thesis work.

Suppose $\boldsymbol{c}_t^{\text{ref}} = \begin{bmatrix} c_{t,1}^{\text{ref}} & c_{t,2}^{\text{ref}} & \cdots & c_{t,D}^{\text{ref}} \end{bmatrix}^\top$ is the frame of mel-cepstrum of dimensionality D at time t from an original recording and $\boldsymbol{c}_t^{\text{syn}} = \begin{bmatrix} c_{t,1}^{\text{syn}} & c_{t,2}^{\text{syn}} & \cdots & c_{t,D}^{\text{syn}} \end{bmatrix}^\top$ is the corresponding frame from a synthesized utterance. Mel-cepstral distortion at the frame level is given by [Kominek et al., 2008, Mashimo et al., 2001]

$$\text{MCD}_\text{f}\left(\boldsymbol{c}_t^{\text{ref}}, \boldsymbol{c}_t^{\text{syn}}\right) = \frac{10\sqrt{2}}{\ln 10} \cdot \sqrt{\sum_{d=1}^{D} \left(c_{t,d}^{\text{ref}} - c_{t,d}^{\text{syn}}\right)^2} \qquad \text{(dB)}, \tag{2.55}$$

and mel-cepstral distortion at the utterance level is given by

$$\text{MCD}\left(\boldsymbol{c}^{\text{ref}}, \boldsymbol{c}^{\text{syn}}\right) = \frac{1}{T} \cdot \sum_{t=1}^{T} \text{MCD}_\text{f}\left(\boldsymbol{c}_t^{\text{ref}}, \boldsymbol{c}_t^{\text{syn}}\right) \qquad \text{(dB)}, \tag{2.56}$$

where T is the total number of frames in the original recording.

Voicing Error Rate, RMSE and Correlation Coefficient of F_0

Suppose the F_0 contours of an original recording and a corresponding synthesized utterance are $\boldsymbol{f}^{\text{ref}} = \begin{bmatrix} f_1^{\text{ref}} & f_2^{\text{ref}} & \cdots & f_T^{\text{ref}} \end{bmatrix}^\top$ and $\boldsymbol{f}^{\text{syn}} = \begin{bmatrix} f_1^{\text{syn}} & f_2^{\text{syn}} & \cdots & f_T^{\text{syn}} \end{bmatrix}^\top$, respectively. An F_0 contour is intrinsically composed of segments with and without F_0 values. It does not make much sense to calculate any distance when f_t^{ref} has a value but f_t^{syn} does not, and vice versa. As a result, two obvious objective metrics are the percentage of voiced-to-unvoiced (V2Uv) and unvoiced-to-voiced (Uv2U) errors in a synthesized utterance.

The V2Uv and Uv2V error rates of $\boldsymbol{f}^{\text{syn}}$ can be calculated as follows:

$$\text{V2Uv}\left(\boldsymbol{f}^{\text{ref}}, \boldsymbol{f}^{\text{syn}}\right) = \frac{1}{T} \cdot \left(\sum_{t=1,\ f_t^{\text{ref}} \text{ has a value},\ f_t^{\text{syn}} \text{ does not}}^{T} 1 \right) \times 100\%, \tag{2.57}$$

$$\mathrm{Uv2V}\left(\boldsymbol{f}^{\mathrm{ref}},\boldsymbol{f}^{\mathrm{syn}}\right)=\frac{1}{T}\cdot\left(\sum_{t=1,\ f_t^{\mathrm{ref}}\text{ does not have a value, }f_t^{\mathrm{syn}}\text{ does}}^{T}1\right)\times 100\%. \tag{2.58}$$

Then clearly there is always no distortion when neither f_t^{ref} nor f_t^{syn} has a value. Only the aligned F_0 frames that are voiced in both the original recording and synthesized utterance are taken into account for other objective metrics, i.e., root-mean-square error (RMSE) and correlation coefficient. Namely, only the F_0 frames at time t that belongs to

$$\mathbb{T}_{\mathrm{V}}=\left\{t\,\middle|\,t=1,2,\cdots,T;\ \text{Both } f_t^{\mathrm{ref}} \text{ and } f_t^{\mathrm{syn}} \text{ have values.}\right\} \tag{2.59}$$

are used for the calculation of RMSE between $\boldsymbol{f}^{\mathrm{ref}}$ and $\boldsymbol{f}^{\mathrm{syn}}$ according to

$$\mathrm{RMSE}\left(\boldsymbol{f}^{\mathrm{ref}},\boldsymbol{f}^{\mathrm{syn}}\right)=\sqrt{\frac{1}{V}\cdot\sum_{t\in\mathbb{T}_{\mathrm{V}}}\left(f_t^{\mathrm{ref}}-f_t^{\mathrm{syn}}\right)^2}\qquad(\mathrm{Hz}), \tag{2.60}$$

and the calculation of the correlation coefficient between $\boldsymbol{f}^{\mathrm{ref}}$ and $\boldsymbol{f}^{\mathrm{syn}}$ according to

$$\mathrm{CorrCoef}\left(\boldsymbol{f}^{\mathrm{ref}},\boldsymbol{f}^{\mathrm{syn}}\right)=\frac{|\mathbb{T}_{\mathrm{V}}|\sum_{t\in\mathbb{T}_{\mathrm{V}}}f_t^{\mathrm{ref}}f_t^{\mathrm{syn}}-\sum_{t\in\mathbb{T}_{\mathrm{V}}}f_t^{\mathrm{ref}}\sum_{t\in\mathbb{T}_{\mathrm{V}}}f_t^{\mathrm{syn}}}{\sqrt{|\mathbb{T}_{\mathrm{V}}|\sum_{t\in\mathbb{T}_{\mathrm{V}}}\left(f_t^{\mathrm{ref}}\right)^2-\left(\sum_{t\in\mathbb{T}_{\mathrm{V}}}f_t^{\mathrm{ref}}\right)^2}\cdot\sqrt{|\mathbb{T}_{\mathrm{V}}|\sum_{t\in\mathbb{T}_{\mathrm{V}}}\left(f_t^{\mathrm{syn}}\right)^2-\left(\sum_{t\in\mathbb{T}_{\mathrm{V}}}f_t^{\mathrm{syn}}\right)^2}}, \tag{2.61}$$

where $|\mathbb{T}_{\mathrm{V}}|$ is the number of elements in $\mathbb{T}_{\mathrm{V}}$. RMSE reflects the microscopic, numerical distortion of $\boldsymbol{f}^{\mathrm{syn}}$ while the correlation coefficient between $\boldsymbol{f}^{\mathrm{ref}}$ and $\boldsymbol{f}^{\mathrm{syn}}$ suggests their macroscopic, geometric similarity.

2.4 Summary

In this chapter we revisited in brief the basics of hidden Markov models, context-dependent modelling, the widely used maximum likelihood linear transformation framework for speaker adaptation and speaker adaptive training, and finally the training, synthesis and evaluation stages of HMM-based speech synthesis. The contents of this chapter serves as a foundation for the entire subsequent research work.

HMM-based speech synthesis provides a nearly language-independent solution to building a speech synthesizer. The only component that is strongly tied to language is the text analyzer and the questions set for decision tree-based state tying. This makes the HMM-based speech synthesis framework particularly well suited to multilingual and cross-lingual speech processing.

Given that HMMs can provide a common foundation for speech recognition and speech synthesis, there is hope to develop a unified framework (e.g., versatile models and features) that may operate for both speech recognition and synthesis and could be particularly useful for personalization of speech-to-speech translation.

3 Cross-Lingual Speaker Adaptation for Speech Synthesis

3.1 Multilingual Speech Processing

HMM-based speech synthesis, visited in the previous chapter, has developed into a mature technology for building monolingual speaker-dependent voices. Unfortunately, application scenarios in real life are not always as simplistic as that. Nowadays, multilingual speech processing, which refers to technology that supports spoken input and output in a large variety of languages at the same time [Schultz and Kirchhoff, 2006, Chapter 1], has caught much interest from the research community. It is hoped that a single "language-independent" system can be developed to handle multiple spoken languages seamlessly.

On the one hand, research on multilingual speech processing is motivated by the increasingly common code-switching phenomenon. Code-switching refers to when people switch between languages while speaking, thereby even a single sentence can contain more than one language. Although a collection of monolingual systems can be effectively employed as a single multilingual system, such a simple combination has problems in tackling transitions from one language to another. In the case of multilingual speech recognition, the performance of such a combined system depends on the accuracy of a language identification module. Therefore the difficulty is to build a highly reliable language identification system in addition to that of the speech recognition itself. In the case of multilingual speech synthesis, it is not trivial to synthesize code-switched sentences naturally because segmental and supra-segmental consistency needs to be maintained around language boundaries, and the voice characteristics should also remain identical across language boundaries.

On the other hand, training data collection can pose a problem for some languages. In particular, there may not exist sufficient training data in languages that are not widely spoken in the world. Training a monolingual system of high quality directly in such an under-resourced language is thus infeasible. Nonetheless, it is possible that a good system in an under-resourced language can be built by bootstrapping from a large amount of training data in other languages that as a whole can more or less cover the acoustic space of the under-resourced one [Vu et al., 2011, Imseng et al., 2012]. Research on multilingual speech processing is also motivated by

this possibility.

The HMM-based framework provides a promising direction to multilingual speech processing and in fact research on multilingual modelling has been conducted for many years. The focus of multilingual modelling is on how to share data and acoustic models among languages, in order that such models may be also applied to languages not seen in the training data. Köhler studied the possibility of creating "multilingual" phoneme models which could be used in a variety of languages by exploiting acoustic-phonetic similarities of sounds [Köhler, 1996]. Byrne et al. proposed to train acoustic models over training data in English, Spanish, Russian and Mandarin Chinese. A recognizer in the Czech language could then be built directly with these acoustic models as well as phoneme mapping rules, and such a speech recognizer could be enhanced by adaptation techniques if a certain amount of data in the Czech language was available [Byrne et al., 2000]. More importantly, they found that even models in training languages that performed poorly when used individually could contribute to the overall combination. Lin et al. explored shared structures embedded in a large collection of speech data spanning a number of spoken languages in order to establish a common set of universal phone models that could be used for large vocabulary speech recognition of all the languages either seen or unseen during training [Lin et al., 2009]. Schultz and Waibel investigated different methods of building multilingual recognition models: through a simple collection of monolingual models, sharing model distributions and Gaussian mixture component weights across languages, or sharing model distributions across languages [Schultz and Waibel, 2001].

Similarly, efforts have been made to develop multilingual models for speech synthesis. Latorre attempted to build a multilingual synthesizer over speech data from multiple speakers and in multiple languages by means of IPA-based phoneme sharing [Latorre, 2006]. He split diphthongs into two in order to facilitate phoneme sharing across training languages. Qian et al. proposed to share HMM state distributions across Mandarin Chinese and English by using language-independent questions for clustering so as to build a bilingual speech synthesizer capable of producing smoother transition at language boundaries [Qian et al., 2009]. In the multilingual synthesis system that Zen et al. developed, sharing happened at the sub-state level: covariance matrices and mean vectors were shared separately across training languages [Zen et al., 2012].

3.2 From "Multilingual" to "Cross-Lingual"

"Multilingual" stresses the ability of a single system to handle more than one language while "cross-lingual" stresses the possibility of transferring some characteristics (e.g. speakers' voices, recording environments, etc) from a language to another. The difference in meaning between "multilingual" and "cross-lingual" seems vague because they are intertwined. For example, Byrne's above-mentioned work can be regarded as both multilingual and cross-lingual [Byrne et al., 2000], as in this case multilingual modelling provided a solution to a cross-lingual

problem of generating acoustic models for a new language using little or no in-language training data.

Recently the research topic of transferring speaker characteristics from one language into another has attracted a great deal of attention. This is an essential technique for personalized speech-to-speech translation. An unavoidable issue in this research topic is that the acoustic spaces of the two languages do not completely overlap because their phoneme inventories and prosodic patterns are normally distinct. The fact that the acoustic space, phoneme inventory, prosodic patterns, articulatory features and so forth of a language partially overlap those of another language is referred to as "language mismatch" in this thesis.

The challenge in transferring speaker characteristics from one language to another is different from that in Byrne's and related work. In this, they faced mismatch between the language to be recognized and the ones in the training data. As for transferring speaker characteristics across languages, the voice characteristics need to be transferred from speech in some language to speech in a different language without inadvertently capturing other language-dependent characteristics, i.e., the language characteristics of models need to remain untouched.

Transferring speaker characteristics can be handled by speaker adaptation techniques described in the previous chapter. The unique challenge is discovering how such techniques can be applied in a cross-lingual fashion, ideally with the same efficiency as equivalent intra-lingual approaches. This challenge is a direct consequence of the fact that state-of-the-art speaker adaptation techniques cannot automatically identify and then single out speaker characteristics. Given the language mismatch between models and adaptation data, not only the voice characteristics but also the language characteristics of models may be adapted towards those in adaptation data.

This thesis is focused on the investigation of cross-lingual speaker adaptation for speech synthesis, more specifically, the influence of language mismatch and how to alleviate this influence. We begin with preparatory issues like state-of-the-art cross-lingual speaker adaptation approaches, model and data preparation and cross-lingual speaker similarity judgement.

It has been noted that researchers have used different terms for the same concepts in cross-lingual speech processing. For example, we have seen at least four terms which refer to the language of adaptation data: input language, source language, adaptation language and first language. For the sake of clarity, the terms in Table 3.1 are adopted throughout this thesis.

3.3 State-of-the-Art Approaches to Cross-Lingual Speaker Adaptation

Unlike intra-lingual speaker adaptation, cross-lingual speaker adaptation adapts the voice characteristics of average voice synthesis models in an output language into those of a target speaker who has provided adaptation data in an input language ($L_{\text{in}} \neq L_{\text{out}}$). The fact that

Table 3.1 – Key terminology for the research on cross-lingual speaker adaptation

Term	Notation	Definition
input language	L_{in}	the language spoken in adaptation data
output language	L_{out}	the language in which spoken output is synthesized
target speaker	—	the person whose voice characteristics are being adapted
target voice	—	the voice of a target speaker

$L_{\text{in}} \neq L_{\text{out}}$ prevents us from directly maximizing the likelihood of synthesis models for the target speaker in the output language. In other words, there is no straightforward way of computing the likelihood in Eq. (3.1) (see Eq. (2.34) for the contrast) though it is possible to obtain a likelihood in practice:

$$\hat{\boldsymbol{W}}_s = \arg\max_{\boldsymbol{W}_s} \sum_{\text{all } \boldsymbol{Q}_s^{L_{\text{out}}}} p\left(\bar{\boldsymbol{O}}_s^{L_{\text{in}}}, \boldsymbol{Q}_s^{L_{\text{out}}} \middle| \boldsymbol{\mu}^{L_{\text{out}}}, \boldsymbol{\Sigma}^{L_{\text{out}}}\right). \tag{3.1}$$

Hence, the inherent difficulty in cross-lingual speaker adaptation is how to extract speaker characteristics from one language and apply them to another without having access to any direct relationships between phonological representations in the input language and underlying state distributions in the output language. Two types of techniques have been investigated so far. Their common key point is to establish the missing relationships, either explicitly or implicitly.

3.3.1 Phoneme Mapping

A phoneme is the smallest contrastive unit in the sound system of a language. It serves to distinguish between meanings of words in the language. Phoneme mapping across two languages may be the most straightforward approach to cross-lingual speaker adaptation. The relationship between the input and output languages is captured explicitly by phoneme mapping pairs according to knowledge of phonetics [Moberg et al., 2004, Latorre et al., 2006, Wu et al., 2008]. To be specific, two phonemes in two respective languages are regarded as identical if they are represented by the same phonetic notation like the International Phonetic Alphabet (IPA) [International Phonetic Association, 1999]. For instance, the English /g/ as in *garden* and the French /g/ as in *garçon* can constitute a phoneme mapping pair. Apart from this kind of phoneme shared across languages, a phoneme existing in only one language gets mapped to one or several phonemes in the other language that either are perceptually the closest or share the most articulatory features. According to phoneme mapping pairs, adaptation data in the input language can be re-transcribed with phonemes of the output language and thus cross-lingual speaker adaptation can be conducted in the intra-lingual fashion.

The main disadvantage of phoneme mapping is that a phoneme is rather a large unit for

mapping construction. It brings difficulty in finding equivalents in two languages, especially when the phonology of the two languages differs to a great extent – this would result in many inaccurate phoneme mapping rules. For example, mapping between Mandarin and English at the phoneme level cannot provide good speech quality after cross-lingual speaker adaptation [Wu et al., 2008].

Moreover, phonetic notations like IPA do not necessarily imply the same acoustic properties across languages. Essentially, they are merely an abstraction of spoken languages that aims to provide common representation of sounds on the basis of a few coarse, language-independent descriptors such as voicing (for consonants), the place and manner of articulation (for consonants), the tongue and lip positions (for vowels), etc. Therefore, phonemes in different languages sharing the same phonetic notation are not necessarily acoustically identical [1], let alone those that do not share the same notation but are mapped to each other.

3.3.2 Bilingual Modelling

The basic idea of bilingual modelling for cross-lingual speaker adaptation is to train models on a corpus including speech data in both the input and output languages, such that the resultant models capture characteristics of the two languages at the same time. The effectiveness of bilingual modelling (and multilingual modelling in a more general sense) has been demonstrated for both speech recognition [Köhler, 2001, Schultz and Waibel, 2001] and synthesis [Latorre et al., 2005, Qian et al., 2009]. Bilingual modelling establishes relationships between the input and output languages in the form of *shared* models. A shared model means that the model distribution is derived from training data in both the input language and the output language. Ideally, all model parameters should be shared between the input and output languages in the case of bilingual modelling.

Using the bilingual modelling technique, cross-lingual speaker adaptation can be treated in the same manner as intra-lingual speaker adaptation. Eq. (3.1) can be converted into the following one:

$$\hat{\boldsymbol{W}}_s^{L_{\text{in}}\&L_{\text{out}}} = \arg\max_{\boldsymbol{W}_s^{L_{\text{in}}\&L_{\text{out}}}} \sum_{\text{all } \boldsymbol{Q}_s^{L_{\text{in}}\&L_{\text{out}}}} p\left(\bar{\boldsymbol{O}}_s^{L_{\text{in}}}, \boldsymbol{Q}_s^{L_{\text{in}}\&L_{\text{out}}} \middle| \boldsymbol{\mu}^{L_{\text{in}}\&L_{\text{out}}}, \boldsymbol{\Sigma}^{L_{\text{in}}\&L_{\text{out}}}\right). \tag{3.2}$$

Now it is possible to directly relate adaptation data to average voice synthesis models and thus cross-lingual speaker adaptation can be carried out.

Because of the greedy top-down manner of decision tree-based state clustering in the training stage, when the input and output languages are substantially dissimilar, questions high up in a decision tree may split distributions along the language boundary, which effectively prevents any language sharing lower down in the decision tree. As a result, the principal drawback of the bilingual modelling technique is its strong dependency on the phonological/acoustic

1. For example, the French /ɡ/ in *garçon* is actually palatalized, which does not happen to the English /ɡ/ in *garden*.

similarity of the input and output languages, which determines the proportion of shared model distributions across the two languages. The smaller the proportion of shared models, the less meaningful cross-lingual speaker adaptation should be expected. Unfortunately, it is difficult to train a truly bilingual model set in the sense that every model distribution is shared by the input and output languages. For instance, it has been reported that only less than 50% of their HMM state distributions were shared across the input and output languages for Spanish & Japanese in [Latorre et al., 2005], and English & Mandarin in [Qian et al., 2009].

3.3.3 Speaker and Language Factorization

Speaker and language factorization proposed in [Zen et al., 2012] shares the basic idea of bilingual modelling: building synthesis models which includes both input and output languages. In speaker and language factorization, language-specific context-dependencies are handled using cluster adaptive training (CAT) [Gales, 2000] and cluster-dependent decision trees [Zen and Braunschweiler, 2009] while acoustic variations caused by voice characteristics of speakers are captured by another layer, CMLLR transforms [Gales, 1998]. At the synthesis stage, models in a target language to be synthesized are created in the form of linear combination of canonical models trained over speech data in several underlying prototype languages, according to language-specific CAT interpolation weights.

To adapt such a speech synthesis system to a target speaker who can speak one of the training languages, firstly, language-adapted models in this training language are composed using the canonical models and pre-estimated L_{in}-specific CAT interpolation weights. Then speaker-dependent CMLLR transforms are estimated. By using these speaker-dependent transforms, the canonical models and pre-estimated L_{out}-specific CAT interpolation weights, speech in any training language can be synthesized in the target speaker's voice. The phonological relationship between the input and output languages is captured by the common set of canonical models trained over speech data in underlying prototype languages and language-dependent CAT interpolation weights.

3.3.4 State Mapping

HMM state mapping across different languages is a similar technique to phoneme mapping. This approach 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). It was introduced into cross-lingual speech synthesis by Qian et al. [Qian et al., 2009]. 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. HMM state mapping works like a function $\mathscr{M}_{L_A \mapsto L_B}(X) = Y$, which captures the relationships between the input and output languages 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 [Yamagishi and Kobayashi, 2007, Yamagishi, 2006], which are speaker-independent, are employed in construction of state mapping rules.

The Kullback-Leibler divergence [Kullback and Leibler, 1951] is typically used as the similarity measure of state emission pdfs during state mapping construction. Given two continuous probability density functions $f(x)$ and $g(x)$, the K-L divergence from $f(x)$ to $g(x)$ is defined as

$$D_{\mathrm{KL}}\big(f(x)||g(x)\big) = \int f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x. \tag{3.3}$$

This original definition is asymmetrical, i.e., $D_{\mathrm{KL}}\big(f(x)||g(x)\big) \neq D_{\mathrm{KL}}\big(g(x)||f(x)\big)$. The symmetrical form of the K-L divergence between $f(x)$ and $g(x)$ is often used, which is

$$\begin{aligned} D_{\mathrm{KL}}\big(f(x),g(x)\big) &= D_{\mathrm{KL}}\big(f(x)||g(x)\big) + D_{\mathrm{KL}}\big(g(x)||f(x)\big) && (3.4)\\ &= \int f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x + \int g(x)\log\frac{g(x)}{f(x)}\,\mathrm{d}x. && (3.5)\end{aligned}$$

If both $f(x)$ and $g(x)$ are Gaussian distributions, there is a closed solution for Eq. (3.5) [Myrvoll and Soong, 2003]:

$$D_{\mathrm{KL}}\big(f(x),g(x)\big) = \frac{\Big(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\Big)^{\top}\Big(\boldsymbol{\Sigma}_f^{-1}+\boldsymbol{\Sigma}_g^{-1}\Big)\Big(\boldsymbol{\mu}_f-\boldsymbol{\mu}_g\Big)}{2} + \frac{\mathrm{tr}\Big(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\Sigma}_f+\boldsymbol{\Sigma}_f^{-1}\boldsymbol{\Sigma}_g\Big)}{2} - N, \tag{3.6}$$

where N is the dimensionality of the random variable x and the function $\mathrm{tr}(\cdot)$ calculates the trace of a matrix.

Wu et al. proposed two manners for utilizing HMM state mapping rules in [Wu et al., 2009]. The *data mapping* manner functions as follows: (1) to apply state mapping rules between the input and output languages to adaptation data such that the adaptation data in the input language is represented as a state sequence in the output language; (2) given the correspondence between the adaptation data and state distributions in the output language, to carry out "intra-lingual" speaker adaptation on the side of the output language. The key point of the above description is visualized in Figure 3.1.

Eq. (3.1) can be converted into

$$\hat{\boldsymbol{W}}_s^{L_{\mathrm{out}}} = \arg\max_{\boldsymbol{W}_s^{L_{\mathrm{out}}}} \sum_{\text{all } \boldsymbol{Q}_s^{L_{\mathrm{out}}}} p\left(\bar{\boldsymbol{O}}_s^{L_{\mathrm{in}}}, \mathscr{M}_{L_{\mathrm{in}}\mapsto L_{\mathrm{out}}}\left(\boldsymbol{Q}_s^{L_{\mathrm{in}}}\right)\middle|\boldsymbol{\mu}^{L_{\mathrm{out}}},\boldsymbol{\Sigma}^{L_{\mathrm{out}}}\right) \tag{3.7}$$

for the data mapping manner. As reported by [Wu et al., 2009], 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 [Wu et al., 2009], conventional intra-lingual

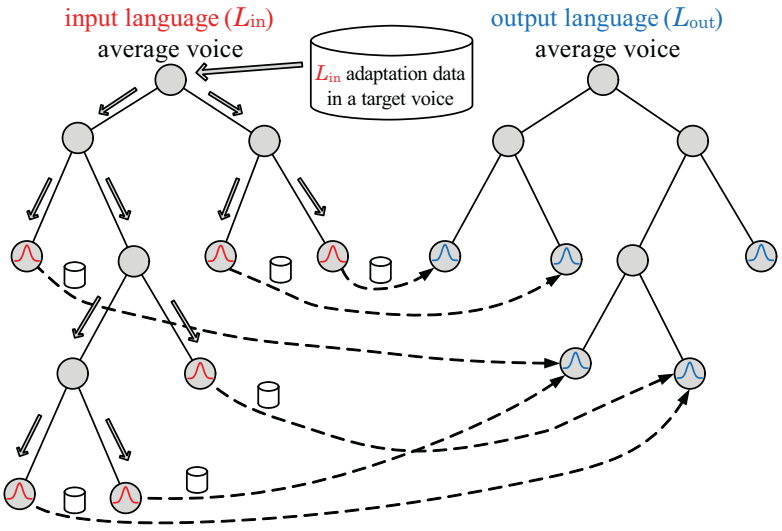


Figure 3.1 – Data mapping manner for cross-lingual speaker adaptation. Small cylinders denote adaptation data segments that are moving from the input language to the output language.

speaker adaptation on the side of the input language is performed first as described below:

$$\hat{\boldsymbol{W}}_s^{L_{\text{in}}} = \arg\max_{\boldsymbol{W}_s^{L_{\text{in}}}} \sum_{\text{all } \boldsymbol{Q}_s^{L_{\text{in}}}} p\left(\bar{\boldsymbol{o}}_s^{L_{\text{in}}}, \boldsymbol{Q}_s^{L_{\text{in}}} \middle| \boldsymbol{\mu}^{L_{\text{in}}}, \boldsymbol{\Sigma}^{L_{\text{in}}}\right). \tag{3.8}$$

Then these resultant speaker-specific transforms $\hat{\boldsymbol{W}}_s^{L_{\text{in}}}$ are associated with state distributions of synthesis models in the output language through state mapping rules between the input and output languages. So the average voice synthesis models in the output language can be adapted with $\hat{\boldsymbol{W}}_s^{L_{\text{in}}}$, which functions as if it were $\hat{\boldsymbol{W}}_s^{L_{\text{out}}}$. The key point of this process is illustrated in Figure 3.2, where $\hat{\boldsymbol{W}}_s^{L_{\text{in}}} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3, \boldsymbol{W}_4, \boldsymbol{W}_5\}$ as an example.

As reported by [Wu et al., 2009], the transform mapping manner provides good speech quality, but speaker similarity is degraded.

3.3.5 Summary

Two types of solutions for cross-lingual speaker adaptation have been reviewed in this section: multilingual modelling and explicit mapping between monolingual models. As for multilingual modelling, speaker and language factorization solves the two problems with direct bilingual modelling [Zen et al., 2012]: (1) All the speech data from different languages and speakers is simply mixed for model training, so acoustic variations among languages as well as speakers are not well dealt with; (2) Only a single decision tree per state is used to represent all the training languages, without taking into account the possibility that each training language might have its exclusive context-dependency, especially for prosody. As for the explicit mapping techniques, state mapping has been shown to be superior to phoneme mapping because it

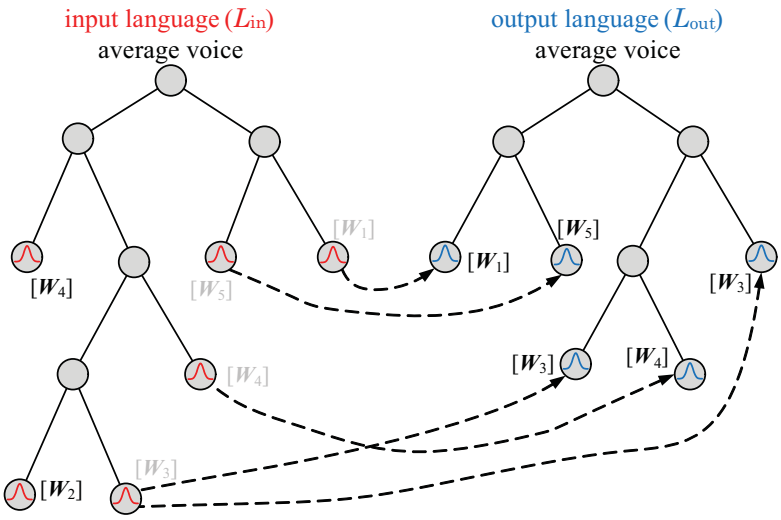


Figure 3.2 – Transform mapping manner for cross-lingual speaker adaptation. "[$\boldsymbol{W}_i$]" indicates a transform estimated by intra-lingual speaker adaptation on the side of the input language.

uses finer grained acoustic units and is based on data-oriented mapping rules [Wu et al., 2008, 2009].

HMM state mapping is theoretically and practically simpler than speaker and language factorization, and is yet to be investigated in depth. In addition, mapping at the phoneme, state or sub-state level (e.g., to map mean vectors and covariance matrices separately) is generally inevitable as long as the language of adaptation data is not one of the training languages of synthesis models. Hence, this thesis is focused on the investigation of state mapping-based cross-lingual speaker adaptation.

3.4 Speech Resources

In this section, speech corpora that have been used in this thesis for the research of cross-lingual speaker adaptation are described. These corpora include training data, adaptation data, test data and data for system enhancement (i.e., development data).

3.4.1 Training Data and Average Voice Synthesis Models

No dedicated speech database was recorded for building average voice synthesis models for the thesis work. Specially designed training data is not necessary since training average voice synthesis models over a speech corpus that was originally designed for continuous speech recognition proved to be viable [Yamagishi et al., 2010a].

As a result, five sets of average voice synthesis models were built on WSJ0 [Paul and Baker, 1992], SPEECON [Iskra et al., 2002], WSJCAM0 [Robinson et al., 1995], GlobalPhone [Schultz,

2002] and PHONDAT1 respectively for subsequent experiments in this thesis. The phoneme sets of these languages can be found in Appendix A. Speech features for training the five model sets included

1. 39th-order STRAIGHT mel-cepstra,
2. one-dimensional $\log F_0$,
3. band aperiodicity (BNDAP),
4. first- and second-order dynamic features (delta and delta-delta coefficients) of the above three kinds of features,

and were extracted from 16kHz WAV files with a window shift of 5 milliseconds. The HMM topology was five-state, left-to-right with no skip and single Gaussian-per-state. Table 3.2 presents their specifics.

Table 3.2 – Specifics of the five average voices employed in the thesis

Average voice ID	AV-ENG-US	AV-CMN-sc
Training corpus	WSJ0 SI84	SPEECON
Language	American English	Mandarin Chinese
# of training speakers used (♂+♀)	43 + 40	97 + 103
# of training utterances used	7085	5914
Total duration (hours)	13.66	12.29
Dimensionality of static BNDAP	5	5
# of tied states of spectrum	3203	2975
System paradigm	HTS-2007 [Yamagishi et al., 2009b]	

Average voice ID	AV-ENG-UK	AV-CMN-gp	AV-DEU
Training corpus	WSJCAM0	GlobalPhone	PHONDAT1
Language	British English	Mandarin Chinese	German
# of training speakers used (♂+♀)	53 + 39	56 + 54	73 + 71
# of training utterances used	9891	5419	10090
Total duration (hours)	18.90	13.37	9.60
Dimensionality of static BNDAP	21	21	21
# of tied states of spectrum	4014	2829	2155
System paradigm	HTS-2010 [Yamagishi and Watts, 2010]		

3.4.2 Adaptation, Test and Development Data

Despite the fact that only monolingual speech data is required by cross-lingual speaker adaptation, a bilingual corpus is considered "indispensable" for research. A bilingual corpus in this thesis refers to a collection of spoken data read by a set of speakers where each speaker has recorded utterances in two languages (there is no mid-utterance code-switching) in the same

chamber. Ideally, average voice synthesis models should be trained over a bilingual corpus for multilingual and cross-lingual speech synthesis research, because, for example, an underlying assumption of the state mapping technique is that speaker variability has been factored out of average voice models so that the average voice in the input language is identical to that in the output language. This assumption is true only when the two average voices are trained over a bilingual corpus. It is unfortunate that training average voice synthesis models over a bilingual corpus is not possible in most cases, mainly because of the difficulty of finding sufficient fluent bilingual speakers.

A bilingual corpus is effectively used for two main purposes in multilingual and cross-lingual speech processing research. Firstly, speech data of a target speaker in the input and output languages is used as adaptation data and test data, respectively. Secondly, since state-of-the-art adaptation techniques always blindly adapt all the aspects (speaker characteristics, background noise characteristics, etc) of synthesis models towards those of adaptation data simultaneously, a bilingual corpus needs to be used to keep speaker (and background noise) characteristics constant such that language characteristics can be focused on. For example, a bilingual corpus is particularly useful as development data in the work in Chapter 5, where it is hoped to enhance HMM state mapping construction and regression class tree growth in order to alleviate negative effects caused by the inherent language mismatch problem with cross-lingual speaker adaptation.

3.4.3 Bilingual Corpora Employed in the Thesis Work

Three bilingual corpora were involved for the thesis work: a pilot corpus in Mandarin and English, a high-quality corpus in Mandarin and English and another high-quality corpus in German and English.

(1) Pilot bilingual corpus (Mandarin and English)

The pilot bilingual (Mandarin and English) corpus contains two male native Mandarin speakers (H and Z) and was recorded in a quiet meeting room in the author's laboratory in 2009. The two speakers speak English well but Z has a pronounced foreign accent when speaking English.

There are 40 adaptation and 22 test utterances per language per speaker in this pilot bilingual corpus. The Mandarin and English prompts were selected from SPEECON and WSJ0, respectively. H and Z read the same prompts.

(2) High-quality bilingual corpora (Mandarin/German and English)

Two high-quality bilingual corpora [2] were recorded in an anechoic studio (German & English [Wester, 2010b], and Mandarin & English [Wester and Liang, 2011]) in the University of Edinburgh in 2010. The speakers are native speakers of German or Mandarin. On the basis of

2. http://www.emime.org/participate/emime-bilingual-database

the English accent rating results in [Wester, 2010b] and [Wester and Liang, 2011], five male and five female speakers are selected from the German-English corpus, six male and five female speakers are selected from the Mandarin-English one and they have the most natural English accent. In addition, a male Mandarin-English speaker whose spoken English is heavily Mandarin-accented was also selected. The 22 speakers are listed in Table 3.3.

Table 3.3 – Bilingual speakers involved in the thesis

Native language	Gender	Bilingual speaker ID						
German	male	GM1	GM2	GM3	GM6	GM7		
German	female	GF1	GF2	GF4	GF6	GF7		
Mandarin	male	MM1	MM3	MM4	MM5	MM6[†]	MM7	MMh
Mandarin	female	MF1	MF2	MF4	MF5	MF7		

[†] the heavily accented Mandarin-English speaker
[a] Pattern of speaker IDs: [*native language* (G/M)] [*gender* (M/F)] [*serial number*]

The 22 speakers read the same English prompts. The 10 German-English speakers read the same German prompts and the 12 Mandarin-English speakers read the same Chinese prompts.

Throughout this thesis, English is always regarded as the output language. The two languages, German and Mandarin Chinese, are regarded as input languages. Table 3.4 lists the partition of the high-quality bilingual data according to the usage of different utterances.

Table 3.4 – Usage of the high-quality bilingual data

Range of utterance IDs	0001~0025	0026~0125
English	DATA-TEST-ENG-25	DATA-DEV-ENG-100[†]
Mandarin	—	DATA-ADP-CMN-100
German	—	DATA-ADP-DEU-100

[a] Pattern: DATA-[*usage*]-[*language*]-[*the number of utterances*]
[b] DEV = development, ADP = adaptation, TEST = test
[†] also used as adaptation data in intra-lingual (English) speaker adaptation

3.5 Synthesis Evaluation in the Context of Cross-Lingual Speaker Adaptation

The synthesis evaluations discussed in Sections 2.3.4 and 2.3.5 can be divided into two groups in the context of cross-lingual speaker adaptation. One group includes all the objective evaluations as well as the naturalness and intelligibility evaluations. The cross-lingual fashion of speaker adaptation is unlikely to have an impact on the evaluation of these metrics except

the essential issues discussed in Sections 2.3.4 [3] and 2.3.5 [4]. By contrast, speaker similarity evaluation may pose additional problems for listeners in the context of cross-lingual speaker adaptation. To be more specific, if reference recordings are presented to listeners in the input language in a listening test, they may find it difficult to judge speaker similarity between a reference recording and a synthesized sample, which is always in the output language. This section describes in detail the different evaluation metrics employed in this thesis for cross-lingual speaker adaptation, with specific attention to speaker similarity.

3.5.1 Objective Evaluation

Original recordings collected in the output language, as discussed in Section 3.4, are used as reference data in objective evaluations. Speech samples are generated by an adapted speech synthesizer in a cross-lingual fashion using durations obtained from forced-alignment of the reference recordings [5]. Then all the four objective metrics, mel-cepstral distortion, the voicing error rate, RMSE and correlation coefficient of F_0, can be easily calculated using the formulae presented in Section 2.3.5.

A potential problem in objective evaluations of cross-lingual speaker adaptation is that original reference recordings in the output language may have an accent different from that of average voice synthesis models, for normally only adaptation data is in the mother tongue of a target speaker. The effectiveness of objective evaluations is thus arguable: If accent is considered a part of speaker identity, objective evaluations would make more sense; otherwise objective evaluations would be less reliable because such foreign-accented evaluation data does not provide an ideal reference. In order to alleviate this problem, speakers who by-and-large had minimal foreign accents when speaking English were chosen from the bilingual corpora.

3.5.2 Subjective Evaluations of Naturalness and Intelligibility

The evaluations of naturalness and intelligibility of a speech synthesizer do not require any original reference recordings. Therefore no matter how a speech synthesizer is built (speaker-dependent, adapted in an intra-lingual fashion, or adapted in a cross-lingual fashion), the naturalness and intelligibility evaluations of the synthesizer follow exactly the same procedure as described in Section 2.3.4.

Note that improving the naturalness and intelligibility of synthesized speech is also important, although this is not the principal goal of research on personalization of speech-to-speech translation, which is improving speaker similarity.

3. I.e., subjective evaluation results could be unintentionally biased due to quite a few factors and a large number of speech samples need to be listened to to ensure evaluation results are representative of the synthesizer and reliable.

4. I.e., objective measures only correlate with human perception loosely.

5. See Section 2.3.5 for the advantage of using time-aligned durations from original recordings.

3.5.3 Subjective Evaluation of Speaker Similarity

Speaker similarity evaluation always requires an original reference recording of the target speaker's voice. In the context of cross-lingual speaker adaptation, ideally, the original reference recording should be in the output language as well. This is possible in a research laboratory since a bilingual corpus has been recorded for this purpose.

However, this is not necessarily possible in actual application scenarios of cross-lingual speaker adaptation. For example, the key motivation of personalization of speech-to-speech translation is to make people "speak" a language that they cannot actually speak, which implies that recordings of the speaker's voice in the output language are not readily available. As a result, no matter whether or not speech data in the output language can be effectively collected from a target speaker, it is necessary to conduct a speaker similarity evaluation of a personalized speech-to-speech translator using the reference speaker's voice in the input language and synthesized speech samples in the output language. This is the only convincing evaluation that reflects the performance of the personalized translator.

An essential question emerges from this: Are people capable of judging the similarity between two voices when they speak different languages? Vocal cords and vocal tract are decisive factors of how a person sounds, but speaking style also plays an important role in his speaker identity. It is likely that one can sound like a different person when speaking a different language, because of the unique phonetic and prosodic patterns of each language. Since personalized speech-to-speech translation is driven by the assumption that the answer to this essential question is *yes*, it is important to obtain confirmation of this assumption. The remainder of this section reports on experiments conducted towards this goal.

This question was already partially addressed in previous studies. [Wester, 2010a] investigated cross-lingual speaker discrimination using natural speech stimuli in two language pairs, German & English and Finnish & English. The experiments in [Wester, 2010a] shows that listeners were able to complete this task well and could discriminate between speakers significantly better than chance. However, listeners performed significantly worse when a pair of speech stimuli contained two different languages than they did when there was only a single language in a pair.

The paper [Winters et al., 2008] shows that listeners could generalize knowledge of speakers' voices across English and German, which are two phonologically similar languages. [Wester, 2010a] involved Finnish, which is from the Uralic language family rather than the Indo-European family like English and German. The results in [Wester, 2010a] shows there was no indication that speaker discrimination between Finnish and English was more difficult for native English listeners than speaker discrimination between German and English.

Listeners' ability to discriminate between speakers when comparing synthesized speech to natural speech within a single language (English) was investigated in [Wester and Karhila, 2011]. It was found that listeners also completed this task well, with speaker discrimination

results being significantly above chance. However, listeners performed significantly worse when a pair of speech stimuli contained two speech types (i.e., synthesized and natural) than they did when there was only one type (either synthesized or natural) in a pair. Furthermore, speaker discrimination across speech types was found to be more difficult for listeners than across languages.

This section investigates how well listeners are able to discriminate between speakers when they have to deal with speech stimulus pairs that cross both language and speech type boundaries, which is exactly the scenario of personalized speech-too-speech translation. It is investigated whether previous findings on the language pairs of German & English and Finnish & English also hold true for English & Mandarin Chinese, which is from the Sino-Tibetan language family. Speaker discrimination experiments with Mandarin and English were conducted, in which native English listeners were presented with natural speech stimuli in English and Mandarin, synthesized speech stimuli in English and Mandarin, or natural Mandarin speech and synthesized English speech stimuli. In each experiment, these listeners were asked to judge whether or not the utterances in a pair were spoken by the same person.

Preparation of Speech Stimuli

The bilingual (Mandarin and English) corpus [Wester and Liang, 2011] mentioned in Section 3.4.3 was used as adaptation data and natural speech stimuli in the speaker discrimination experiments. Synthesized speech stimuli in English/Mandarin were all speaker-adapted samples on the basis of AV-ENG-US/AV-CMN-sc. Five females and five males with the least degree of foreign accent in their spoken English were selected. An accent rating task was used to decide the degree of foreign accent of each speaker [Wester and Liang, 2011].

(1) Stimuli Obtained by Intra-Lingual Speaker Adaptation

The two average voices were adapted to each of the 10 selected speakers with 105 English and 60 Mandarin adaptation utterances (i.e., on average, 86060 English and 84715 Mandarin speech frames per speaker), respectively. The difference of 45 utterances was due to the fact that Mandarin sentences were much longer than English ones. To ensure the amount of adaptation data in the two languages was comparable, the number of Mandarin adaptation utterances was limited.

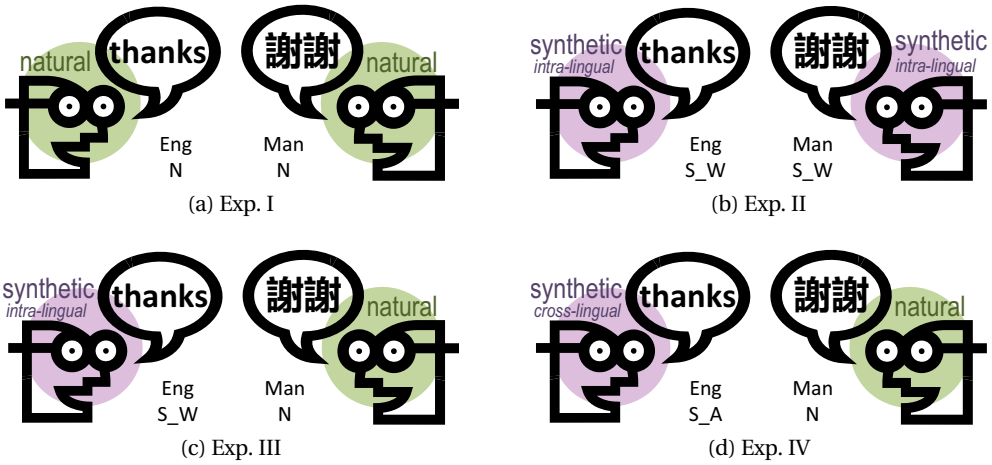
The speaker adaptation procedure was applied in the supervised intra-lingual manner. The CSMAPLR algorithm [Nakano et al., 2006, Yamagishi et al., 2009a] was employed for transform estimation. For speech stimulus generation, global variances calculated on the adaptation data and duration models of the average voices were used. The use of the average voice duration models was aimed at ensuring synthesized speech stimuli had natural prosody and were not affected by foreign prosody present in the adaptation data.

(2) Stimuli Obtained by Cross-Lingual Speaker Adaptation

The English average voice AV-ENG-US was adapted once again to each of the 10 selected speakers using their 60 Mandarin adaptation utterances. The cross-lingual speaker adaptation procedure was applied in the supervised data-mapping manner. Likewise, the CSMAPLR algorithm, global variances calculated on the adaptation data and duration models of AV-ENG-US were employed for transform estimation and speech stimulus generation.

Design of Listening Experiments

Four listening experiments (Exp. I ~ Exp. IV) were conducted to examine people's ability of discriminating between speakers across languages and/or across speech types, as shown in Figure 3.3.



(a) Exp. I (b) Exp. II

(c) Exp. III (d) Exp. IV

Figure 3.3 – Configurations of the four listening experiments

Each listening experiment consisted of two parts: one test on female speakers and one test on male speakers. So there were a total of eight listening tests, none of which included speech stimulus pairs across genders. 40 English and 40 Mandarin sentences from newspaper text were used in each listening test. None of the 80 sentences had been used as adaptation data. Each listening test consisted of 160 stimulus pairs (i.e., 320 utterances in total). Each sentence occurred four times in a listening test – twice in matched-speaker pairs and twice in mixed-speaker pairs. The two sentences within a pair were always different. Each of the five male (or female) speakers was presented in combination with every other male (or female) speaker twice and counterbalanced for order. It was also ensured that the number of mixed-language pairs was equal to that of matched-language pairs.

Eighty native English listeners with no known hearing, speech or language problems, 20-30 years of age, were recruited at the University of Edinburgh. Each listener participated in one of the eight listening tests (thus 10 listeners per listening test). This took between 35 and 45 minutes. The listeners were asked to judge whether the two utterances in each pair were uttered by the same speaker or two different speakers. In addition, they were asked to

indicate on a 3-point scale how sure they were of their judgements. Listeners were paid for their participation.

Experimental Results

In all box plots in this section, a median is indicated by a solid bar across a box which shows quartiles; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented by circles.

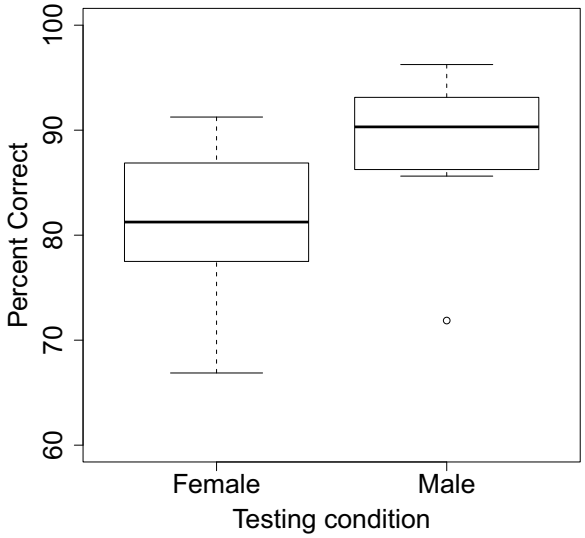


Figure 3.4 – Percent correct in Exp. I (i.e., only natural speech stimuli)

Results from all the 10 listeners in each of the eight listening tests were pooled. Figure 3.4 shows the results of Exp. I, where only natural speech stimuli were presented to listeners. An analysis of variance (ANOVA) with speaker gender as the between-test factor shows that there was a significant main effect of speaker gender at the 5% significance level [$F(1,18) = 6.49,\ p = 0.02014$]. Therefore, results on male and female speakers are presented separately in the following analysis.

Figure 3.5 shows box plot results of all the four experiments. The order of presentation of the mixed-language pairs – "Eng/Man" and "Man/Eng" – did not have a significant effect on percent correct, so they were combined. ANOVAs with the type of language pair ("Eng/Eng", "Man/Man" and "Eng/Man") as the within-test factor were conducted for all the four experiments. In all cases, a significant main effect of the type of language pair was found. Tukey's HSD tests show that listeners performed significantly worse when listening to mixed-language pairs than they did when listening to matched-language pairs. For both female and male speakers in Exp. IV, there was also a significant difference between "Man/Man" and "Eng/Eng". This was in contrast to previous experiments, in which no significant differences between

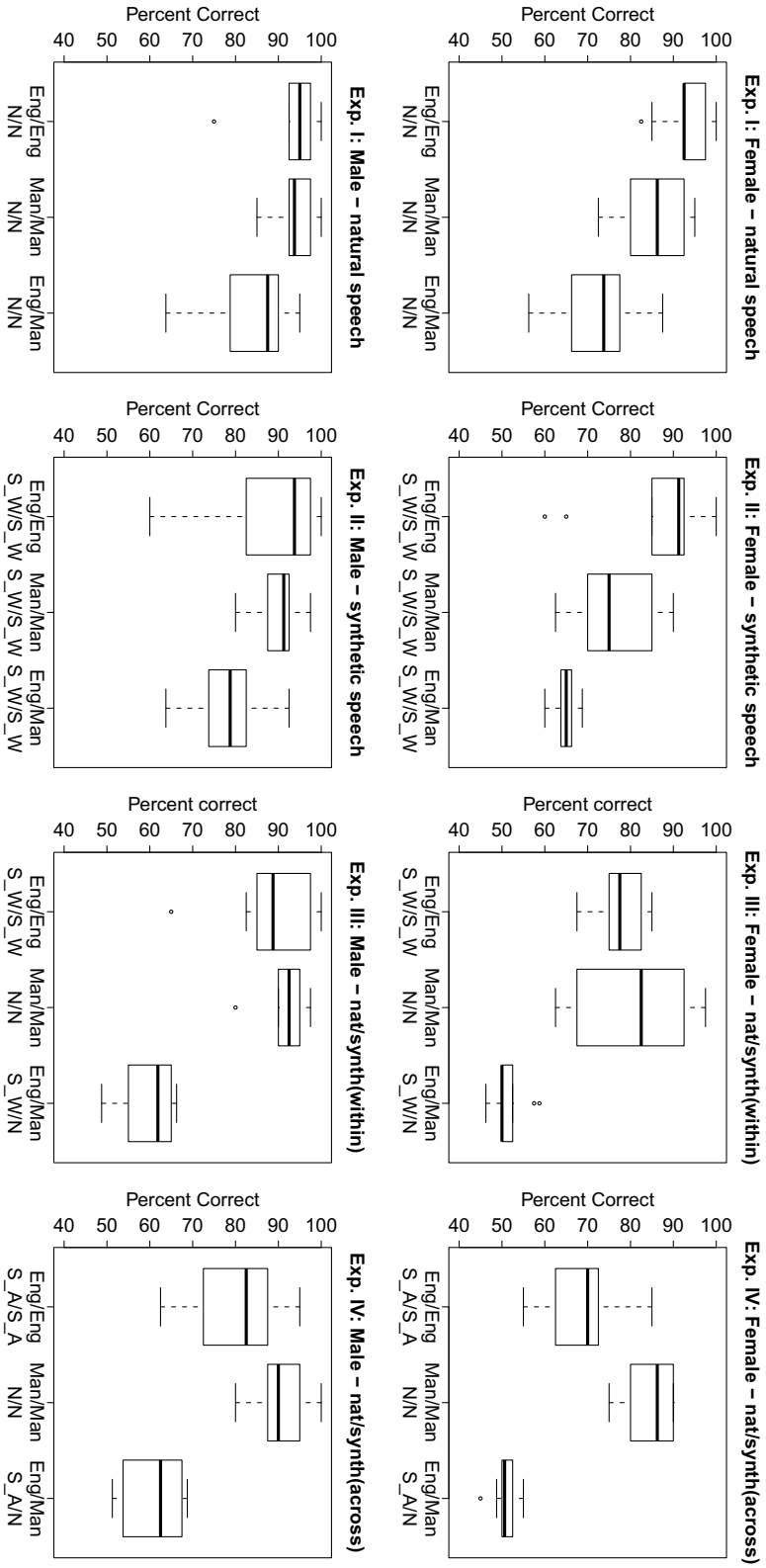


Figure 3.5 – Percent correct in the eight listening tests (N=Natural speech, S=Synthesized speech, _W=Within-language adaptation, _A=Across-language adaptation)

matched-language pairs were found, irrespective of the speech being natural or synthesized.

Table 3.5 shows the results in terms of mean percent correct for each of the four experiments. Differences in terms of percent correct between these experiments are also given.

Table 3.5 – Mean percent correct in all the four experiments

Speaker gender	Experiment ID	Language pair (%)		
		Eng/Eng	Man/Man	Eng/Man
female	I (Eng N, Man N)	92.8	85.5	72.6
	II (Eng S_W, Man S_W)	86.3	76.3	64.6
	III (Eng S_W, Man N)	77.3	81.0	51.5
	IV (Eng S_A, Man N)	69.3	84.5	50.6
(difference)	*I – II*	***6.5***	***9.2***	***8.0***
	II – III	*9.0*	*-4.7*	***13.1***
	III – IV	***8.0***	*-3.5*	***0.9***
male	I (Eng N, Man N)	94.0	94.0	84.0
	II (Eng S_W, Man S_W)	89.3	89.8	78.1
	III (Eng S_W, Man N)	88.3	92.3	60.4
	IV (Eng S_A, Man N)	80.5	90.8	61.1
(difference)	*I – II*	***4.7***	***4.2***	***5.9***
	II – III	*1.0*	*-2.5*	***17.7***
	III – IV	***7.8***	*1.5*	***-0.7***

Discussions

It was shown in [Wester, 2010a] that when comparing speech stimuli across languages (English & German and English & Finnish), listeners' performance dropped on average by 10 percentage points, from 90-100% correct (matched-language) to 80-90% correct (mixed-language). Exp. I shows a similar picture. For the male Mandarin-English speakers, listeners followed this pattern exactly. For the female Mandarin-English speakers, the results were about 10% lower.

Speaker discrimination using Mandarin & English does not seem to be more difficult for native English listeners than that using German & English or Finnish & English, when we look at the cases of using male speakers. However, significant differences are found between the results of listeners on female Mandarin-English speakers and other female speakers (German-English and Finnish-English), as well as between the results of listeners on female Mandarin-English speakers and the male German-English speakers. The most likely explanation would be that the five female Mandarin-English speakers are intrinsically more confusable than other speakers.

To illustrate this, Figure 3.6 shows a non-metric multi-dimensional scaling (MDS) plot of the judgements given by the 80 native English listeners. The plots are 2-dimensional projections of a 4-dimensional space (stress = 0.02 for the results on male speakers and 0.014 for those on the female speakers).

The MDS plot can be interpreted as follows: The proximity between a speaker's English and Mandarin data points indicates how well listeners distinguished between speakers across the two languages. A large distance between a speaker's English and Mandarin data points indicates that they were difficult to recognize as the same person. The MDS plot also shows which speakers were most confusable, as their data points are close together. Note, however, that it is not clear from this initial analysis what the acoustic correlates of the dimensions are.

In the plot with respect to female speakers, the data points of speakers 1 and 4 totally overlap, meaning that listeners were not able to distinguish between the two speakers. Speaker 2's English and Mandarin data points are quite far away from each other. Speaker 3's English and Mandarin data points merge but are quite close to speaker 5's data points. Three out of the five female speakers were clearly difficult for the listeners. Compare this to the plot with respect to male speakers in which speakers 2, 3, 4 and 5 all have Mandarin and English data points that are near each other, i.e., listeners were able to recognize these speakers well across the two languages. Only speaker 1 seems more difficult to identify across the two languages and is more confusable with speaker 3 in Mandarin and speaker 2 in English.

When going from Exp. I to Exp. II, i.e., from natural speech to synthesized speech, we observe small drops in listeners' performance of 7-9% on female speakers and of 4-6% on male speakers. The synthesized speech created using intra-lingual speaker adaptation led to speaker identities that were recognized as individuals in matched-language pairs. The results on synthesized speech are very similar to those found on natural speech.

In Exp. III and Exp. IV, the focus was on mixed-language pairs. Going from Exp. II to Exp. III, we see a drop of 13% in listeners' performance on female speakers and a drop of 18% on male speakers. When applying cross-lingual speaker adaptation, there was no further drop in performance in mixed-language pairs. But in this case, for female speakers, listeners already performed at near chance levels. There was a drop of about 8% in the results with respect to matched-language (English) pairs, when intra-lingual speaker adaptation became cross-language speaker adaptation.

Conclusions about Speaker Similarity Evaluation

It has been confirmed that listeners are able to carry out speaker discrimination tasks well, that is, deciding whether or not a speaker in one language sounds similar to the original speaker in another language. The current study has shown that native English listeners did not experience more difficulties with Mandarin than Finnish or German in such a speaker discrimination task.

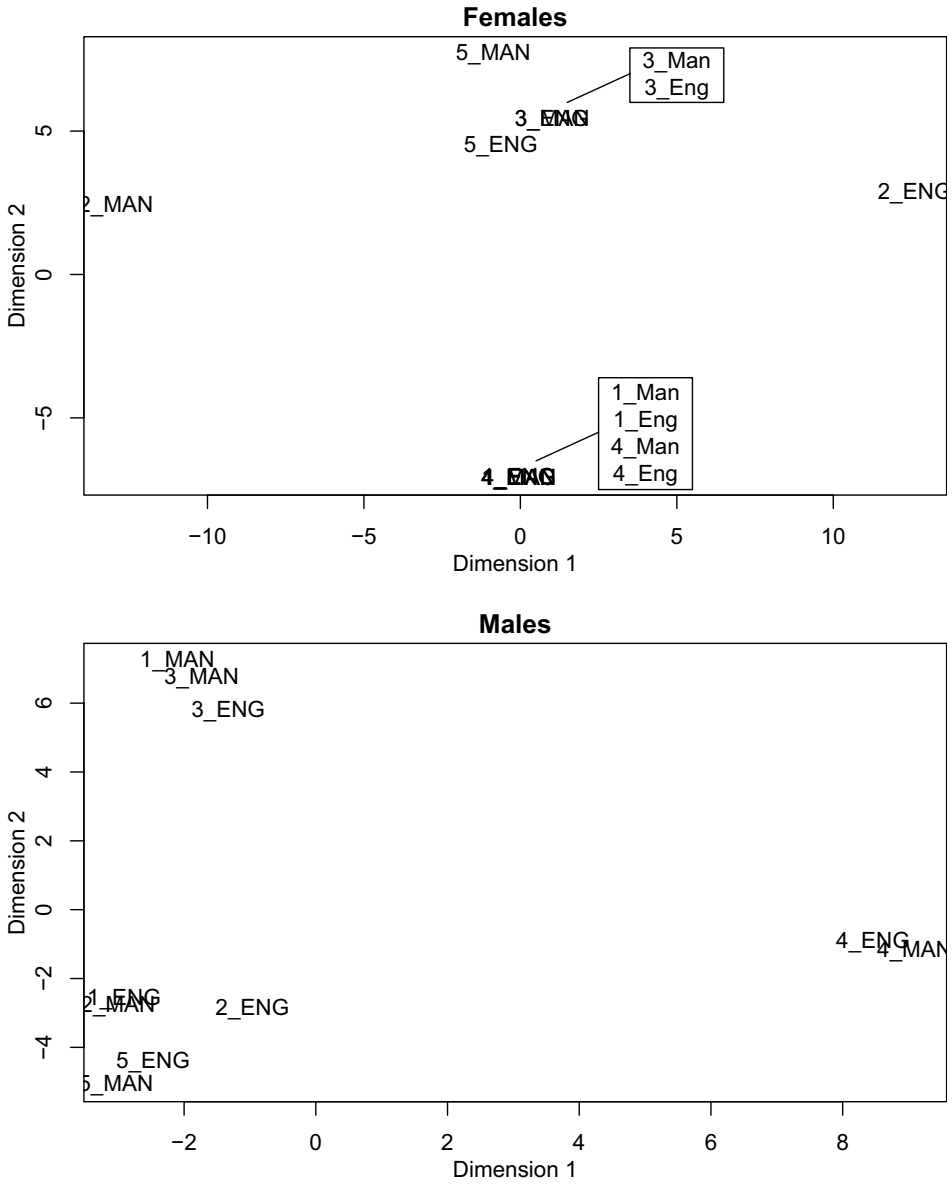


Figure 3.6 – MDS plots of the judgements of the 80 listeners

[Wester, 2010a] showed that listeners were well able to compare natural stimuli across languages (on average, 82-90% correct). The discrimination study in [Wester and Karhila, 2011] showed that listeners were also reasonably able to discriminate between speakers across speech types (synthesized and natural) *within* a language (on average, 69-73% correct). The experiments in this section show that when, in addition to comparing different speech types, listeners also had to contend with pairs across languages, their ability to correctly discriminate between speakers suffered quite substantially (on average, 51-61% correct). To summarize, listeners are able to discriminate between speakers across languages *or* across speech types, but the combination of these two factors leads to a speaker discrimination task that is too difficult for listeners to perform successfully. Consequently, future research in personalized speech-to-speech translation will need to be concentrated on further improving a speaker's synthetic voice so as to achieve the goal of sounding like the original speaker. This provides ample motivation for the work conducted in the following chapters.

3.6 Summary

In this chapter, we revisit speech processing in the multilingual and cross-lingual contexts and then discuss preparatory issues for research on personalization of speech-to-speech translation and cross-lingual speaker adaptation: state-of-the-art cross-lingual speaker adaptation approaches, model and data preparation, and evaluation of adaptation performance in the cross-lingual circumstance. A critical issue in evaluation in the cross-lingual circumstance, which is the capability of people distinguishing between speakers across languages, was investigated. It was confirmed in our experiments that cross-language speaker discrimination/identification is indeed feasible, though with some caveats.

The contribution presented in this chapter was a piece of collaborative work with Dr Mirjam Wester based in the Centre for Speech Technology Research (CSTR), the University of Edinburgh and originally published in the following conference paper:

– Mirjam WESTER and Hui LIANG, "Cross-Lingual Speaker Discrimination Using Natural and Synthetic Speech", *Proc. of Interspeech*, pp. 2481–2484, August 2011.

4 Analysis of State-of-the-Art Cross-Lingual Speaker Adaptation

4.1 Overview

The previous chapter presented an overview of cross-lingual speaker adaptation for text-to-speech synthesis, and more importantly, provides us with evidence that people are capable of distinguishing between speakers across languages, even if the languages are considerably dissimilar in terms of their phonology (e.g., Mandarin Chinese and English). This conclusion suggests that personalization of speech-to-speech translation is an attainable objective of research and deserves further attention and efforts. Meanwhile, the major difficulty has been also revealed in the previous chapter: It is the poor quality of synthesized speech through speaker adaptation (even intra-lingual speaker adaptation, let alone cross-lingual speaker adaptation) that hampers listeners' judgement when they compare voices across both languages and speech types. Therefore, the main focus in the rest of this thesis work should be to improve the performance of cross-lingual speaker adaptation, such that it can be comparable to that of intra-lingual speaker adaptation. After that, it can be assumed that improvements to monolingual speech synthesis and intra-lingual speaker adaptation will also carry over to the cross-lingual scenario.

As discussed in the previous chapter, throughout this thesis cross-lingual speaker adaptation is applied using the HMM state mapping technique. Application of HMM state mapping to cross-lingual speaker adaptation for speech synthesis is a fairly new approach (proposed in [Chen et al., 2009] and [Wu et al., 2009] in 2009) and thus has not been yet investigated in depth. It has been observed that its performance is inferior to that of intra-lingual speaker adaptation [Chen et al., 2009], but what exactly causes the gap in performance between intra-lingual and cross-lingual speaker adaptation has not been revealed by earlier work. In order to advance the state of the art, it is important that we can quantify the differences between cross-lingual and intra-lingual speaker adaptation in terms of their impacts on the quality of synthesized speech and speaker similarity that can be reproduced.

Intuitively, it is expected that the major cause of the gap in performance is the inherent mismatch between the languages of adaptation data and synthesis models used in cross-lingual

speaker adaptation, since such mismatch does not exist in intra-lingual speaker adaptation. However, it has not been analyzed how this mismatch between languages affects the performance of cross-lingual speaker adaptation for speech synthesis. In order to work out how the state of the art of cross-lingual speaker adaptation can be improved, an in-depth analysis of the impact of the inherent language mismatch is conducted in this chapter, with the goal of understanding the underlying mechanism.

Apart from the inherent issue of language mismatch in cross-lingual speaker adaptation itself, there exists another potential issue due to the scenario of personalized speech-to-speech translation. Unsupervised speaker adaptation is necessary for personalization of speech-to-speech translation, as it can help to adapt the average voice synthesis models of a speech-to-speech translator towards a user's voice characteristics as the user continues to use the translator. Nevertheless, since transcriptions of adaptation data produced by a speech recognizer may contain errors, it is possible that the unsupervised fashion is detrimental to speaker adaptation in a cross-lingual setting. Hence, an investigation is carried out in this chapter in order to examine the possibility of utilizing unsupervised cross-lingual speaker adaptation in the scenario of personalized speech-to-speech translation.

This chapter begins with unsupervised cross-lingual speaker adaptation and then an investigation into the inherent language mismatch between adaptation data and synthesis models follows. This order of presentation is due to the fact that the conclusion on unsupervised cross-lingual speaker adaptation can help to decide which fashion of adaptation (supervised or unsupervised) should be employed in subsequent research. Namely, the investigation into the inherent language mismatch will be affected by the findings with respect to unsupervised cross-lingual speaker adaptation.

4.2 Unsupervised Cross-Lingual Speaker Adaptation

As discussed previously, an additional challenge exists in the context of personalization of speech-to-speech translation, that is, *unsupervised* cross-lingual speaker adaptation. To date, research has only been conducted into unsupervised intra-lingual speaker adaptation [King et al., 2008] and supervised cross-lingual speaker adaptation [Chen et al., 2009, Wu et al., 2009] separately for speech synthesis.

In this section, two techniques, decision tree marginalization (see Section 4.2.1 for an overview) and HMM state mapping (see Section 3.3.4 for an overview), are combined in order to achieve unsupervised cross-lingual speaker adaptation and this combination is evaluated. In brief, eight speaker adaptation systems (various combinations of supervised versus unsupervised, intra-lingual versus cross-lingual) were built and their performance was compared using objective and subjective evaluations.

4.2.1 Decision Tree Marginalization

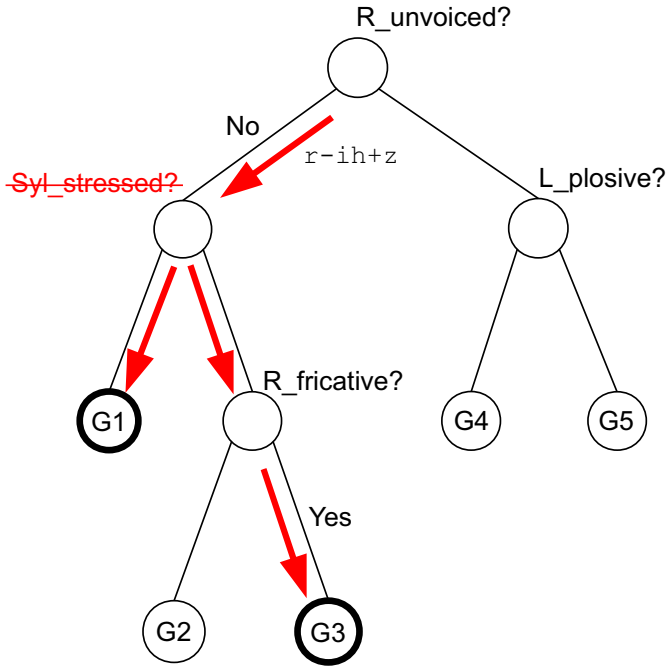
A simple and obvious method of conducting unsupervised speaker adaptation for speech recognition is to transcribe adaptation data with a well-trained, speaker-independent, triphone model-based speech recognizer and then to adapt these recognition models with the resultant transcriptions in the supervised fashion. This method can be also applied to unsupervised speaker adaptation for speech synthesis. However, it is less straightforward for speech synthesis, since we have to transcribe adaptation data at the word level using recognition models and then to extract rich context-dependent labels using a speech synthesis front-end, potentially introducing additional sources of error. As a result, the decision tree marginalization technique [Dines et al., 2009] was proposed, by which speech *synthesis* models can be used for transcribing adaptation data – in other words, adaptation data can be *directly* associated with distributions of synthesis models.

Decision tree marginalization allows the derivation of speech recognition models from a rich context-dependent speech synthesis model set according to given triphone labels. Hence, the first stage is to train a conventional HMM-based speech synthesis system from scratch, in which typically, each HMM state emission distribution is composed of a single Gaussian pdf and decision trees for state tying are central phoneme-independent.

Normally, a synthesis model with new contexts can be generated by traversing the decision trees of a synthesis model set according to a new context-dependent label and eventually assigning one leaf node (i.e., one Gaussian pdf) to the context-dependent label. The basic idea of decision tree marginalization is fairly straightforward in the sense that it generates a triphone recognition model in almost the same manner. The only difference from adding a new synthesis model is that both children of a decision tree intermediate node of a synthesis model set are traversed when the question associated with this intermediate node is irrelevant to any triphone context. So finally a triphone label is associated with more than one leaf node, which form a state emission distribution of multiple Gaussian components. In other words, a triphone recognition model constructed by decision tree marginalization can be viewed as a linear combination of context-dependent single Gaussian synthesis models. No model parameters (mean vectors and covariance matrices) are changed during the whole process.

Figure 4.1 visualizes the basic idea of decision tree marginalization by showing how to create a recognition model for a triphone label "`r-ih+z`" from a tiny synthesis model set consisting of merely five Gaussian distributions. It is apparent that $p(\boldsymbol{o}_t|\mathrm{G1})$ and $p(\boldsymbol{o}_t|\mathrm{G3})$ correspond to Gaussian distributions in the synthesis model set. The prior probabilities, $P(\mathrm{G1}|\mathtt{r}-\mathtt{ih}+\mathtt{z})$ and $P(\mathrm{G3}|\mathtt{r}-\mathtt{ih}+\mathtt{z})$, are defined as *normalized* occupation counts for G1 and G3 obtained during the training stage of the synthesis model set [Dines et al., 2009], i.e., the summation of $P(\mathrm{G1}|\mathtt{r}-\mathtt{ih}+\mathtt{z})$ and $P(\mathrm{G3}|\mathtt{r}-\mathtt{ih}+\mathtt{z})$ should be equal to one. With a well-trained synthesis model set and such prior probabilities, a set of triphone recognition models can be easily constructed.

The decision tree marginalization process described above is actually a special case. It can



$$p(\boldsymbol{o}_t|\mathtt{r-ih+z}) = P(\mathrm{G1}|\mathtt{r-ih+z}) \cdot p(\boldsymbol{o}_t|\mathrm{G1}) + P(\mathrm{G3}|\mathtt{r-ih+z}) \cdot p(\boldsymbol{o}_t|\mathrm{G3})$$

Figure 4.1 – Illustration of decision tree marginalization, showing how the new recognition model "r-ih+z" is derived from the decision tree of a tiny speech synthesis system ("L_" / "R_": left/right phone context; "G1"~"G5": clustered state emission Gaussian distributions; $\boldsymbol{o}_t$: the feature observation at time t)

be extended to marginalizing out an arbitrary set of contexts in order to create models from a normal set of synthesis models. For instance, tonal monophone models can be created by marginalizing out all the contexts that are unrelated to the base phone and tone information. Apart from marginalizing out non-triphone contexts to create recognition models, the following experiments also involve marginalizing out English-specific contexts so as to construct new models as per given Mandarin labels from a normal set of English models.

4.2.2 System Description

Decision tree marginalization makes it possible to perform *unsupervised* intra-lingual speaker adaptation and HMM state mapping makes it possible to perform supervised *cross-lingual* speaker adaptation. It is thus expected that their combination should enable unsupervised cross-lingual speaker adaptation.

HMM state mapping rules and eight synthesis systems were prepared on the basis of the two average voices AV-ENG-US and AV-CMN-sc in order to verify the feasibility of the combination of these two techniques. The eight synthesis systems were paired, half of them being supervised and the other half being unsupervised. Speech data for adaptation and evaluation was from the pilot bilingual corpus (see Section 3.4.3) containing two male native Mandarin speakers

(H and Z) comprising 40 adaptation and 22 test utterances each.

Table 4.1 – Naming rules of systems to be compared

Pattern of system names: (**S/U**) (**1/2**) - (**D/T/M**)	
S/U	supervised / unsupervised
1/2	cross-lingual / intra-lingual
D/T	data mapping / transform mapping [Wu et al., 2009]
M	Decision tree marginalization was used instead of HMM state mapping. AV-CMN-sc was therefore unnecessary.

Following the rules in Table 4.1, the eight synthesis systems were named S2, S1-M, S1-T, S1-D, U2, U1-M, U1-T and U1-D:

S2 A conventional supervised intra-lingual speaker adaptation system in English.

S1-M All the English-specific contexts were marginalized out first. In other words, only language-independent questions were left in the decision trees of AV-ENG-US. As a result, each of given Mandarin context-dependent labels was associated with more than one English state distribution. Then Mandarin adaptation data could be treated as English data for "intra-lingual" speaker adaptation on the English side.

S1-T A supervised cross-lingual speaker adaptation system using transform mapping, as described in Section 3.3.4.

S1-D A supervised cross-lingual speaker adaptation system using data mapping, as described in Section 3.3.4.

U2 An unsupervised intra-lingual speaker adaptation system in English. Recognition models were constructed from AV-ENG-US through decision tree marginalization in order to generate triphone labels of English adaptation data. Then model distributions of AV-ENG-US were adapted in the supervised fashion.

U1-M All the non-triphone contexts of AV-ENG-US were marginalized out and then Mandarin adaptation data was recognized as if it were English data, thereby Mandarin adaptation data getting associated with Gaussian pdfs of AV-ENG-US. Then model distributions of AV-ENG-US were adapted in the supervised and "intra-lingual" fashion.

U1-T Speech recognition was performed with the help of decision tree marginalization on AV-CMN-sc in order to obtain estimated triphone transcriptions of Mandarin adaptation data. Once estimated triphone transcriptions of adaptation data were available, cross-lingual speaker adaptation was conducted using transform mapping in the supervised fashion.

U1-D The same approach as U1-T except that data mapping was used instead of transform mapping.

Note that as decision tree marginalization was engaged in all the four unsupervised systems

as well as S1-M, their transforms were estimated over multiple Gaussian component models instead of single Gaussian models.

The CSMAPLR [Nakano et al., 2006, Yamagishi et al., 2009a] algorithm and all the 40 adaptation utterances were used to adapt the eight synthesis systems. Global variances were calculated on the adaptation data. A simple phoneme loop was adopted as the language model for recognition, for there was no language model trained along with the acoustic, average voice synthesis models. The average phoneme error rate was around 75%. It is hypothesized that besides the effect of the simple language model, this high phoneme error rate was due to the fact that (i) the models for recognition were actually derived from the average voice synthesis models by decision tree marginalization and (ii) only a single decision tree per emitting state per stream instead of central phoneme-specific decision trees was constructed for state tying during the training stage of these synthesis models (in other words, multiple phonemes may correspond to the same state distribution for synthesis). However, the underlying purpose of recognition here was to associate adaptation data with distributions of these synthesis models rather than produce correct transcriptions of adaptation data.

4.2.3 Objective Evaluation

Mel-cepstral distortion as well as the RMSE and correlation coefficient (CorrCoef) of F_0 was calculated on all the 22 test sentences for objective evaluation. The results are presented in Table 4.2.

Table 4.2 – Objective evaluation results (supervised versus unsupervised)

	MCEP		F_0			
	MCD (dB)		RMSE (Hz)		CorrCoef	
	H	*Z*	*H*	*Z*	*H*	*Z*
the average voice	8.55	8.78	26.0	35.9	0.46	0.49
S2	6.36	6.40	11.8	9.6	0.46	0.56
U2	6.49	6.61	13.0	14.0	0.47	0.54
S1-T	7.58	7.48	20.0	12.6	0.47	0.51
U1-T	7.59	7.74	21.1	16.5	0.48	0.53
S1-D	6.97	7.02	19.5	12.6	0.47	0.51
U1-D	6.92	6.94	22.7	17.3	0.48	0.55
S1-M	6.77	6.85	25.9	22.3	0.48	0.54
U1-M	6.74	6.83	25.1	21.0	0.48	0.53

Table 4.2 confirms that the performance of unsupervised adaptation is comparable to that of supervised adaptation no matter which approach was applied in spite of the high phoneme

error rates that were recorded. According to Table 4.2, the following observations can be made:

(1) Intra-lingual systems S2 and U2 provide the best performance, which makes sense as there was not any kind of mismatch.

(2) It is not surprising that S1-T and U1-T provide worse performing spectrum adaptation, because the transforms were estimated on Mandarin model distributions but used to adjust English synthesis model parameters. There was obvious mismatch between the transforms and the English synthesis models.

(3) In contrast, for S1-D and U1-D where data mapping was used, mapping rules were applied to the Mandarin adaptation data before transform estimation. Since transforms were directly estimated on Mandarin data and English model distributions, there was no mismatch between the resulting transforms and the English synthesis models. Mel-cepstral distortion thus decreased.

(4) In S1-M and U1-M, without any explicit state mapping rules, the Mandarin adaptation data was directly associated with Gaussian pdfs of the English average voice synthesis models by prior phonetic knowledge and in an ML-based data-driven manner, respectively. This can be regarded as a "soft" mapping process. So S1-M and U1-M could be slightly better than S1-D and U1-D in terms of spectrum adaptation performance.

(5) Unfortunately, the great prosody distinction between English and Mandarin meant F_0 adaptation was not nearly as effective in the case of cross-lingual adaptation.

4.2.4 Subjective Evaluation

Initially speech samples for subjective evaluation were synthesized with adapted pitch contours, but unnatural pitch patterns resulting from unsupervised cross-lingual speaker adaptation were perceived. In addition, Table 4.3 confirms that the prosody of English (i.e. stress-timed & atonal) is distinct from that of Mandarin (i.e. syllable-timed & tonal). Hence, pitch and duration of utterances to be subjectively evaluated were synthesized from the English average voice AV-ENG-US. Then each synthesized pitch contour was shifted such that its mean F_0 value was equal to that of the corresponding bilingual speaker (H or Z). So our listening test merely focused on the performance of spectrum adaptation.

Table 4.3 – F_0 statistics (Unit: Hz)

Speaker	Language	Mean	StD	Min	Max
H	Mandarin	137.9	25.2	72.9	236.3
H	English	128.7	11.8	64.1	222.6
Z	Mandarin	117.9	15.4	58.1	182.1
Z	English	112.0	10.3	59.3	186.1

The listening test consisted of two sections: naturalness and speaker similarity. In the naturalness section, a listener was presented with a natural utterance first and then utterances synthesized by the eight systems as well as vocoded speech in random order. Having listened to a synthesized utterance, the listener was requested to score what he/she heard on a 5-point scale of 1 through 5, where 1 meant "completely unnatural" and 5 meant "completely natural". The speaker similarity section was designed in the same fashion, except that a listener was requested to listen to an additional utterance which was synthesized directly from AV-ENG-US and the 5-point scale was such that 1 meant "sounds like a totally different person" and 5 meant "sounds like exactly the same person".

Twenty listeners participated in our listening test. Because of the anonymity of our listening test, only two native English speakers can be confirmed among the 20 listeners. The results in Figure 4.2 and Figure 4.3 suggest that unsupervised cross-lingual speaker adaptation is comparable to or sometimes better than the supervised case in terms of naturalness. It is noted that in the case of intra-lingual speaker adaptation with speaker Z's English adaptation data, the supervised system S2 outperformed the unsupervised one U2. This is probably because speaker Z speaks Mandarin-accented English while speaker H has a more natural English accent. In order to avoid the potential effect of non-standard English accents [1], only speaker H was involved in the speaker similarity evaluation.

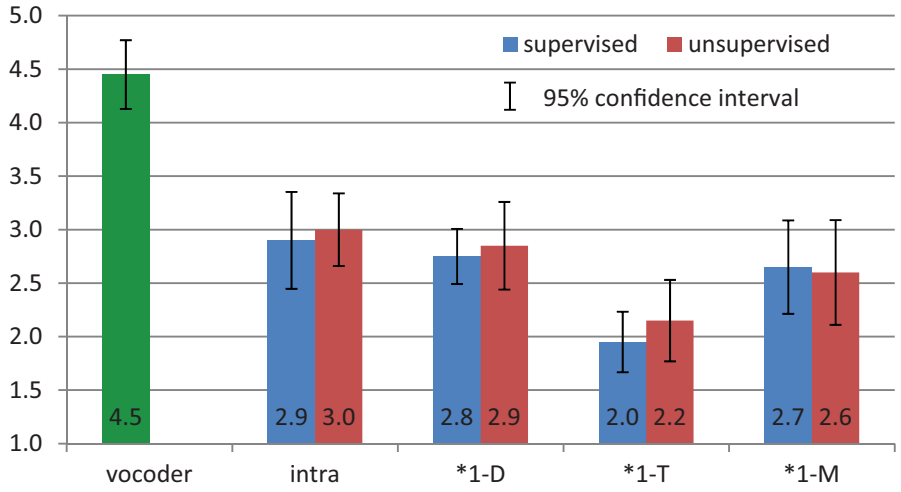


Figure 4.2 – Naturalness score (speaker H)

It is observed from both objective and subjective evaluation results that for speaker H, *1-D and *1-M followed the intra-lingual adaptation systems closely while *1-T evidently underperformed. Reviewing the analysis of Table 4.2, we note the state emission pdfs of *1-D, *1-M and the intra-lingual systems for transform estimation were all in English, which was the output language, and that the difference was just the language of their respective adaptation data. By contrast, both the state emission pdfs and adaptation data of *1-T for transform estimation were in Mandarin, which was not the output language. Hence, it would appear that the use

1. As mentioned previously, a foreign accent might be considered a part of speaker identity.

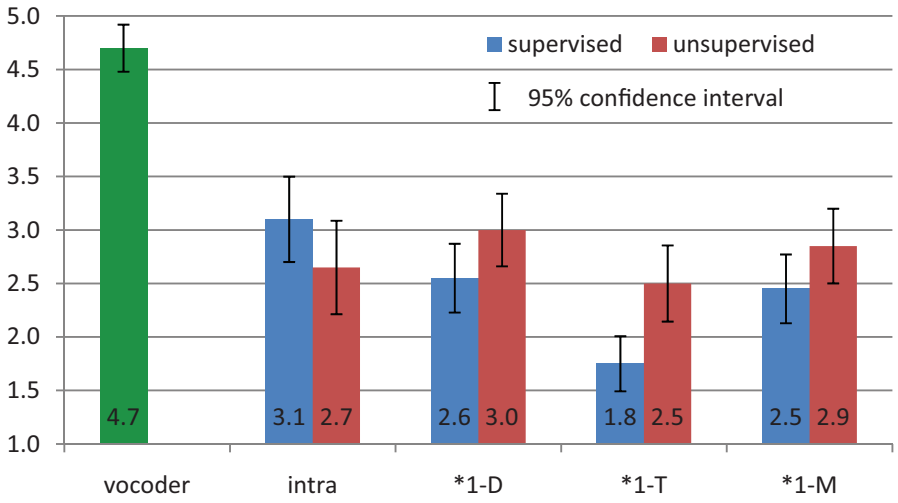


Figure 4.3 – Naturalness score (speaker Z)

of model distributions of the output language for estimation of adaptation transforms in the cross-lingual setting leads to the best results. In other words, the language of adaptation data is less important than that of a model set to be adapted.

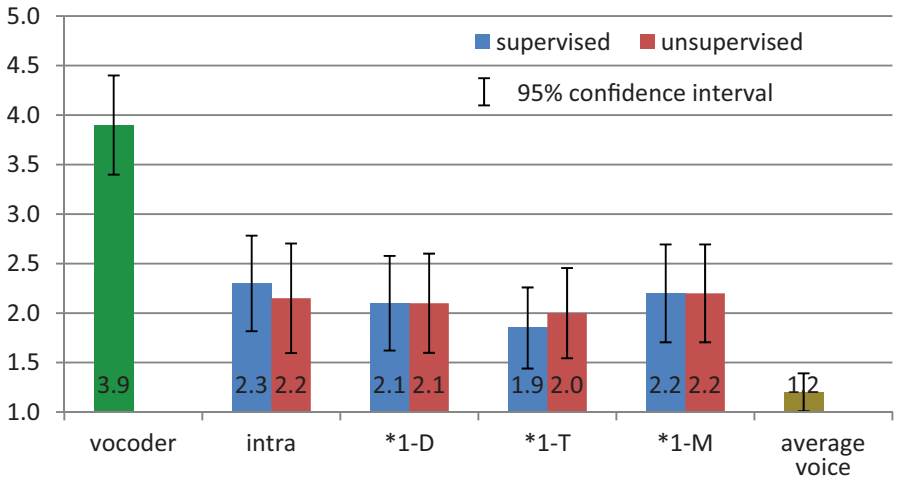


Figure 4.4 – Speaker similarity score (Mandarin reference uttered by speaker H)

The results in Figure 4.4 were obtained according to the application scenario of personalized speech-to-speech translation – speaker similarity is compared between natural speech in the input language and synthesized speech in the output language. This figure shows unsupervised speaker adaptation is comparable to the supervised case in terms of speaker similarity. However, Figure 4.5, where both natural and synthesized speech samples were in English, shows an interesting contrast in that supervised adaptation outperformed the unsupervised case. We attribute this phenomenon to human perception being affected by different cues, some of which do not transfer across languages. Namely, because the prompt of a natural English utterance was the same as that of synthesized ones, and thus they were uttered with

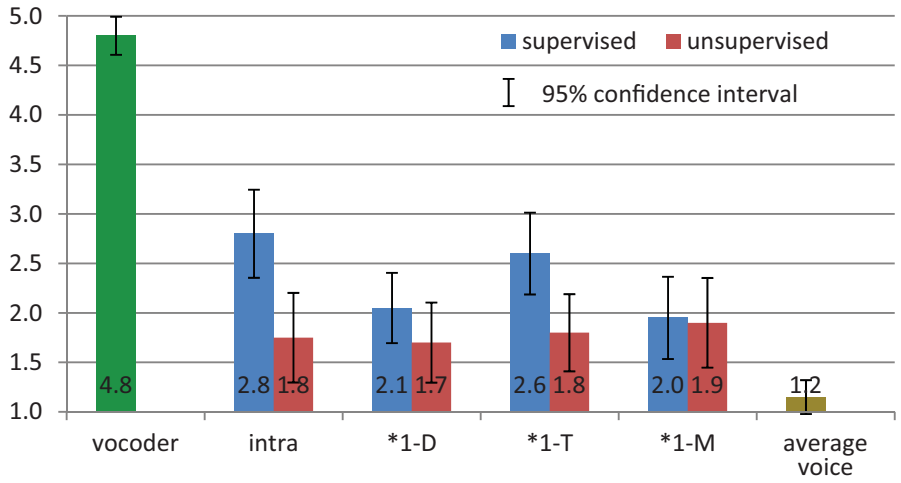


Figure 4.5 – Speaker similarity score (English reference uttered by speaker *H*)

close prosody, the listeners could more easily perceive how similar/dissimilar a synthetic voice was to a natural one, and tended to grade supervised adaptation with higher scores. In the case shown by Figure 4.4, the language difference made it more difficult for the listeners to compare a synthesized utterance with a natural one. The listeners did not think either synthetic voices (obtained in the supervised or unsupervised fashion) sounded more similar/dissimilar to the natural one. This explanation needs to be confirmed by further experiments and analysis.

The contrast between Figure 4.4 and Figure 4.5 is consistent with the conclusion of speaker discrimination experiments in Section 3.5.3, i.e., judging the similarity between two voices across both languages and speech types is a challenging task for listeners. Nevertheless, this difficulty could be considered a merit: It desensitizes human perception of speaker similarity so that it indirectly eases the development of an automated personalized speech-to-speech translator to some extent.

4.3 Impact of Mismatch between Adaptation & Synthesis Languages

The previous section demonstrates that the performance of unsupervised cross-lingual speaker adaptation is comparable to that of the supervised fashion in terms of spectrum adaptation in the scenario of personalized speech-to-speech translation. In addition, the comparability between supervised and unsupervised cross-lingual speaker adaptation is also observed in [Oura et al., 2010], where unsupervised adaptation was achieved by employing standard speech recognition models. In the remainder of this thesis, adaptation experiments were performed only in the supervised fashion, since these results indicate that the accuracy of adaptation labels is not the key determining factor in the effectiveness of cross-lingual speaker adaptation. Therefore the focus of research in this chapter moves on to the investigation into the impact of the language mismatch in cross-lingual speaker adaptation.

Cross-lingual speaker adaptation has an inherent challenge aside from the obvious lack of correspondence between adaptation data and average voice synthesis models. This challenge lies in the fact that we would like to apply adaptation algorithms such as maximum likelihood linear transformation [Gales, 1998], so that maximizing the likelihood of given adaptation data in an input language should also generalize to an increase of the likelihood (as well as objective/subjective synthesis quality) of unseen adaptation data in an output language. Although in practice adaptation algorithms employed to date have been found to work acceptably well (see [Wu et al., 2009] and Section 4.2), they make no such guarantee of generalization. The fact that conventional adaptation algorithms do not typically factor out speaker characteristics from other characteristics such as channel, noise, accent and language could be a major hindrance to such generalization.

Alleviating the influence of the language mismatch factor should improve the performance of HMM state mapping-based cross-lingual speaker adaptation and eventually make it comparable to that of intra-lingual adaptation. However, it is firstly necessary to clarify how this factor impacts cross-lingual speaker adaptation. An investigation of the effects of language mismatch on cross-lingual speaker adaptation is detailed in this section in order to fully understand the underlying mechanism and to discover potential directions for further improvements.

As mentioned in Section 3.3.4, state mapping rules are established on the basis of two sets of average voice synthesis models that are speaker-independent in order to preclude effects of speaker-specific information. The underlying assumption here is that the two sets of average voice synthesis models have an identical "voice" and overlapping acoustic space. This assumption may not be necessarily true, since the training procedure of average voice synthesis models in the EM fashion cannot guarantee such consistency, which highly depends on the method of model initialization and training corpora themselves. Such potential inconsistency between two sets of average voice synthesis models is considered one of the contributing factors to language mismatch that are looked into in this section.

4.3.1 Various Implementations of State Mapping-Based Cross-Lingual Speaker Adaptation

A set of experiments involving four ways of utilizing HMM state mapping rules constructed over two sets of average voice synthesis models was designed for the purpose of finding out how the language mismatch between average voice synthesis models and adaptation data affected cross-lingual speaker adaptation. The two approaches proposed in [Wu et al., 2009] were employed:

Data mapping

1. Establish a set of HMM state mapping rules $\mathscr{M}_{\mathrm{d}}$ over the two sets ($\mathbb{S}_{\mathrm{in}}$ and $\mathbb{S}_{\mathrm{out}}$) of average voice state distributions of the input and output languages:

$$\mathscr{M}_{\mathrm{d}}\left(S_{\mathrm{in}}^{i}\right) = S_{\mathrm{out}}^{j}, \qquad S_{\mathrm{in}}^{i} \in \mathbb{S}_{\mathrm{in}},\ S_{\mathrm{out}}^{j} \in \mathbb{S}_{\mathrm{out}}. \tag{4.1}$$

This direction of mapping rules is aimed at guaranteeing each adaptation data segment is assigned a state distribution in $\mathbb{S}_{\text{out}}$.

2. Associate all the adaptation data segments in the input language with state distributions in the output language according to $\mathscr{M}_{\text{d}}$.
3. Perform "intra-lingual" speaker adaptation on the side of the output language.

In brief, this procedure means transferring adaptation data in the input language to the output language side and then estimating transforms on the side of the output language. Figure 3.1 on page 40 visualizes the key point of data mapping.

Transform mapping

1. Establish a set of HMM state mapping rules $\mathscr{M}_{\text{t}}$ over the two sets ($\mathbb{S}_{\text{in}}$ and $\mathbb{S}_{\text{out}}$) of average voice state distributions of the input and output languages:

$$\mathscr{M}_{\text{t}}\left(S_{\text{out}}^{j}\right) = S_{\text{in}}^{i}, \qquad S_{\text{in}}^{i} \in \mathbb{S}_{\text{in}},\ S_{\text{out}}^{j} \in \mathbb{S}_{\text{out}}. \tag{4.2}$$

This direction of mapping rules is aimed at guaranteeing each state distribution in $\mathbb{S}_{\text{out}}$ is assigned a transform.

2. Perform intra-lingual speaker adaptation on the side of the input language.
3. Associate each of the state distributions in the output language with a transform obtained in Step 2 according to $\mathscr{M}_{\text{t}}$.

In brief, this procedure means estimating transforms on the side of the input language and then transferring the resulting transforms to the output language side. Figure 3.2 on page 41 visualizes the key point of transform mapping.

In order to obtain a full picture of the influence of the language mismatch between average voice synthesis models and adaptation data, two other methods of utilizing HMM state mapping rules are proposed:

Regression class tree mapping

1. According to the state mapping rules $\mathscr{M}_{\text{t}}\left(S_{\text{out}}^{j}\right) = S_{\text{in}}^{i}$, add each state distribution in the output language S_{out}^{j} into the regression class which the state distribution in the input language, $\mathscr{M}_{\text{t}}\left(S_{\text{out}}^{j}\right)$, belongs to.
2. Remove state distributions in the input language from regression classes of the input language, and then remove empty regression class tree leaf nodes of the input language.
3. Like the data mapping approach, associate adaptation data in the input language with average voice state distributions in the output language.
4. Estimate transforms over average voice state distributions in the output language and the regression class tree structure of the input language.

Conceptually, this is equivalent to transferring the regression class tree structure of the input language to the output language side and then estimating transforms on the output language side. Figure 4.6 visualizes the key point of regression class tree mapping.

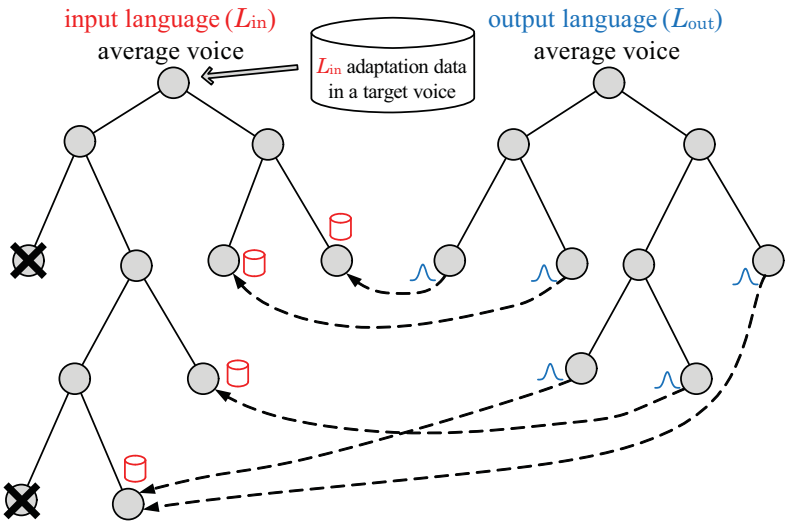


Figure 4.6 – Regression class tree mapping manner for cross-lingual speaker adaptation. Small cylinders denote adaptation data segments.

Distribution mapping

1. According to the state mapping rules $\mathcal{M}_{\text{d}}\left(S^i_{\text{in}}\right) = S^j_{\text{out}}$, add each state distribution in the input language S^i_{in} into the regression class which the state distribution in the output language, $\mathcal{M}_{\text{d}}\left(S^i_{\text{in}}\right)$, belongs to.
2. Remove state distributions in the output language from regression classes of the output language, and then remove empty regression class tree leaf nodes of the output language.
3. Estimate transforms over average voice state distributions in the input language and the regression class tree structure of the output language.
4. As transforms are associated with regression classes rather than state distributions, average voice state distributions in the output language are assigned transforms automatically.

Conceptually, this is equivalent to transferring average voice state distributions in the input language to the output language side and then estimating transforms on the output language side. Figure 4.7 visualizes the key point of distribution mapping.

As a summary, Table 4.4 presents the languages which the state distributions (StateDist), regression class trees (RegTree) and adaptation data (AdaptData) involved in the above-mentioned implementations are derived from.

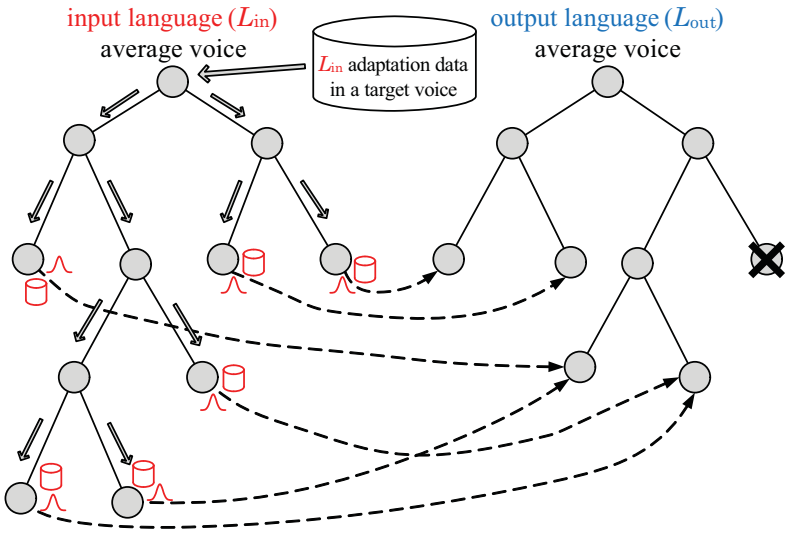


Figure 4.7 – Distribution mapping manner for cross-lingual speaker adaptation. Small cylinders denote adaptation data segments that are moving from the input language to the output language.

Table 4.4 – Overview of languages involved in the different implementations

	For transform estimation			For synthesis
	AdaptData	StateDist	RegTree	StateDist
transform mapping	L_{in}	L_{in}	L_{in}	L_{out}
distribution mapping	L_{in}	L_{in}	L_{out}	L_{out}
data mapping	L_{in}	L_{out}	L_{out}	L_{out}
regression class tree mapping	L_{in}	L_{out}	L_{in}	L_{out}

4.3.2 Isolating Sources of Language Mismatch

On the surface, language mismatch in the context of cross-lingual speaker adaptation refers to the mismatch between the language of adaptation data (L_{data}, i.e., the input language) and that of average voice state emission pdfs for *synthesis* (L_{pdf}^{syn}, i.e., the output language). This is however a vague description. In effect, language mismatch in cross-lingual speaker adaptation occurs in four possible ways:

1. between L_{data} and L_{pdf}^{adapt} during transform estimation
2. between L_{data} and L_{reg}^{adapt} during transform estimation
3. between L_{pdf}^{syn} and L_{pdf}^{adapt} during synthesis
4. between L_{pdf}^{syn} and L_{reg}^{adapt} during synthesis

$L_{\text{pdf}}^{\text{adapt}}$ and $L_{\text{reg}}^{\text{adapt}}$ refer to the languages of average voice state emission pdfs and the regression class tree that are used for transform estimation, respectively. Table 4.5 presents where language mismatch occurs in each of the four approaches described in Section 4.3.1.

Table 4.5 – Language mismatch overview ("×": mismatched; "∘": matched)

	L_{data}		$L_{\text{pdf}}^{\text{syn}}$	
	$L_{\text{pdf}}^{\text{adapt}}$	$L_{\text{reg}}^{\text{adapt}}$	$L_{\text{pdf}}^{\text{adapt}}$	$L_{\text{reg}}^{\text{adapt}}$
transform mapping	∘	∘	×	×
distribution mapping	∘	×	×	∘
data mapping	×	×	∘	∘
regression class tree mapping	×	∘	∘	×
intra-lingual	∘	∘	∘	∘
pseudo intra-lingual †	∘	×	∘	×

† This is almost the same as the intra-lingual setting, except that its regression class tree is replaced purposely with one from another synthesis system in a different language. Also see Section 4.3.3 for more information.

As a result, the four implementations described in Section 4.3.1 as a whole can comprehensively reflect the impact of language mismatch in state mapping-based cross-lingual speaker adaptation. The impact is quantified and analyzed in the following subsection.

4.3.3 Setup of Main Speaker Adaptation Experiments

The two average voices AV-ENG-US and AV-CMN-sc were used in the experiments in this section. Speech data for adaptation and evaluation was DATA-ADP-CMN-100/DATA-DEV-ENG-100 and DATA-TEST-ENG-25 uttered by the male native Mandarin speaker MMh, who has a reasonably natural English accent. The CSMAPLR [Nakano et al., 2006, Yamagishi et al., 2009a] algorithm was used and all the CSMAPLR transforms were estimated for six iterations. Global variances for synthesis were calculated on DATA-ADP-CMN-100. The main focus was on cross-lingual adaptation of mel-cepstrum and thus mel-cepstrum distortion was employed as the objective measure of adaptation performance.

Experiments on Intra-Lingual Speaker Adaptation

There is no language mismatch in intra-lingual speaker adaptation (see the fifth row of Table 4.5). Consequently, adaptation should behave in a "normal" fashion: It should reduce mel-cepstrum distortion of synthesized speech and provide further improvements as more regression class-specific transforms are estimated, given enough adaptation data. Several sets of transforms were estimated for confirmation and subsequent comparison. The description of experiments in the intra-lingual setting is as follows:

1. Each HMM stream was assigned a single transform. So there was only one global transform for mel-cepstrum adaptation.
2. Each state of each HMM stream was assigned a single transform. So there were five global transforms in all for mel-cepstrum adaptation.
3. Transforms in various quantities were estimated by setting different thresholds of transform generation.

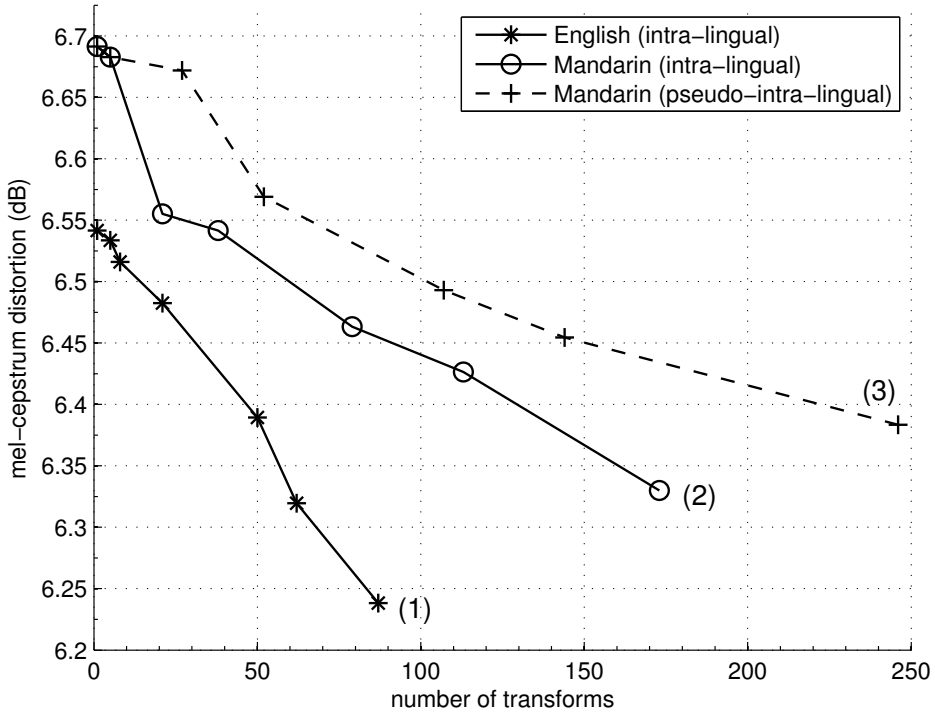


Figure 4.8 – Mel-cepstral distortion of the intra-lingual speaker adaptation systems using DATA-ADP-CMN-100 or DATA-DEV-ENG-100 in MMh's voice

It can be confirmed from the two solid lines in Figure 4.8 that a larger number of transforms can better characterize the voice of a target speaker in the intra-lingual context. Since transforms generated by distribution mapping were effectively estimated over average voice synthesis models in Mandarin, Mandarin speech was also synthesized with these transforms for further analysis. This is the *pseudo* intra-lingual case, as its $L_{\text{reg}}^{\text{adapt}}$ is English. It is involved for evaluating the impact of the source of a regression class tree (i.e., whether to be generated from synthesis models in the input or output language), given all else is matched.

Experiments on Cross-Lingual Speaker Adaptation

Cross-lingual speaker adaptation in the form of the four HMM state mapping-based implementations detailed in Section 4.3.1 was carried out. In each case, adaptation transforms

were generated in various quantities, as what was previously done for intra-lingual speaker adaptation. Objective evaluation results of cross-lingual speaker adaptation experiments are presented in Figure 4.9.

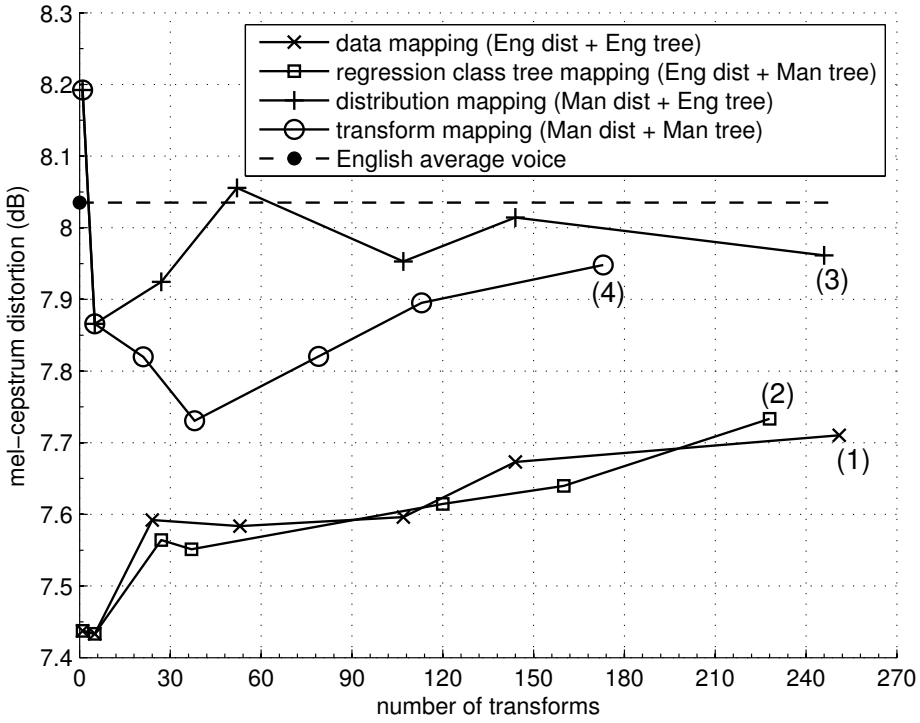


Figure 4.9 – Mel-cepstral distortion of the cross-lingual speaker adaptation systems using DATA-ADP-CMN-100 in MMh's voice

4.3.4 Analysis of the Influence of Language Mismatch

Overall Impact

The seven polylines in Figures 4.8 and 4.9 can be divided into three groups:

(a) All the polylines in Figure 4.8:
All the intra-lingual speaker adaptation systems had similar behaviour, though the deliberate misuse of an English regression class tree in the pseudo intra-lingual system introduced the mismatches between L_{data} and $L_{\text{reg}}^{\text{adapt}}$ and between $L_{\text{pdf}}^{\text{syn}}$ and $L_{\text{reg}}^{\text{adapt}}$ that resulted in worse adaptation performance.

(b) Polylines 1 and 2 in Figure 4.9:
These results pertain to cross-lingual speaker adaptation using state emission pdfs mapped from the English average voice models for both transform estimation and speech parameter generation. Both systems gave the lowest MCD values and did not

appear to be impacted by the regression class tree structure.

(c) Polylines 3 and 4 in Figure 4.9:
These English synthesis systems used adaptation transforms estimated over state emission pdfs of the Mandarin average voice models. The worst performance was achieved with the distribution mapping system, which involved language mismatch during both transform estimation and synthesis.

It is apparent that the different sources of language mismatch can have a significant impact on cross-lingual speaker adaptation. The most severe mismatch appears to be that between the distributions used to estimate adaptation transforms and those to which the transforms are applied during synthesis (i.e., between $L_{\text{pdf}}^{\text{adapt}}$ and $L_{\text{pdf}}^{\text{syn}}$). The language mismatch related to regression class tree structure appears to be less severe and less predictable in their severity.

Influence of the Number of Transforms

Polyline 4 in Figure 4.9 and Polyline 2 in Figure 4.8 actually correspond to the same transforms, which were applied to English (cross-lingual speaker adaptation) and Mandarin (intra-lingual speaker adaptation) synthesis respectively. The monotonically decreasing Polyline 2 in Figure 4.8 is what we would expect (and desire) from using an increasing number of transforms. However, when the same transforms were applied to synthesizing English speech, quite different behaviour is noted – the performance was firstly improved and then degraded after a certain number of transforms was estimated (see Polyline 4 in Figure 4.9). Likewise, the performance of data and regression class tree mapping, corresponding to Polylines 1 and 2 in Figure 4.9, was degraded immediately when more than one transform per state were estimated. This behaviour can be explained in terms of over-fitting.

When adapting average voice synthesis models, the resulting combination of models and transforms should match adaptation data. In the speaker adaptation scenario, transforms would ideally be learning only speaker-dependent characteristics to transform average voice models to speaker-dependent models, but in practice, language-dependent characteristics are also captured. In the case of transform mapping, whereby transforms are estimated over average voice models in the input language, speaker-only characteristics are better captured in the transforms since there is no language mismatch during transform estimation. As a result, using multiple regression class-specific transforms can be beneficial up to a certain point, after which the transforms become more and more language-specific and adaptation performance is degraded. In the case of data and regression class tree mapping, there is inherent language mismatch between average voice distributions for transform estimation and adaptation data. Hence, transforms immediately begin to be strongly influenced by this mismatch and using multiple regression class-specific transforms is immediately detrimental.

Despite the apparent advantage of transform mapping better taking advantage of multiple regression class-specific transforms, it still performs worse than data and regression class tree mapping. It would appear that transform mapping, while capturing fewer characteristics of

the input language, is less suitable for adapting models in the output language. Thus, data mapping and regression class tree mapping seem to provide the best way forward, but the challenge will be to develop techniques that are able to take advantage of a larger quantity of adaptation data by using regression class-specific transforms. Primarily, this would require a means to separate the effects of language and speaker mismatches that are both captured at present.

4.3.5 Subjective Evaluation

In this study we have been mainly interested in objective measures, as they relate to the adaptation criterion most closely and thus should be a more sensitive reflection of the impacts of language mismatch. Nonetheless, objective measures generally only weakly correlate with human perception [Gray Jr. and Markel, 1976, Barnwell III, 1980, Yamagishi et al., 2010a]. We performed an informal listening test for confirmation.

In the case of intra-lingual speaker adaptation, we noted speech quality was always good and that with an increasing number of regression class-specific transforms speaker similarity improved. The fact that the target speaker MMh did not have an American accent (to match the average voice models) made the use of a regression class tree particularly important – His own accent became noticeable when enough regression class-specific transforms were estimated. In all cases of cross-lingual speaker adaptation, speaker similarity was noticeably worse than that in intra-lingual speaker adaptation. For transform mapping, voice quality was maintained, but speaker similarity was poor. For data mapping and regression class tree mapping, speaker similarity was better, but voice quality was degraded (a "muddy" quality that reflects the adaptation towards Mandarin). Furthermore, synthesized speech became distorted as more regression class-specific transforms were estimated, which confirms the results obtained from the objective evaluations.

4.3.6 Follow-Up 1: Effects of the Quantity of Adaptation Data

The effects of the quantity of adaptation data on cross-lingual speaker adaptation are also worth investigating. Since data mapping using global transforms provides the best adaptation performance amongst all the cross-lingual systems, the effects of the quantity of adaptation data was looked into by conducting another data mapping experiment using global transforms: AV-ENG-US was adapted with different quantities of adaptation utterances from DATA-ADP-CMN-100 in MMh's voice. Objective evaluation results on DATA-TEST-ENG-25 are presented in Figure 4.10. Due to the size of our bilingual corpus, no more than 100 adaptation utterances could be used.

Figure 4.10 shows a rough trend that more adaptation data helps to improve cross-lingual adaptation performance. Unfortunately, the use of global transforms limits the benefits of using more adaptation data, which can be seen in the very small improvements that were

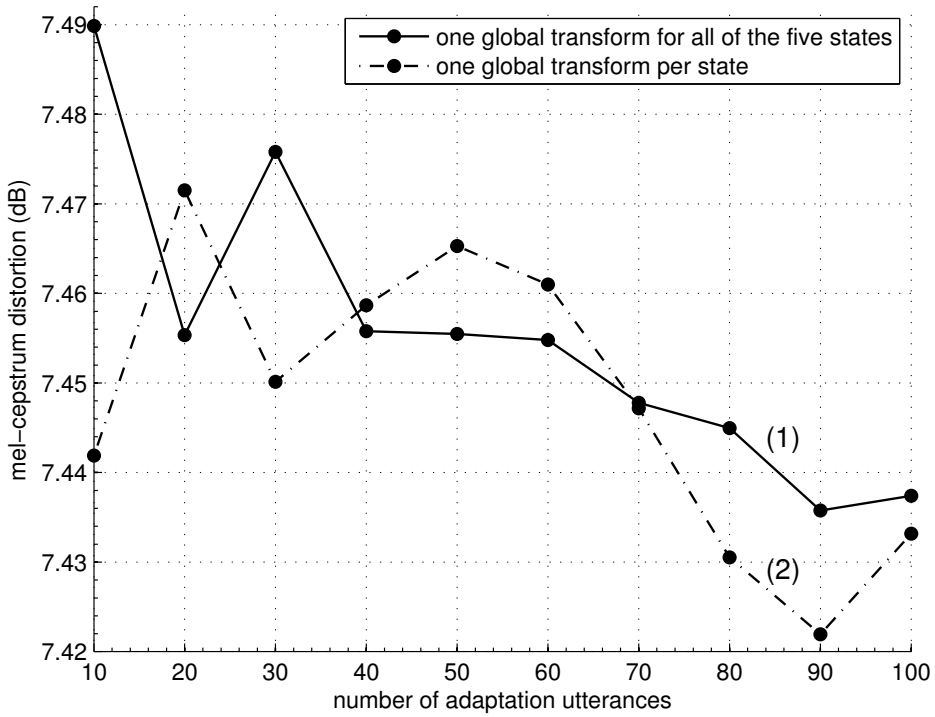


Figure 4.10 – MCD with respect to various quantities of adaptation utterances
Mel-cepstral distortion with respect to various quantities of adaptation utterances

achieved. These results further justify the need for developing new techniques which can take advantage of a large quantity of adaptation data and a regression class tree in transform estimation.

4.3.7 Follow-Up 2: Effects of the Number of Iterations of Transform Estimation

It has been demonstrated that multiple regression class-specific transforms in the data mapping-based system captured more undesirable language information than a single global transform did and thus led to worse adaptation performance. Thus it is realized that likewise, re-estimating a certain number of transforms iteratively could also add more undesirable language information in a data mapping-based system.

An experiment was carried out in order to verify the impact of the number of transform estimation iterations. Cross-lingual speaker adaptation by data mapping was carried out on the average voice AV-ENG-UK with adaptation data DATA-ADP-CMN-100 and DATA-ADP-DEU-100 in 20 speakers' voices. Two sets of CSMAPLR transforms for the synthesis of DATA-TEST-ENG-25, one containing a single global transform and the other containing multiple regression class-specific transforms, were estimated for one to six iterations in turn. Mel-

cepstral distortion on the test data set DATA-TEST-ENG-25 was calculated for the 20 target speakers and is presented in Figure 4.11.

As we anticipated, estimating adaptation transforms by data mapping in an iterative manner is detrimental to cross-lingual speaker adaptation most of the time. In particular, as Figure 4.11 shows, mel-cepstral distortion on DATA-TEST-ENG-25 consistently increases (i) when the input language is substantially phonologically distinct from the output language (e.g., Mandarin to English adaptation), regardless of whether a global or multiple regression class-specific transforms are estimated, and (ii) even when the languages are much closer (e.g., German to English adaptation) if multiple regression class-specific transforms are estimated.

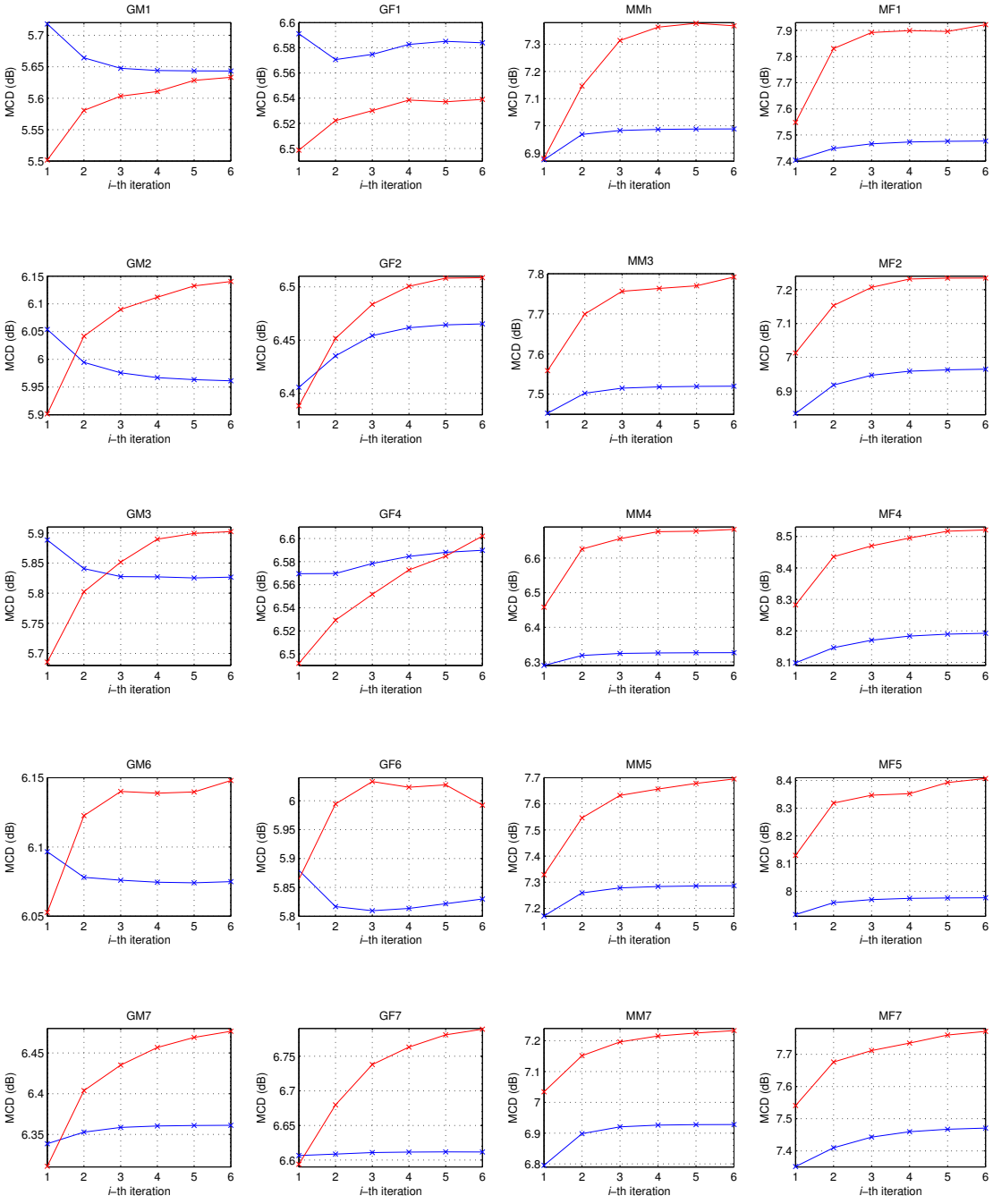
4.4 Conclusions

Two main issues have been covered in this chapter. Firstly, the possibility of employing cross-lingual speaker adaptation in the *unsupervised* fashion in the context of personalized speech-to-speech translation was investigated.

Unsupervised cross-lingual speaker adaptation was implemented by combining recently developed decision tree marginalization and HMM state mapping techniques. It was observed that unsupervised cross-lingual speaker adaptation was comparable to the supervised fashion in terms of spectrum adaptation in the scenario of personalized speech-to-speech translation, even though automatically obtained transcriptions of adaptation data had a very high phoneme error rate. This is what was hoped for – In subsequent research on personalization of speech-to-speech translation, researchers can simply focus on the supervised fashion.

Then we move on to the second issue, i.e., the investigation of how language mismatch degrades HMM state mapping-based cross-lingual speaker adaptation. In this chapter, it is demonstrated how the various sources of language mismatch impacted the different adaptation systems. From these results, it can be concluded that though HMM state mapping is an effective method to relate two different languages, it remains sensitive to the negative impacts of language mismatch. Reducing this mismatch is thus a key to advancing the state of the art. Currently, HMM state mapping rules are always constructed based on the minimum K-L divergence criterion. Alternative mapping criteria have not been investigated.

Moreover, the impacts of the number of regression class-specific transforms and the quantity of adaptation data on cross-lingual speaker adaptation have been investigated. It was found that the performance of cross-lingual speaker adaptation was degraded when many regression class-specific transforms are estimated. From the results of this part of study, it becomes clear that current approaches are largely unable to take advantage of a large quantity of adaptation data, mainly because the language mismatch between average voice synthesis models and adaptation data introduces too much unwanted language-specific information. In order to better reduce the negative impact of language mismatch and in so doing enable the effective use of a regression class tree, it is necessary to introduce new techniques that model speaker



(a) male Germans (b) female Germans (c) male Chinese (d) female Chinese

Figure 4.11 – Mel-cepstral distortion of data mapping systems on DATA-TEST-ENG-25 with respect to the number of iterations of transform estimation. The blue and red polylines correspond to estimating a single global and multiple regression class-specific transforms, respectively.

characteristics and inherent differences between languages separately, or to find a new method of growing a regression class tree.

Lastly, it is found in both investigations that the data mapping approach outperforms the transform mapping approach. Consequently, only the data mapping approach will be investigated in the following work. It was also found that estimating adaptation transforms iteratively in the data mapping approach is detrimental to the performance of cross-lingual speaker adaptation. Thus, in the experiments in Chapter 5 only a single iteration of transform estimation is employed, unless otherwise stated.

The contributions presented in this chapter were originally published in the following conference papers:

– Hui LIANG, John DINES and Lakshmi SAHEER, "A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-Based Speech Synthesis", *Proc. of ICASSP*, pp. 4598–4601, March 2010.

– Hui LIANG and John DINES, "An Analysis of Language Mismatch in HMM State Mapping-Based Cross-Lingual Speaker Adaptation", *Proc. of Interspeech*, pp. 622–625, September 2010.

5 Data-Driven Adaptation Framework Using Phonological Knowledge

In the previous chapter, HMM state mapping with the K-L divergence as a measure of the similarity between state distributions has been shown to be a simple and effective technique that enables cross-lingual speaker adaptation for text-to-speech synthesis. Meanwhile, the weakness of this technique is also noticeable: it constructs state mapping rules only based on means and variances of HMM state distributions, ignoring any other information that may positively contribute to state mapping construction, for example, the phoneme(s) which an HMM state represents. In this chapter, a jointly data-driven and phonological knowledge-guided approach that produces enhanced state mapping rules is presented: HMM state distributions derived from the input and output languages are clustered according to broad phonetic categories using a decision tree, and state mapping rules are then constructed only within each resultant phonologically consistent cluster as per the minimum K-L divergence criterion.

Apart from this, the previous chapter showed that regression class trees which followed the decision tree structure for state tying provided minimal benefits and usually resulted in degradation of synthesis quality. Thus the basic idea of the jointly data-driven and phonological knowledge-guided approach is also applied to regression class tree growth as well: HMM 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 cross-lingual speaker adaptation.

In this chapter, HMM state mapping is presented from the *data mapping* perspective since the previous chapter has shown a preference for this approach, though the proposed jointly data-driven and phonological knowledge-guided approach may equally generalize to other state mapping approaches as well. Adaptation of spectrum, which is the dominant component of speaker identity [Türk and Arslan, 2003], is the focus of this research.

There exists a potential confusion in this chapter: Two sets of decision trees are touched upon here, one of which is obtained in the normal training stage of synthesis models while the other is generated during the enhancement of state mapping rules by the jointly data-driven and

phonological knowledge-guided approach. The two sets of decision trees are involved for completely distinct purposes. Furthermore, the trees derived for enhanced state mapping rules are also distinct from those derived for enhanced regression classes.

5.1 Preliminary Investigations

First of all, two preliminary experiments were carried out, in order to test the hypothesis on the sub-optimality of the minimum K-L divergence criterion for determining state mapping rules between average voice synthesis models of two languages.

5.1.1 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 was conducted: State mapping rules between AV-ENG-US and AV-CMN-sc were constructed and then AV-ENG-US was adapted with DATA-ADP-CMN-100 in speaker MMh's voice. A slight difference in this experiment was that this time HMM state mapping rules defined by the k-th best match in the output language were used for each state in the input language, instead of always selecting the best match satisfying the minimum KLD criterion (i.e., $k \equiv 1$).

Table 5.1 – Results obtained under the k-th best match criterion for cross-lingual speaker adaptation in the data mapping manner

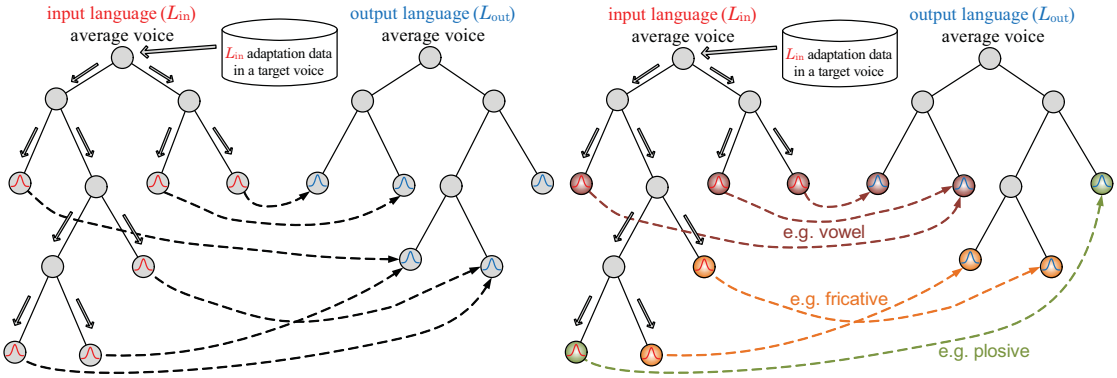
k	MCD (dB)	k	MCD (dB)
1	7.67	10	7.76
2	7.64	20	7.98
3	7.64	30	8.16
4	7.64	40	8.38
5	7.80	50	8.48

Ten values of k were evaluated in turn and corresponding mel-cepstral distortion was calculated on DATA-TEST-ENG-25. Measurements in Table 5.1 show that while mel-cepstral distortion does generally increase with increasing k, this is only apparent for $k > 5$. This phenomenon suggests that while the K-L divergence 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.

5.1.2 Introduction of 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 state mapping rules may be avoided by taking advantage of the knowledge of underlying phoneme categories.

Taking the case of $k = 1$ in Table 5.1 (i.e., the baseline data mapping approach), state distributions of AV-ENG-US and AV-CMN-sc were categorized according to seven broad phoneme categories (silence, vowel [1], 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. Figure 5.1 visualizes the difference between the baseline and this simple phonological knowledge-guided approach.



(a) baseline method (b) phonological knowledge-guided

Figure 5.1 – HMM state mapping construction for cross-lingual speaker adaptation in the data mapping manner. The dashed lines refer to state mapping rules.

AV-ENG-US was adapted using DATA-ADP-CMN-100 in speaker MMh's voice and the new set of state mapping rules. Then mel-cepstral distortion was calculated on DATA-TEST-ENG-25. Objective evaluation results are presented in Table 5.2.

Table 5.2 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,

1. The reason why there was only one category for vowels is that unlike consonants, there does not exist any apparent gap in the vowel quadrilateral (see Appendix B). It is less straightforward how to categorize vowels appropriately, especially those like /æ/, /ɪ/, /ʊ/, etc.

Table 5.2 – 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 investigation of further means to exploit phonological knowledge was pursued as detailed in the remainder of this chapter.

5.2 Data-Driven & Phonological Knowledge-Guided State Mapping Construction

In the previous section, 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 the input language could only be mapped to its phonologically consistent counterpart in the output language 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 [Wu et al., 2008]. 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 the thesis work in a similar fashion to that in synthesis model training. Well-trained HMM state distributions of average voice synthesis models in the input and output languages 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 the output language is minimized.

5.2.1 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, central phoneme and right phoneme. Consequently, the triphone contexts are considered an essential factor for grouping average voice state distributions of the input and output languages. In addition, we continue to 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 (listed in Table 5.3) for 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”

Table 5.3 – All the questions used in the jointly data-driven and phonological knowledge-guided approach

	Left phoneme	Central phoneme	Right phoneme
Silence	L_silence	C_silence	R_silence
Vowel	L_Vowel	C_Vowel	R_Vowel
Plosive	L_Plosive	C_Plosive	R_Plosive
Fricative	L_Fricative	C_Fricative	R_Fricative
Affricate	L_Affricate	C_Affricate	R_Affricate
Approximant	L_Approximant	C_Approximant	R_Approximant
Nasal	L_Nasal	C_Nasal	R_Nasal

and "C_plosive" if it is tied to context-dependent phones `*-ch+*`, `*-k+*` and `*-p+*`.

A table of mapping from phonemes in German, American English, British English and Mandarin Chinese to the seven phoneme categories can be found in Appendix A.

5.2.2 Question Selection Criterion

Several criteria have been employed in decision tree-based clustering during synthesis model training for selecting the best question to split a node, such as maximum likelihood [Young et al., 1994] and minimum description length [Shinoda and Watanabe, 2000]. Nonetheless, the goal of speech synthesis is to generate speech as close as natural speech, which is only achieved indirectly through optimization criteria like maximum likelihood or minimum description length.

The minimum generation error criterion was proposed [Wu and Wang, 2006] 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 minimum generation error criterion has been applied to training synthesis model parameters [Wu and Wang, 2006] as well as decision tree-based state clustering [Wu et al., 2006]. According to this criterion, the question selected to split a decision tree node should be the one which minimizes a predefined measure of distortion over a particular set of speech data (the training data set of synthesis models or a new set of development data) – this idea is used in the jointly data-driven and phonological knowledge-guided 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 the output language based on adaptation of synthesis models using data in the input language. Therefore a bilingual corpus is required in the jointly data-driven and phonological knowledge-guided approach. The bilingual corpus does not need to be large as it is not used for model training like in [Qian et al., 2009].

5.2.3 Procedure for Enhancing HMM State Mapping Construction

Bilingual data from a fixed number of speakers is selected such that adaptation data in the input language is used to estimate adaptation transforms and development data in the output language 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. Form N root nodes by pooling all average voice state distributions from the input and output languages for each of the N HMM emitting states.
2. Find the next non-terminal leaf node X across the N decision trees in the manner of breadth-first search (see Figure 5.2).

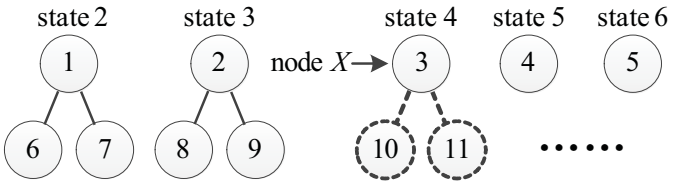


Figure 5.2 – Breadth-first search in enhanced HMM state mapping construction

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. For instance, the decision tree of state 4 may end up looking like Figure 5.3.

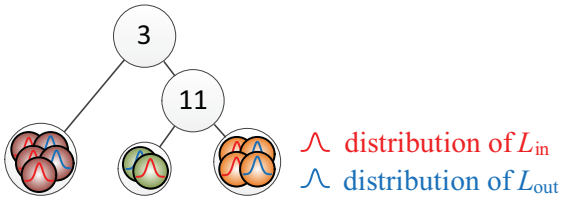


Figure 5.3 – Imaginary final structure of the decision tree of state 4

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 visualized in Figure 5.4, where node 3 in Figure 5.2 is taken as an example.

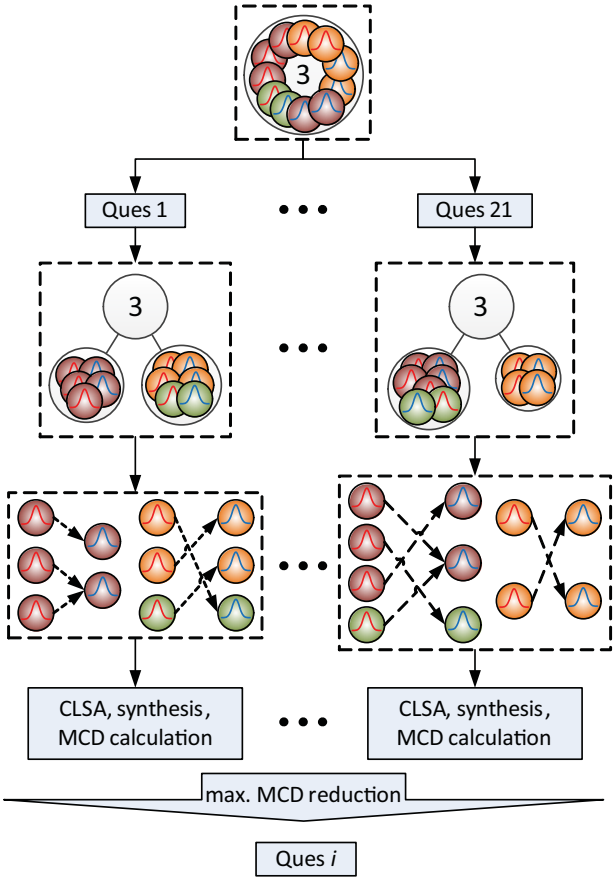


Figure 5.4 – Procedure of finding the best question to split a decision tree node under the MGE criterion for HMM state mapping construction

As [Wu and Wang, 2006] and [Wu et al., 2006] 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 jointly data-driven and phonological knowledge-guided 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).

5.3 Data-Driven & Phonological Knowledge-Guided Regression Class Tree Construction

In previous experiments it was demonstrated that regression class trees derived using the usual approaches based on either state tying [Yamagishi et al., 2004] or Euclidean clustering

[Young et al., 2009, Chapter 9] did not lead to effective cross-lingual speaker adaptation. Thus it is proposed to apply the jointly data-driven and phonological knowledge-guided approach elaborated in Section 5.2 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 jointly data-driven and phonological knowledge-guided 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 the output language.
2. Find the next non-terminal leaf node Y of 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 procedure is visualized in Figure 5.5.

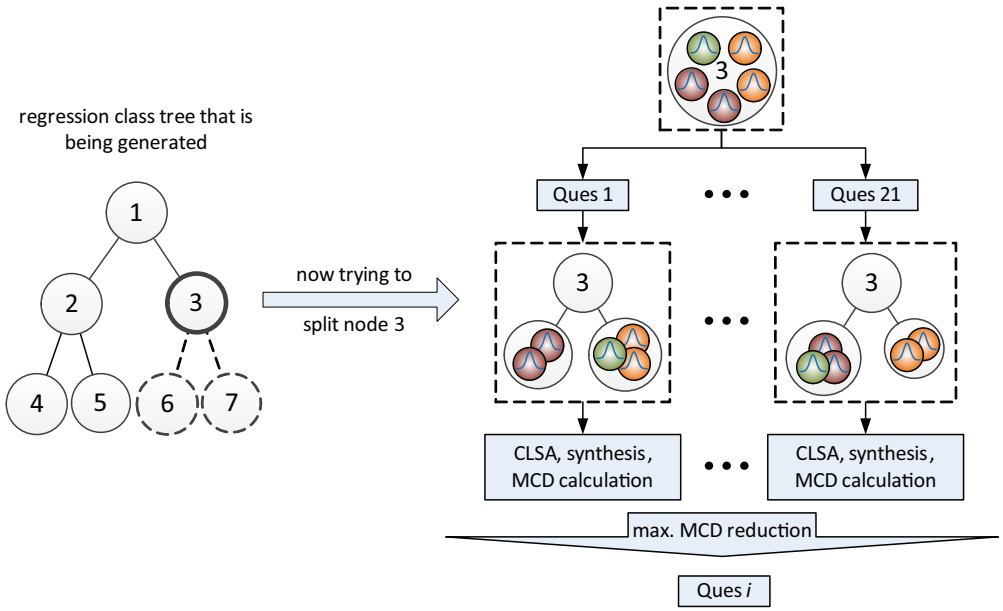


Figure 5.5 – Procedure of finding the best question to split a node of a regression class tree under the MGE criterion

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 will be limited, as we would expect for two very disparate languages.

5.4 Speaker-Dependent Experiments

5.4.1 Experimental Setup

The two average voices AV-ENG-US and AV-CMN-sc were used in speaker-dependent [2] experiments, Mandarin and English being the input and output languages respectively. The two average voices were adapted by the CSMAPLR [Nakano et al., 2006, Yamagishi et al., 2009a] algorithm for only one iteration. Global variances for synthesis were calculated on adaptation data.

Speakers and Speech Data

Three male (MMh, MM3 and MM6) and two female (MF2 and MF7) speakers were selected for speaker-dependent experiments. 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 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 the five speakers was the set DATA-ADP-CMN-100. Development data of the four training speakers was the set DATA-DEV-ENG-100. Test data of the five speakers was the set DATA-TEST-ENG-25.

Systems for Comparison

Four groups of experiments were conducted. Within each group, state mapping rules for mel-cepstra between AV-ENG-US and AV-CMN-sc were derived from *one* of the four training speakers by means of the jointly data-driven and phonological knowledge-guided approach while those for $\log F_0$, 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 American English average voice AV-ENG-US towards each of the four remaining speakers. $\varepsilon_{\Delta\mathrm{MCD}}$ was set to 0.0005dB. The baseline system merely involved the minimum KLD criterion in construction of state mapping rules for all the streams of the state emission pdfs.

In these speaker-dependent experiments, only global transform-based adaptation was investigated. Investigation of regression class-based adaptation is provided in the following

2. "Speaker-dependent" in this section means HMM state mapping rules are enhanced on the basis of development data from a single speaker.

section.

5.4.2 Objective Evaluation

Original recordings of DATA-TEST-ENG-25 of the five speakers were aligned using the average voice models AV-ENG-US 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 Figure 5.6 and Table 5.4. These MCD measurements were calculated on the entire test data set of each of the five speakers.

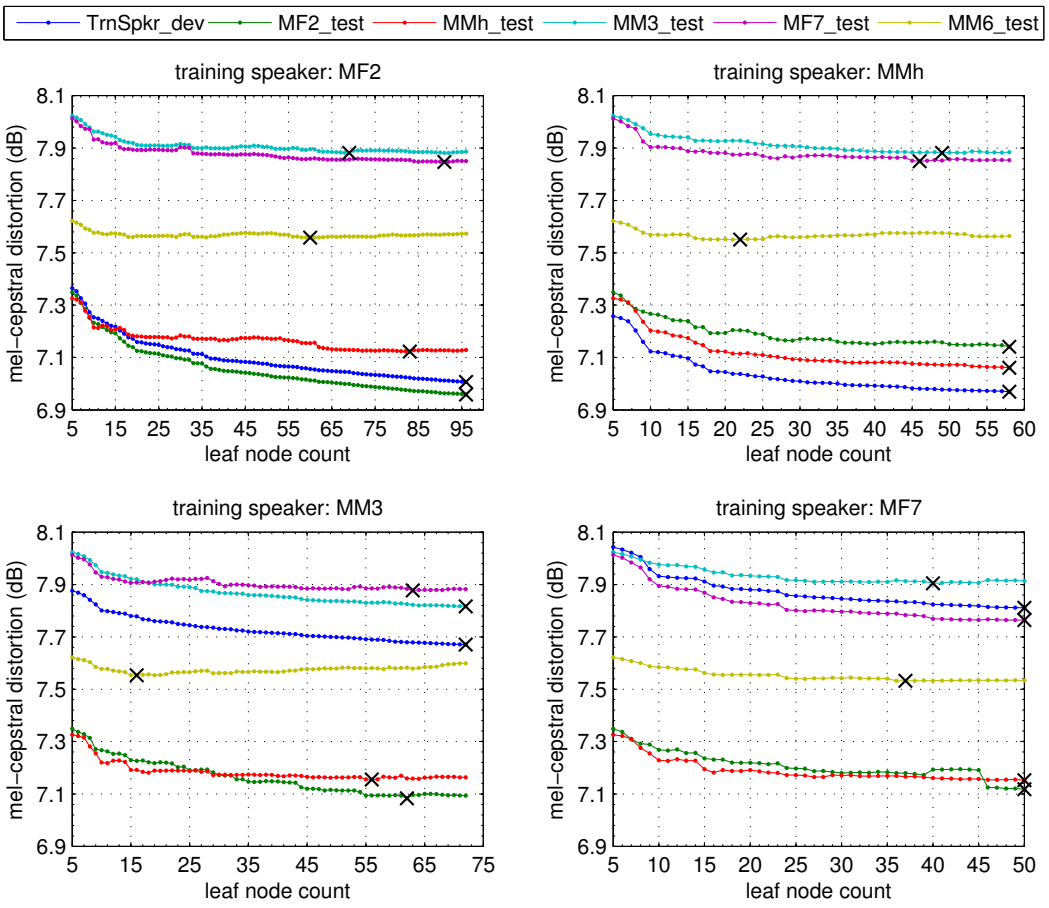


Figure 5.6 – Mel-cepstral distortion in relation to the leaf node count during decision tree generation. Crosses indicate minimums on the curves. "TrnSpkr_dev" refers to the development data of respective training speakers. "_test" refers to test data. The six points on the vertical axis in each sub-figure come from the baseline.

It can be seen from Figure 5.6 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 still have a nearly monotonically decreasing tendency. In other

Table 5.4 – MCD reduction in dB produced by the jointly data-driven and phonological knowledge-guided approach, i.e., the difference of the leftmost and rightmost values on each curve in Figure 5.6.

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

words, mapping rules constructed from a single speaker still maintain 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 American English retain their American 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. These scores are less reliable and misleading, as discussed in Section 3.5.1.

5.4.3 Impact of Phonological Knowledge on State Mapping Rules

A total of 2975 mapping rules were constructed, one for each of the states in the Mandarin average voice AV-CMN-sc. Figure 5.7 shows how k varies under the data-driven use of phonological constraints (see the definition of k in Section 5.1.1). Two common traits are observed across the four histograms in this figure.

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

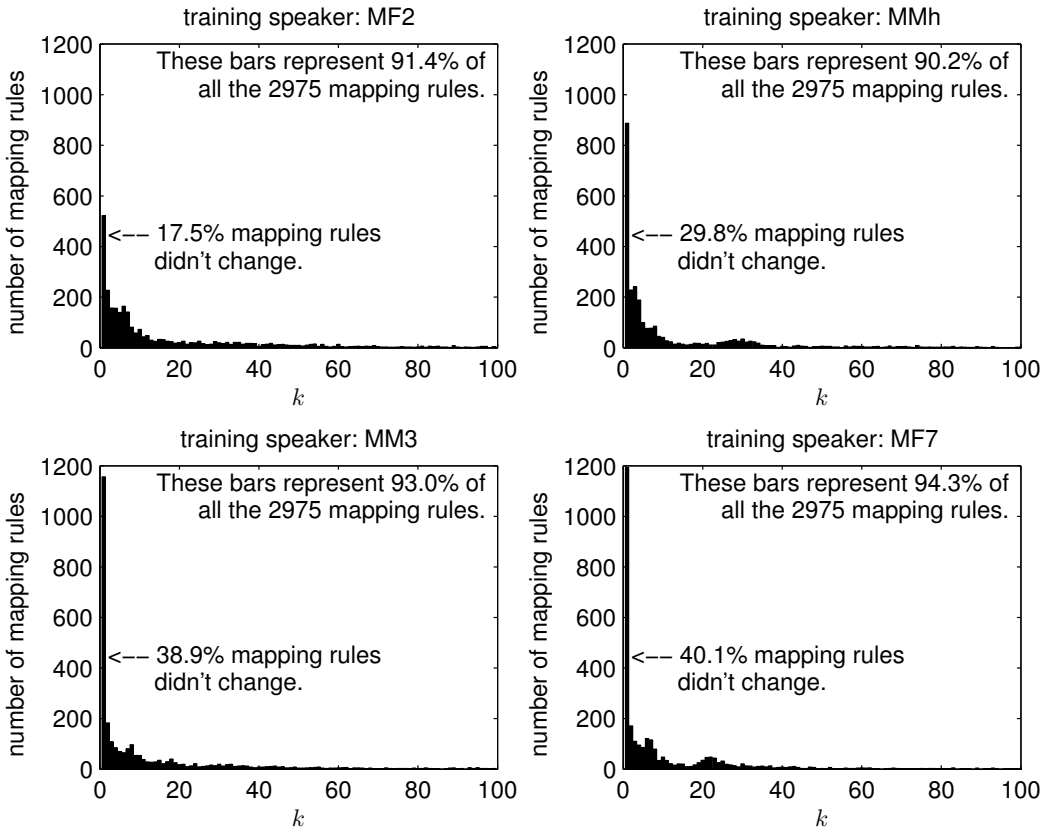


Figure 5.7 – Histogram of the KLD rank (k) using the jointly data-driven and phonological knowledge-guided approach

selected with $k \neq 1$ 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 Section 5.1.1. It is also interesting to note from both Table 5.4 and Figure 5.7 that the proposed approach has the most impact on the truly bilingual speaker MF2, in terms of the number of changed mapping rules, MCD reduction and providing the best generalization to other speakers (except MM6, as discussed previously).

5.4.4 Questions Used for Root Node Splitting

One means to analyze the generalization of the proposed jointly data-driven and phonological knowledge-guided approach is to consider the questions that have yielded the greatest MCD reduction. Table 5.5 shows the questions associated with the root node of each decision tree (which also gave the greatest MCD reduction) for each of the training speakers.

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

Table 5.5 – Root node questions for emitting states at each of the five positions (2~6) in an HMM

	MF2	MMh	MM3	MF7
State 2	L_nasal	L_nasal	L_nasal	L_nasal
State 3	C_nasal	C_nasal	C_vowel	C_nasal
State 4	C_nasal	C_nasal	C_affricate	C_affricate
State 5	R_fricative	C_affricate	C_nasal	C_affricate
State 6	L_silence	L_plosive	L_plosive	L_silence

5.4.5 Subjective Evaluation

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 Figure 5.6, since MF2 seems to provide the best generalisation to other speakers. Using the baseline and the proposed jointly data-driven and phonological knowledge-guided approach, five sentences from DATA-TEST-ENG-25 were synthesized for each of the five speakers. Note that unadapted duration models of the English average voice AV-ENG-US were used. The evaluation comprised a total of 50 AB/ABX comparisons. Original reference speech in the speaker similarity test was in English since this should lead to better discrimination between systems, as discussed previously. Subjective evaluation results are shown in Figure 5.8.

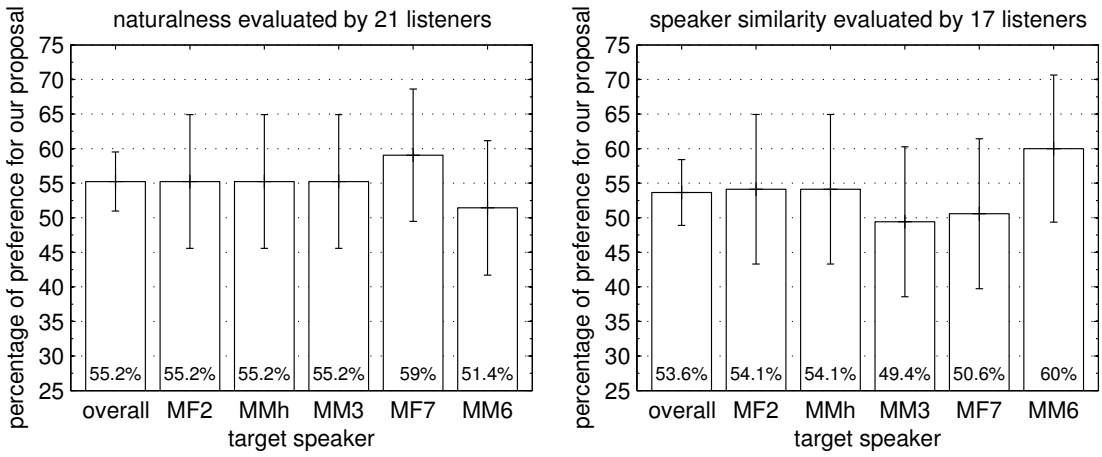


Figure 5.8 – Subjective evaluation results of the jointly data-driven and phonological knowledge-guided approach using MF2-dependent state mapping rules. Whiskers indicate 95% confidence intervals.

From informal listening, it is noted that speaker similarity was not greatly impacted by the proposed approach, but naturalness was improved (speech was produced with less "muffled" characteristics by the proposed approach). This observation is reflected in Figure 5.8.

5.5 Speaker-Independent Experiments

The effectiveness and generalization across speakers of jointly data-driven and phonological knowledge-guided state mapping construction have been demonstrated in Section 5.4. 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.

5.5.1 Experimental Setup

Three average voices were used in the speaker-independent [3] experiments: AV-ENG-UK, AV-CMN-gp and AV-DEU. The input language was either German or Mandarin Chinese. The output language was always British English. Mandarin and German were chosen as input languages as they are "far from" and "close to" English respectively. This should give some insights into the extent to which the dissimilarity of input and output languages can affect the performance of cross-lingual speaker adaptation. All of the speaker-independent cross-lingual speaker adaptation experiments were performed using the CSMAPLR [Nakano et al., 2006, Yamagishi et al., 2009a] algorithm, transforms being estimated from one iteration. Global variances for synthesis were calculated on adaptation data.

Ten Mandarin-English speakers (Chinese) and ten German-English (Germans) speakers were used in the speaker-independent experiments. They were grouped as shown in Table 5.6. The groupings were used for cross validation since the number of available bilingual training speakers was limited.

Table 5.6 – Grouping of 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 was either DATA-ADP-CMN-100 or DATA-ADP-DEU-100 for Mandarin and German speakers respectively. Development data was DATA-DEV-ENG-100 and test data was

3. "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.

DATA-TEST-ENG-25.

5.5.2 Effect of the Number of Transforms

First of all, the experiments in Section 4.3 that employed the conventional data mapping approach were repeated: the British English average voice AV-ENG-UK was adapted with either DATA-ADP-CMN-100 or DATA-ADP-DEU-100 with various regression class occupation thresholds [4] that have the effect of adjusting the number of resulting transforms. The regression class tree followed the decision tree structure of AV-ENG-UK [Yamagishi et al., 2004]. Adaptation performance is presented in Figure 5.9 in the form of mel-cepstral distortion.

In *intra-lingual* speaker adaptation, it is accepted that more adaptation data leads to improved synthesis quality via the estimation of more regression class-specific transforms. As Figure 5.9 clearly shows, this is not the case in *cross-lingual* speaker adaptation: the MCD curves never have a decreasing tendency and the optimal number of transforms varies as the phonological/acoustic similarity between input and output languages varies. When the two languages are modestly different (e.g., German to English), a regression class tree that follows the decision tree structure for state tying could be of help to a certain extent. In more extreme cases (e.g., Mandarin to English), there seems to be no benefit from the generation of multiple regression classes. It can be hypothesized that given sufficient adaptation data, the number of transforms that produces the smallest MCD in HMM state mapping-based cross-lingual speaker adaptation might be a measure of the phonological/acoustic similarity between two languages.

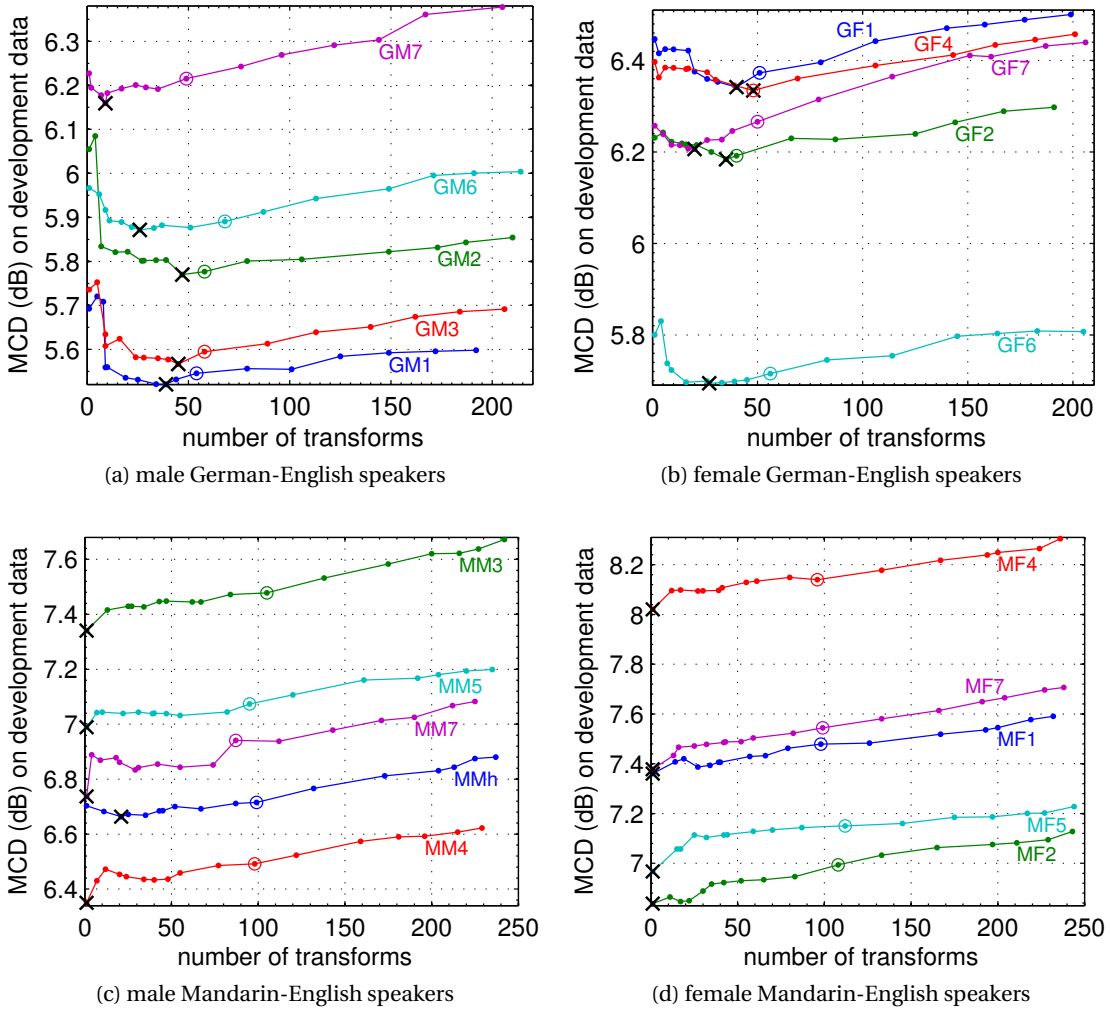
5.5.3 Systems for Analysis of the Proposed Approach

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 5.6) 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 for mel-cepstra between English and German/Mandarin were derived from the training partition by the proposed jointly data-driven and phonological knowledge-guided approach, while those for $\log F_0$, band aperiodicity and duration were still constructed under the minimum KLD criterion. These mapping rules were used for cross-lingual adaptation of the British English average voice AV-ENG-UK towards each of the test speakers.

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

4. These thresholds were 20000, 15000, 12000, 10000, 8000, 7500, 6000, 5000, 3500, 2460, 1500, 1000, 750, 650, 550 and 450. Among all these thresholds, 2460 is the one by default in the HTS-2010 system [Yamagishi and Watts, 2010], which is 1.5 times the size of a transformation matrix plus a transformation vector and is empirically a good choice.



(a) male German-English speakers

(b) female German-English speakers

(c) male Mandarin-English speakers

(d) female Mandarin-English speakers

Figure 5.9 – MCD with respect to the number of transforms. A cross refers to the minimum and a circle refers to the transform generation threshold being equal to 2460.

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 AV-ENG-UK towards each of the test speakers.

Four settings (S-m1, S-m2, S-r1 and S-r2) were evaluated in the speaker-independent experiments that are described in Table 5.7. $\varepsilon_{\Delta\text{MCD}}$ was set to 0.0005dB. The four settings in Table 5.7 with a grey background were used as system contrasts.

5.5.4 Objective Evaluation

Original recordings of development and test data of the 20 speakers were aligned using the English average voice AV-ENG-UK and speech samples for objective evaluation were synthe-

Table 5.7 – Settings of speaker-independent experiments

System ID	State mapping construction	Regression class tree growth
S-m1	proposed approach	global transform
C-m1	minimum KLD criterion	
S-m2	proposed approach	decision trees from AV-ENG-UK[†]
C-m2	minimum KLD criterion	
S-r1	minimum KLD criterion	proposed approach
C-r1		global transform
S-r2	proposed approach	proposed approach
C-r2		global transform

[†] The transform generation threshold was set to 2460.

sized using resultant durations. Results of objective evaluation on the development data set are presented in Tables 5.8 and 5.9.

Table 5.8 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 5.9 shows that the proposed jointly data-driven and phonological knowledge-guided 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 Section 5.5.2 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 Figures 5.10, 5.11, 5.12 and 5.13, 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.

Figures 5.10 and 5.11 confirm that the best solution in the case of Mandarin and English

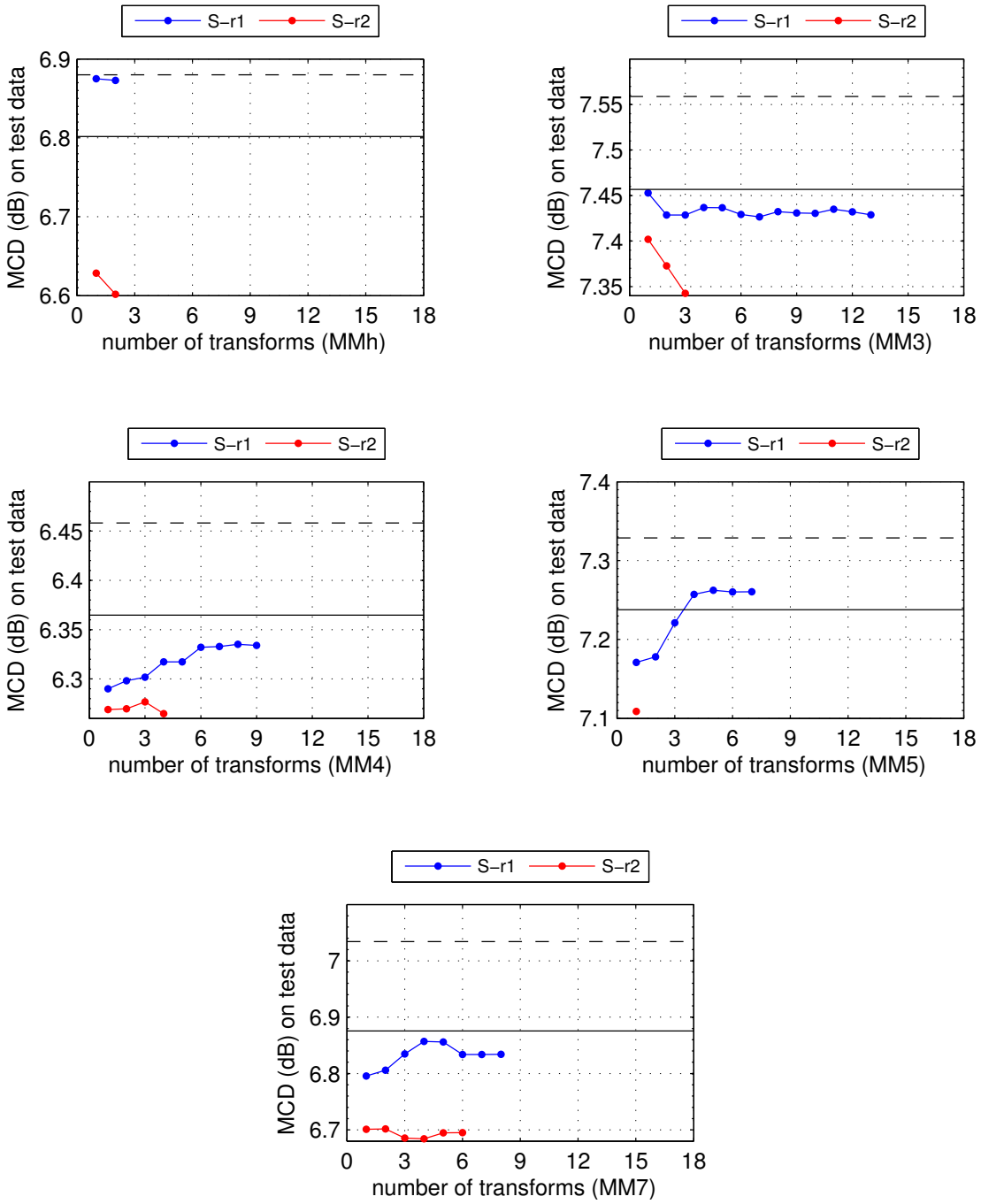


Figure 5.10 – MCD measurements in relation to the number of transforms in various conditions. The five speakers are male Chinese. The leftmost point on each red curve indicates the result of S-m1 and the leftmost point on each blue curve indicates the result of C-m1. The solid black horizontal lines indicate the results of S-m2 and the dashed black horizontal lines indicate the results of C-m2.

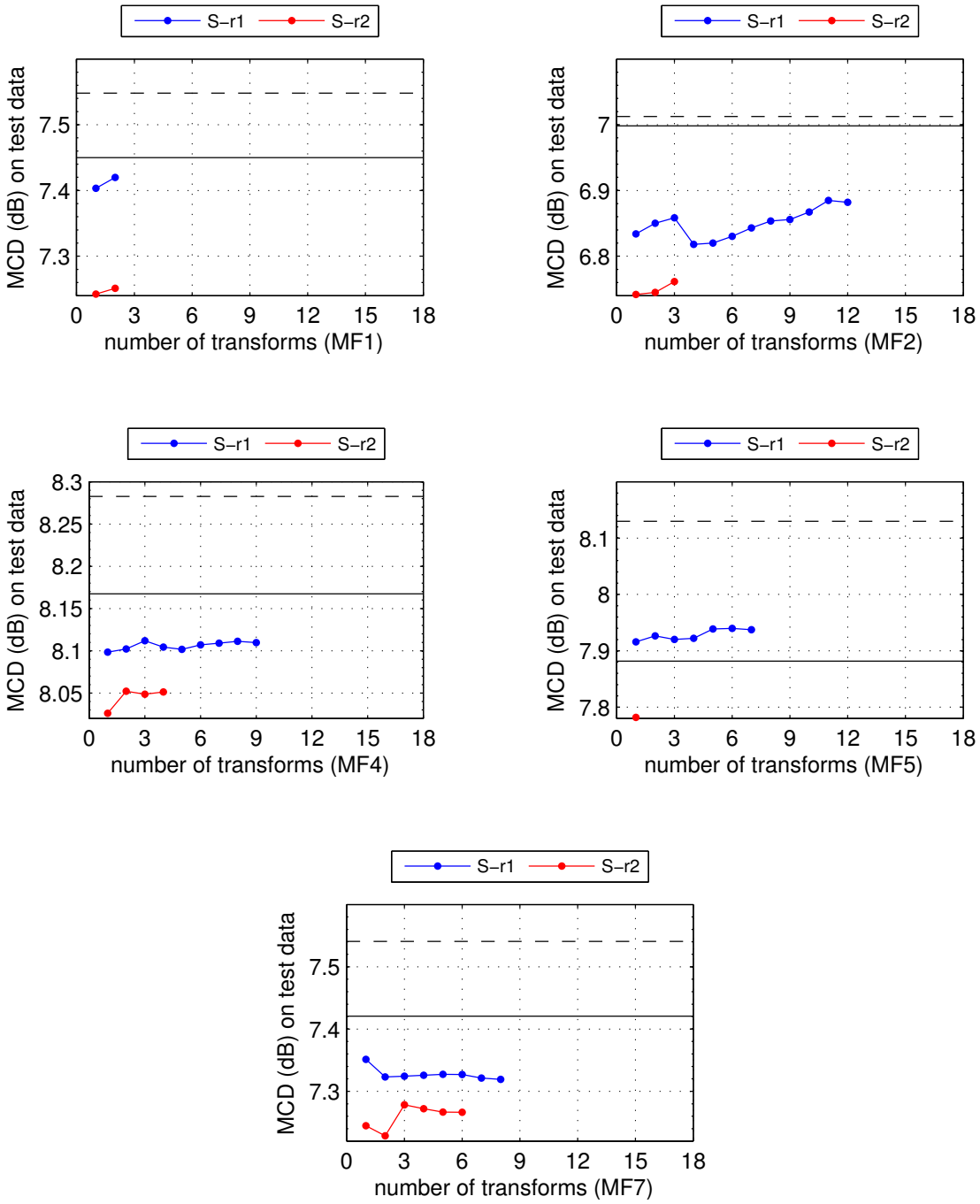


Figure 5.11 – MCD measurements in relation to the number of transforms in various conditions. The five speakers are female Chinese. The leftmost point on each red curve indicates the result of S-m1 and the leftmost point on each blue curve indicates the result of C-m1. The solid black horizontal lines indicate the results of S-m2 and the dashed black horizontal lines indicate the results of C-m2.

Table 5.8 – MCD (dB) on the development data of the training partition & the percentage of mapping rules that remained unchanged

Language	L_{in} = German, L_{out} = British English					
Training speaker groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	average
C-m1	6.04	6.13	6.08	6.07	6.08	6.08
S-m1	5.93	6.04	5.98	6.00	5.99	5.99
difference	0.11	0.09	0.10	0.07	0.09	0.09
% of unchanged mappings	50.2%	56.8%	45.5%	49.3%	52.1%	50.8%
C-m2	5.93	6.04	6.00	5.99	6.00	5.99
S-m2	5.82	5.94	5.88	5.91	5.92	5.89
difference	0.11	0.09	0.12	0.09	0.08	0.10
% of unchanged mappings	54.4%	47.6%	45.5%	54.2%	60.0%	52.3%
Language	L_{in} = Mandarin, L_{out} = British English					
Training speaker groups	6-7-8-9	6-7-8-0	6-7-9-0	6-8-9-0	7-8-9-0	average
C-m1	7.07	7.09	7.04	7.06	7.08	7.07
S-m1	6.96	6.97	6.91	6.93	6.97	6.95
difference	0.11	0.12	0.13	0.13	0.10	0.12
% of unchanged mappings	39.4%	25.6%	29.3%	35.7%	22.8%	30.6%
C-m2	7.19	7.22	7.17	7.19	7.23	7.20
S-m2	7.06	7.08	6.99	7.02	7.10	7.05
difference	0.13	0.14	0.18	0.17	0.13	0.15
% of unchanged mappings	41.7%	46.1%	41.7%	47.5%	42.4%	43.9%

was achieved by only applying the jointly data-driven and phonological knowledge-guided 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 between Mandarin and English, it is not surprising that the quality of synthesized English is degraded immediately after the number of adaptation transforms grew.

As for German and English, Figures 5.12 and 5.13 show that the proposed jointly data-driven and phonological knowledge-guided approach can be applied to state mapping construction

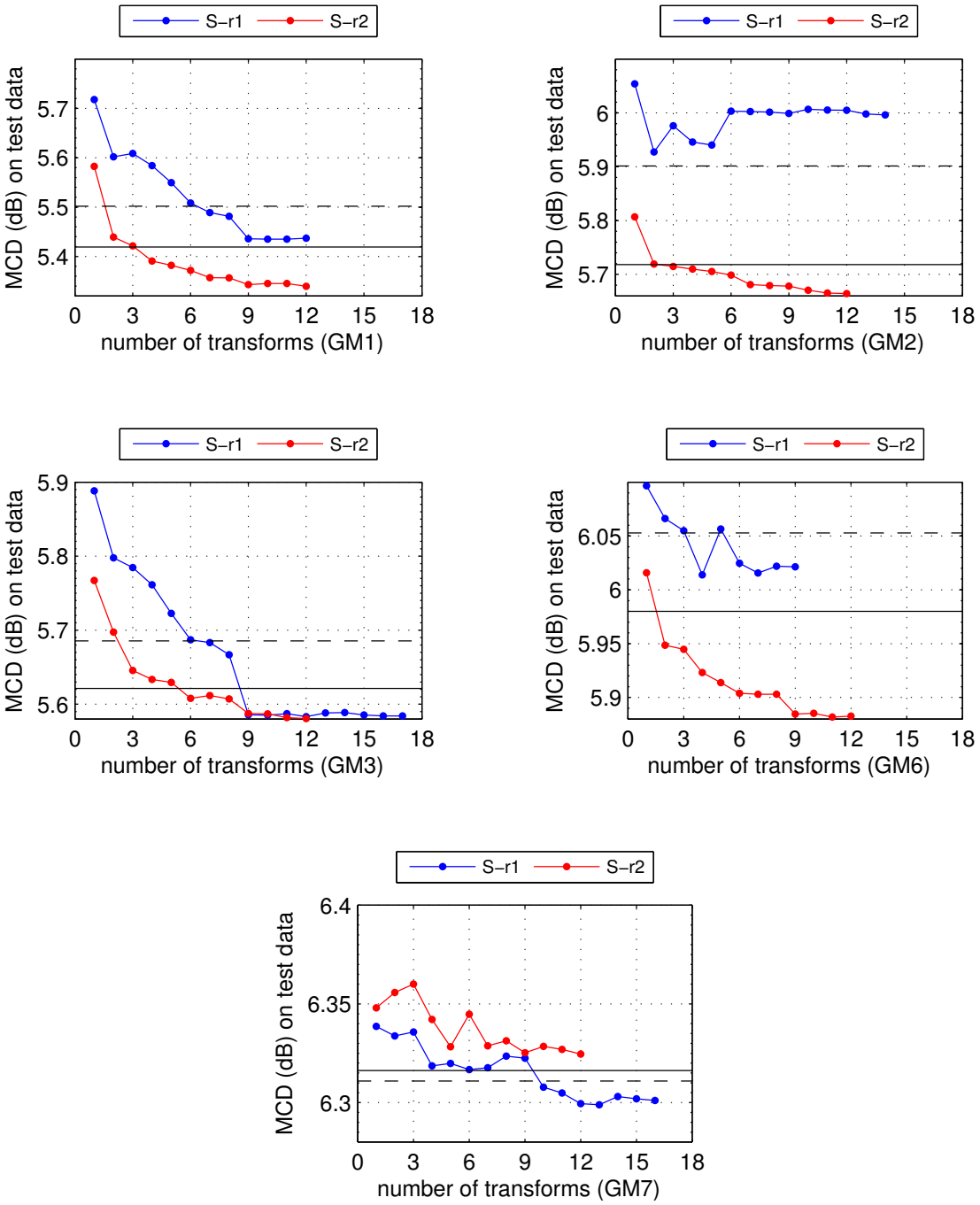


Figure 5.12 – MCD measurements in relation to the number of transforms in various conditions. The five speakers are male Germans. The leftmost point on each red curve indicates the result of S-m1 and the leftmost point on each blue curve indicates the result of C-m1. The solid black horizontal lines indicate the results of S-m2 and the dashed black horizontal lines indicate the results of C-m2.

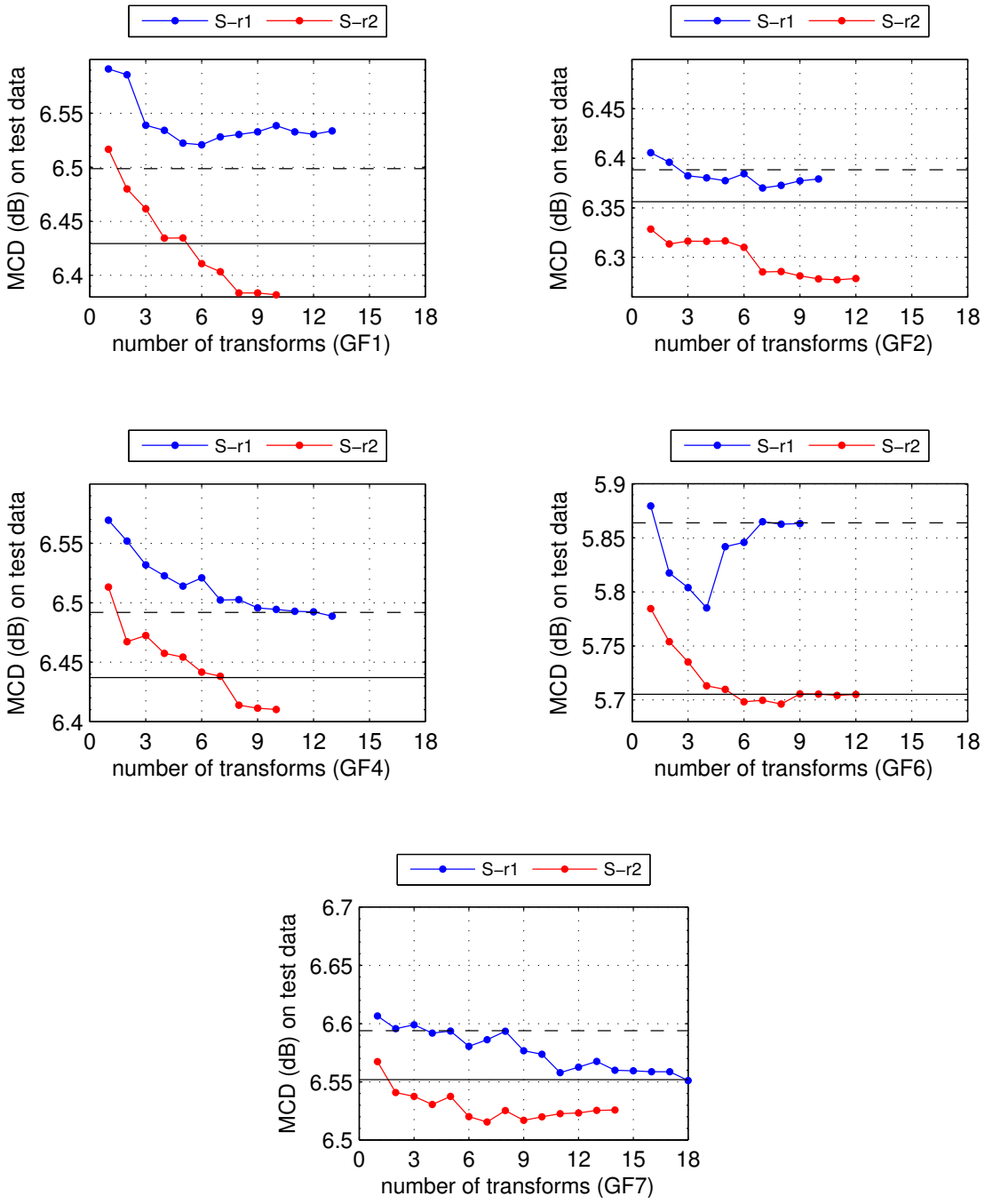


Figure 5.13 – MCD measurements in relation to the number of transforms in various conditions. The five speakers are female Germans. The leftmost point on each red curve indicates the result of S-m1 and the leftmost point on each blue curve indicates the result of C-m1. The solid black horizontal lines indicate the results of S-m2 and the dashed black horizontal lines indicate the results of C-m2.

Table 5.9 – MCD (dB) on the development data of the training partition & the number of regression class tree leaves

Language	L_{in} = German, L_{out} = British English					
Training speaker groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	average
C-r1	6.04	6.13	6.08	6.07	6.08	6.08
S-r1	5.87	6.00	5.94	5.93	5.95	5.94
difference	0.17	0.13	0.14	0.14	0.14	0.14
# of regression classes	19	9	18	14	14	14.8
C-r2	5.93	6.04	5.98	6.00	5.99	5.99
S-r2	5.79	5.92	5.86	5.86	5.87	5.86
difference	0.15	0.12	0.13	0.13	0.12	0.13
# of regression classes	14	12	12	12	12	12.4
Language	L_{in} = Mandarin, L_{out} = British English					
Training speaker groups	6-7-8-9	6-7-8-0	6-7-9-0	6-8-9-0	7-8-9-0	average
C-r1	7.07	7.09	7.04	7.06	7.08	7.07
S-r1	7.05	7.07	7.01	7.03	7.07	7.05
difference	0.02	0.02	0.03	0.03	0.01	0.02
# of regression classes	8	7	9	13	2	7.8
C-r2	6.96	6.97	6.91	6.93	6.97	6.95
S-r2	6.95	6.97	6.91	6.91	6.97	6.94
difference	0.01	0.00	0.01	0.02	0.01	0.01
# of regression classes	6	1	4	3	2	3.2

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. Figures 5.12 and 5.13 also show: (1) the MCD scores produced by applying the proposed approach to both state mapping construction and regression class tree growth (S-r2, the red curves) are more likely to decrease further than those produced by applying the proposed approach to regression class tree growth only (S-r1, the blue curves); (2) when using enhanced state mapping rules, enhanced regression class trees generated by the proposed approach (S-r2, the red curves) eventually produced MCD scores smaller than those the regression class tree following the decision tree structure of AV-ENG-UK produced (S-m2, 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.

5.5.5 Iterative Enhancement

The jointly data-driven and phonological knowledge-guided 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" hereinafter.
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" hereinafter.

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 5.10.

Table 5.10 – 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

Language	L_{in} = German, L_{out} = British English					
Training speaker groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	average
baseline†	5.79	5.92	5.86	5.86	5.87	5.86
using 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%
using 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 S-r2 (i.e., from the first iteration).

Then the 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 5.11.

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

Table 5.11 – MCD (dB) on the development data of the training partition and the number of regression class tree leaves after regression class tree growth in the second iteration

Language	L_{in} = German, L_{out} = British English					
Training speaker groups	1-2-3-4	1-2-3-5	1-2-4-5	1-3-4-5	2-3-4-5	average
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 S-r2 (i.e., from the first iteration).

for German and English.

5.5.6 Subjective Evaluation

Naturalness and speaker similarity of speech which was synthesized by the proposed jointly data-driven and phonological knowledge-guided approach applied to both state mapping construction and regression class tree growth (i.e., system S-r2) 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, C-m1 (i.e., using the minimum KLD criterion plus a single global transform) and C-m2 (i.e., using the minimum KLD criterion plus a regression class tree following the decision tree structure of AV-ENG-UK). Each listener was presented with 60 utterance pairs in total: 3 (pairs)×10 (test speaker groups)×2 (tests). The sentence of each pair was randomly selected from the 25 test sentences in DATA-TEST-ENG-25. 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 Figure 5.14.

Firstly, it is noted that the jointly data-driven and phonological knowledge-guided approach mainly improved naturalness of synthesized speech in the speaker-independent experiments, as observed in the previous speaker-dependent experiments in Section 5.4. Thinking back on the speaker discrimination experiments in Section 3.5.3, we hypothesize that a limiting factor in these experiments is the quality of speech generated by cross-lingual speaker adaptation, which hinders listeners' judgement of speaker identity.

Secondly, it is observed that applying the proposed approach to both state mapping construction and regression class tree growth produced a significantly better system than using the minimum KLD criterion and a regression class tree following the decision tree structure for state tying. The proposed approach can automatically generate a suitable regression class tree structure for cross-lingual speaker adaptation so that input language-specific information

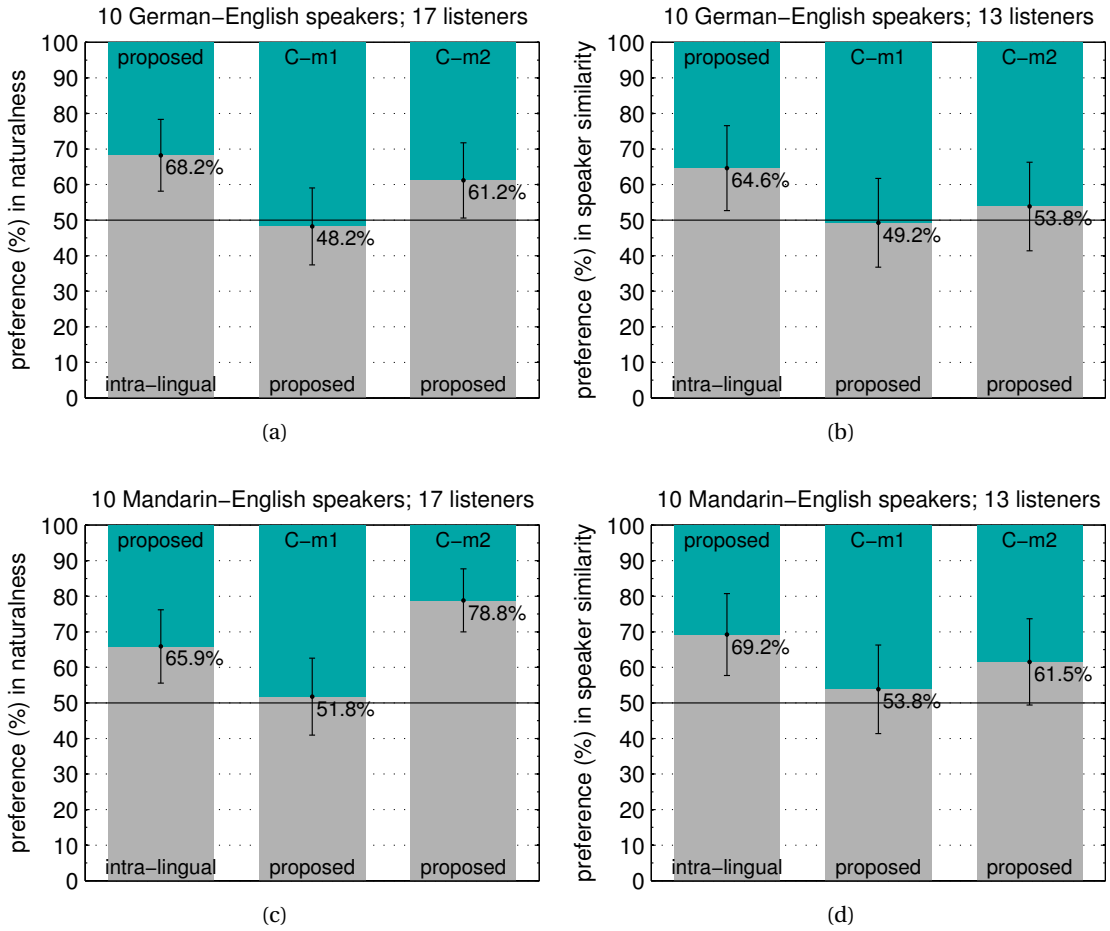


Figure 5.14 – Results of subjective evaluations on the jointly data-driven and phonological knowledge-guided approach. Whiskers indicate 95% confidence intervals.

from adaptation data can be suppressed as much as possible. The contrast between Figures 5.14a and 5.14c appears to suggest that the jointly data-driven and phonological knowledge-guided approach is more effective for a pair of languages which are more phonologically dissimilar.

Lastly, Figure 5.14 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 jointly data-driven and phonological knowledge-guided approach alleviated some of the negative effects.

5.6 Conclusions

A jointly data-driven and phonological knowledge-guided approach was proposed in this chapter. It was applied to HMM state mapping construction such that phonologically in-

consistent 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.

In addition, it is noted that given sufficient amount of adaptation data, the number of transforms that produces the smallest MCD in HMM state mapping-based cross-lingual speaker adaptation may be a measure of the phonological/acoustic similarity between two languages. This hypothesis needs to be examined once bilingual speech data in other language pairs are available.

The contribution presented in this chapter was originally published in the following papers:

- Hui LIANG and John DINES, "Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation", *Proc. of Interspeech*, pp. 1825–1828, August 2011.
- Hui LIANG and John DINES, "Jointly Data-Driven and Phonological Knowledge-Guided Enhancement of State Mapping Based Cross-Lingual Speaker Adaptation", submitted to *IEEE Transactions on Audio, Speech and Language Processing*.

6 Hierarchical Transformation Framework

A data-driven and phonological knowledge-guided approach was proposed in Chapter 5 to tackle the language mismatch between average voice synthesis models and adaptation data by enhancing the processes of HMM state mapping construction and regression class tree growth. While providing improvements, experiments showed that further effort is necessary in order to achieve the performance of intra-lingual adaptation in cross-lingual scenarios. Since the findings in Chapter 4 led to the conclusion that new techniques that model speaker characteristics and inherent differences between languages separately should be introduced into cross-lingual speaker adaptation, research on this direction is conducted in this chapter. In particular, a two-layer hierarchical transformation framework is investigated.

In this chapter, cross-lingual speaker adaptation experiments are conducted by *data mapping* through state mapping rules constructed under the minimum K-L divergence criterion. In case a regression class tree is involved, it follows the decision tree structure from state tying [Yamagishi et al., 2004]. In order to simplify the analysis, state mapping rules and regression class trees generated by the jointly data-driven and phonological knowledge-guided approach are not incorporated, though it would be trivial to do so.

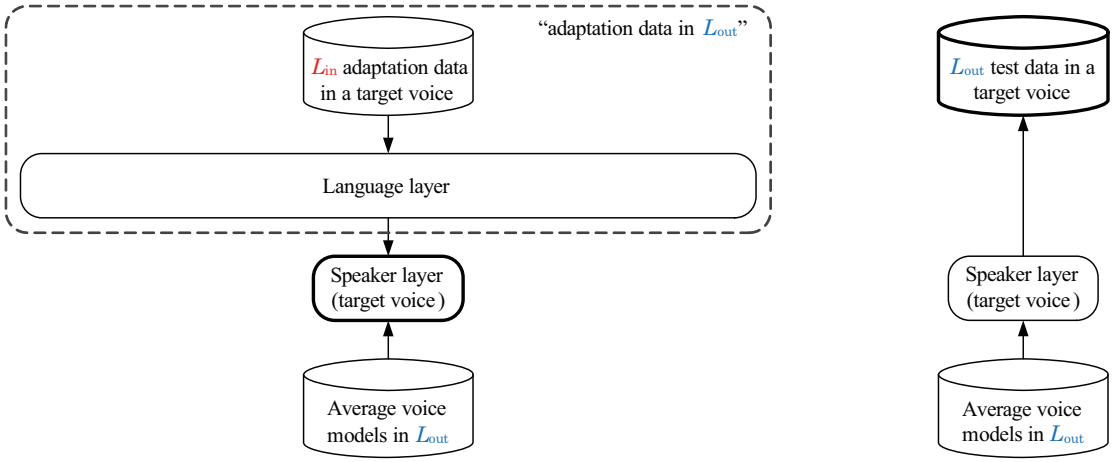
6.1 Two-Layer Hierarchy

Relevant work has been carried out. For example, the speaker and language factorization technique proposed in [Zen et al., 2012] is effectively a two-layer hierarchical transformation framework. However, it involves cluster adaptive training and cluster-dependent decision trees besides CMLLR. It is of interest whether language and speaker characteristics can be captured separately using only linear transforms (CMLLR or CSMAPLR) for cross-lingual speaker adaptation.

Previous work confirmed the possibility of the separation of speaker characteristics from age (an adult voice to a child voice) [Karhila et al., 2012], accent [Smit and Kurimo, 2011], or environmental characteristics [Seltzer and Acero, 2011] using only CMLLR/CSMAPLR

transforms. Their common hierarchy operates in a way that in the training stage, transforms of the non-speaker layer were estimated alone first and then employed as parent transforms to estimate those of the speaker layer; in the recognition/synthesis stage, the transforms of the speaker layer were applied to recognition/synthesis models first, then those of the non-speaker layer were applied to the adapted models, and finally recognition/synthesis was performed with the twice-adapted models.

In order to improve state mapping-based cross-lingual speaker adaptation, Peng et al. proposed to estimate a global transform which minimized the K-L divergence between distributions of average voice synthesis models in the input and output languages, aiming to compensate for the actual differences in terms of voice characteristics between the two model sets [Peng et al., 2010]. At the synthesis stage, the global transform was applied to the average voice models in the output language before target speaker-specific transforms obtained in the intra-lingual manner on the side of the input language were applied. Although the global transform was meant to compensate for the difference of voice characteristics, the language mismatch between the two sets of average voice models was also captured in the global transform. The fact that their baseline built by normal transform mapping [Wu et al., 2009] outperformed the proposed approach, presumably, implies that the average voice models in the output language were adapted towards the input language by the global transform in the synthesis stage. Their work provides a clue that the layer handling language mismatch should be probably involved only in the training stage, i.e., the language characteristics of average voice models in the output language should be maintained in the synthesis stage. This clue results in a distinction from the hierarchy employed in [Karhila et al., 2012, Smit and Kurimo, 2011, Seltzer and Acero, 2011], where there were still two respective layers for speaker and age/accent/environment characteristics in the recognition/synthesis stage.



(a) Step 2 – speaker layer training by CMLLR/CSMAPLR

(b) Step 3 – parameter generation

Figure 6.1 – Two-layer hierarchy for cross-lingual speaker adaptation

The previous work discussed above has made the two-layer hierarchy for cross-lingual speaker adaptation clearer, especially how it should function in the stages of speaker transform training and parameter generation in a target voice. The two steps are illustrated in Figure 6.1.

6.2 Language Layer Training

The key problem is how to estimate transforms of the language layer. Namely, how to achieve Step 1 is the goal of this section.

6.2.1 Direct Estimation

The work in [Smit and Kurimo, 2011] involved an approach whereby transforms of the accent layer were trained over speaker-independent models and accented data from multiple speakers, using a large number of regression classes. The purpose of the accent layer is synonymous with the estimation of transforms of the language layer in this chapter, as illustrated in Figure 6.2.

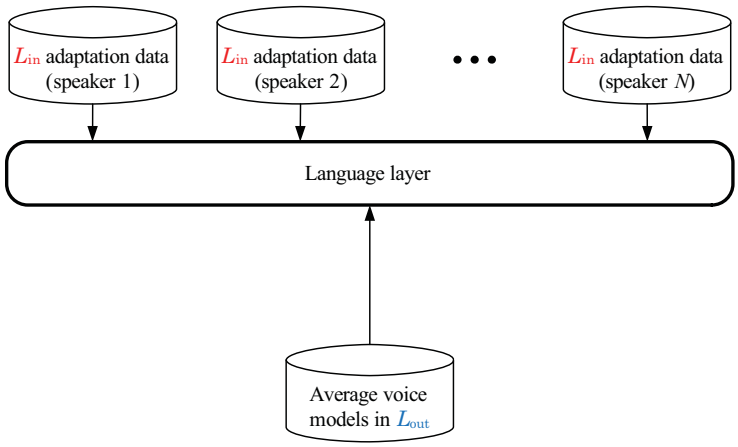


Figure 6.2 – Direct language layer training by CMLLR/CSMAPLR

In order to verify the feasibility of this approach, a system was built according to the following steps:

1. Construct state mapping rules based on AV-ENG-US and AV-CMN-sc;
2. Estimate transforms of the language layer by CSMAPLR over AV-ENG-US and SPEECON (the training data of AV-CMN-sc) using the resultant mappings;
3. Estimate speak-specific transforms using DATA-ADP-CMN-100 and DATA-TEST-ENG-25 in MF2's, MMh's, MM3's and MF7's voices, and synthesize speech with adapted models, as Figure 6.1 shows.

Another two systems were also built for comparison: (1) adapting AV-ENG-US with DATA-DEV-ENG-100 in the intra-lingual fashion; (2) adapting AV-ENG-US with DATA-ADP-CMN-100

by the normal data mapping approach. All the systems used regression class tree-based adaptation (i.e., multiple transforms were estimated) and the same regression class tree was employed for both layers and all the systems. As SPEECON is quite a large corpus, 1018 transforms of the language layer were generated for mel-cepstrum.

Table 6.1 – Mel-cepstral distortion (dB) comparison in direct estimation of the language layer

Speaker ID	MF2	MMh	MM3	MF7
hierarchical (Fig. 6.2)	8.38	8.63	8.87	9.43
data mapping	7.78	7.67	8.21	8.39
intra-lingual	6.65	6.32	7.41	7.51

Objective evaluation results can be found in Table 6.1. The MCD measurements indicate that directly estimating transforms of the language layer is not appropriate. In fact, what was captured by such a language layer is not clear. This is not an issue in the work in [Smit and Kurimo, 2011]. Since their accent layer was applied to models in the recognition stage, it was not necessary to fully factorize speaker and accent information. In the case of cross-lingual speaker adaptation, the language layer should capture only language characteristics in order that the speaker layer estimated in Step 2 can be applied independently of the language layer for synthesis in Step 3.

6.2.2 Estimation in a Speaker-Adaptive Fashion

Compared with Figure 6.2, the speaker layer is added into Figure 6.3. Estimating the language layer in a speaker-adaptive fashion as shown in Figure 6.3 could be of help since it is more likely that the language layer in Figure 6.3 captures only language characteristics.

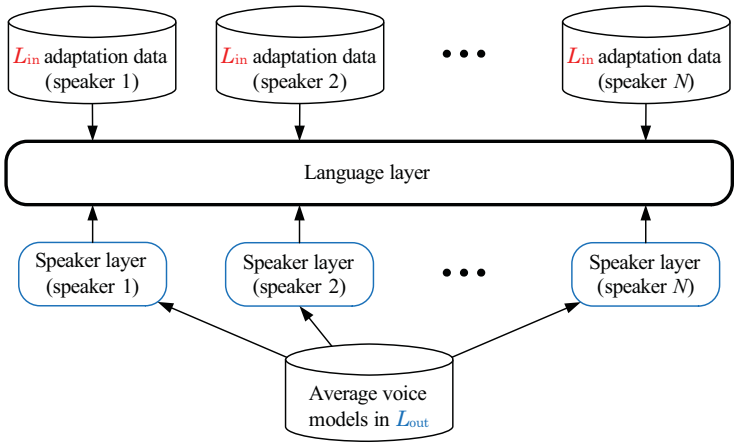


Figure 6.3 – Language layer training by CMLLR/CSMAPLR in a speaker-adaptive fashion. Blue rounded rectangles denote speaker-specific transforms obtained in intra-lingual speaker adaptation in the output language.

Two difficulties are present in this speaker-adaptive approach: (1) it requires speaker-specific transforms estimated in intra-lingual speaker adaptation in the output language, thereby a bilingual corpus in N training speakers' voices being necessary; (2) the language layer needs to be estimated on adapted synthesis models with full covariance matrices. In the initial research in this chapter, the second difficulty is avoided by diagonalizing the full covariance matrices, though it is possible to use full covariance matrices directly [Ghoshal et al., 2010].

In order to verify the feasibility of this approach, a simple experiment was conducted: rather than using adaptation data from N speakers, only the test speaker in Step 2 was involved for training the language layer. As a result, it should make the performance of the hierarchical transformation framework very close to that of intra-lingual speaker adaptation on this test speaker's data.

This experiment was conducted using average voice synthesis models AV-ENG-US and AV-CMN-sc as well as speech data DATA-ADP-CMN-100 and DATA-DEV-ENG-100 in MMh's voice, according to the steps shown in Figure 6.3 and Figure 6.1. This time only global transforms were estimated by CSMAPLR for both layers. Mel-cepstral distortion was calculated on his data DATA-TEST-ENG-25 (see Table 6.2).

Table 6.2 – Mel-cepstral distortion (dB) comparison in speaker-dependent estimation of the language layer

Speaker ID	MMh
data mapping	7.44
hierarchical (Fig. 6.3)	6.84
intra-lingual	6.54

It is clear that this approach is comparable to intra-lingual speaker adaptation even though diagonalization of adapted covariance matrices was employed. Basically this result suggests that the two-layer hierarchical transformation framework for cross-lingual speaker adaptation should be constructed with three steps which are illustrated in Figures 6.3, 6.1a and 6.1b respectively.

6.2.3 Speaker-Independent Estimation

The above experiments has helped to determine how a two-layer hierarchy for cross-lingual speaker adaptation should function. In order to obtain a speaker-independent language layer that works for target speakers unseen in the training data of the language layer, multiple training speakers are needed as Figure 6.3 shows.

DATA-ADP-CMN-100 in ten speakers' voices (MM1, MM3, MM4, MM5, MM7, MF1, MF2, MF4, MF5 and MF7, thus gender-balanced) was used for language layer training according to Figure 6.3. After transforms of the language layer were obtained, DATA-ADP-CMN-100 and DATA-

TEST-ENG-25 in five speakers' voices (MM3, MF2, MF7, MMh and MM6 [1]) were used for target speaker-specific transform estimation and speech parameter generation according to Figure 6.1. The adaptation algorithm was CSMAPLR. The two layers shared the same regression class tree. Objective evaluation results are presented in Table 6.3.

Table 6.3 – Mel-cepstral distortion (dB) comparison in speaker-independent estimation of the language layer

number of transforms of the language layer for MCEP		1	136	691	930	intra-lingual	data mapping
MF2	Global	7.19	6.99	6.91	**6.89**	6.81	7.61
	RegTree	7.49	7.26	6.96	**6.89**	6.65	7.78
MM3	Global	7.99	7.88	7.81	**7.80**	7.62	8.14
	RegTree	8.17	8.10	7.82	**7.75**	7.41	8.21
MF7	Global	8.17	8.01	7.89	**7.88**	7.60	8.39
	RegTree	8.31	8.29	8.01	**7.96**	7.51	8.39
MMh	Global	**7.37**	7.40	7.41	7.43	6.54	7.44
	RegTree	7.69	7.75	**7.67**	7.68	6.32	7.67
MM6	Global	7.78	**7.76**	7.77	7.78	7.17	7.72
	RegTree	7.98	8.03	**7.88**	7.88	6.87	7.80

[a] "Global" means only one transform was estimated for mel-cepstrum for each speaker in Step 2. "RegTree" means regression class-specific transforms were estimated for mel-cepstrum for each speaker in Step 2.

The test speakers, MF2, MM3 and MF7, were actually training speakers of the language layer. Table 6.3 shows that the language layer was beneficial to their voices and could make the performance of cross-lingual speaker adaption comparable to that of intra-lingual adaptation. In addition, their MCD measurements consistently decrease as transforms of the language layer become more specific. This demonstrates once again that the two-layer hierarchy designed in Figures 6.3 and 6.1 should be appropriate.

For MMh and MM6 who were not present in Step 1, Table 6.3 shows that transforms of the language layer had neither positive nor negative impact. Presumably, the lack of generalization is caused by the limited number of training speakers for estimation of the language layer. It is also possible that there may not exist a set of universal transforms of the language layer that applies to every single target speaker and it would be necessary to select training speakers of the language layer carefully. In any case, speech data needs to be recorded from a large number of bilingual speakers for more in-depth analysis.

1. MM6's spoken English is heavily accented but there are only a limited number of good bilingual speakers in this bilingual corpus. So objective evaluation was still performed on his English test data.

6.3 Summary

Initial research was conducted into the hierarchical transformation framework for state mapping-based cross-lingual speaker adaptation. A two-layer hierarchy was designed, where one layer captures target speaker-specific characteristics and the other compensates for the mismatch between the input and output languages. This hierarchy was found to be promising to make the performance of cross-lingual speaker adaptation comparable to that of intra-lingual adaptation. Unfortunately, due to the shortage of bilingual speakers, especially fluent bilingual speakers with natural-sounding accents in both languages, an optimal method for the estimation of speaker-independent transforms of the language layer has not yet been confirmed. Further investigation will be required using a corpus with a significant number of good bilingual speakers.

The experimental results presented in this chapter have not yet been published elsewhere.

7 Conclusions

First of all, experiments were conducted in the thesis to investigate (i) the ability of people to discriminate between speakers across languages, (ii) unsupervised cross-lingual speaker adaptation and (iii) the effect of the inherent problem of language mismatch on state mapping-based cross-lingual speaker adaptation. Then a data-driven and phonological knowledge-guided approach for alleviating the negative effect of language mismatch was proposed. Finally, a two-layer hierarchy aimed at capturing speaker characteristics and language information separately was examined. The original research work is summarized below.

7.1 Summary of Contributions

The main contributions of the thesis work to the state of the art of cross-lingual speaker adaptation for speech synthesis include the following:

(1) *Exploring the ability of people to distinguish between speakers across different languages*

Firstly, the ability of people to distinguish between speakers across different languages was explored in this thesis. Experimental results show that the difference in language between two utterances leads to additional difficulty in discriminating between speakers, in comparison to the intra-lingual setting without such difference. Aside from that, the quality of synthesized speech is found to play a significant role in distinguishing between speakers. It leads to even more noticeable difficulty than language difference. Therefore it is concluded that differentiating between speakers across languages is an achievable task, but this becomes very difficult in the context of personalized speech-to-speech translation for the moment (i.e., when difference in language is combined with that in speech type), given the current quality of speech synthesized through cross-lingual speaker adaptation. Thus the main future research should be focused on how to improve synthesis quality.

(2) *Examining unsupervised cross-lingual speaker adaptation for personalized speech-to-speech translation*

The possibility of employing cross-lingual speaker adaptation in the unsupervised fashion was investigated. Both objective and subjective evaluation results demonstrate that the performance of unsupervised cross-lingual speaker adaptation is comparable to that of supervised cross-lingual speaker adaptation. Hence the major difficulty in building a personalized speech-to-speech translator does not lie in the use of unsupervised adaptation.

(3) *In-depth analysis of the impacts of the language mismatch between adaptation data and synthesis models*

The impacts of undesirable language information that adaptation transforms capture as a result of the language mismatch between adaptation data and average voice synthesis models were analyzed. The HMM state mapping technique requires two sets of average voice synthesis models in the input and output languages, respectively. Depending on how to utilize HMM state mapping rules, adaptation transforms can be estimated over synthesis model distributions in either language. Meanwhile, a regression class tree can be also derived from either language. Experimental results show that it is preferable to estimate transforms directly over synthesis model distributions in the output language. The language from which a regression class tree is derived appears to be of secondary importance.

It is also revealed that there appears to be little advantage to estimating multiple adaptation transforms via a regression class tree. In contrast, regression class-specific adaptation transforms are actually detrimental to the performance of cross-lingual speaker adaptation. The greater number of regression class-specific transforms that are generated, the greater the degradation to adaptation performance. It can be concluded that language information needs to be eliminated or reduced from transforms that are meant for speaker adaptation only.

(4) *Jointly data-driven and phonological knowledge-guided enhancement under the minimum generation error criterion*

The approach of jointly data-driven and phonological knowledge-guided enhancement under the minimum generation criterion was proposed in this thesis. It was applied to both HMM state mapping construction and regression class tree growth. This approach guarantees that state mapping rules are always meaningful in the phonological sense and automatically generates a regression class tree with an appropriate structure. The minimum KLD criterion is found to be sub-optimal for state mapping-based cross-lingual speaker adaptation and it is observed that its reliability depends on the phonological/acoustic similarity between the input and output languages. The usefulness of a regression class tree in cross-lingual speaker adaptation for speech synthesis is observed being also dependent on the phonological/acoustic similarity between the input and output languages. This gives some insight into the phonological/acoustic similarity of two languages. Furthermore, improved experimental results (i.e., MCD reductions) demonstrate the effectiveness and generalization across speakers of the proposed approach when it is applied to HMM state mapping construction and regression class tree growth.

(5) *Two-layer hierarchical transformation framework*

A two-layer hierarchical transformation framework was developed. The two layers of linear transforms are applied in such a way as to capture speaker characteristics and language information respectively. How this hierarchy should operate was investigated and determined: firstly, estimate transforms in the intra-lingual manner on the output language side; secondly, estimate the language layer based on adaptation data in the input language and synthesis models adapted by these intra-lingual transforms; thirdly, estimate target speaker-specific transforms with those of the language layer used as parent transforms; lastly, synthesize speech using only the target speaker-specific transforms. Consequently, the challenge is restricted to the estimation of a speaker-independent language layer.

7.2 Limitations and Future Work

Apart from the contributions mentioned above, some limitations of the thesis work can be noted in the previous chapters. They could be considered directions of future research related to cross-lingual speaker adaptation.

Firstly, the thesis work was focused on cross-lingual adaptation of spectral features, since spectrum was considered the dominant aspect that contributed to speaker identity. In fact, prosodic patterns of a particular speaker in different languages may share common traits and thus contribute to speaker identity, although each language has its own prosodic patterns. It is worth investigating cross-lingual adaptation of prosodic features, more specifically, pitch and duration.

Secondly, only 21 questions were involved in the proposed data-driven and phonological knowledge-guided approach. This question set can be extended. For example, unlike categories concerning consonants, there was only one category with respect to vowels. It is worth investigating how to split this vowel category into finely grained ones according to articulatory features (mostly the tongue and lip positions), i.e., how to partition the vowel quadrilateral appropriately. Furthermore, this approach may be applied to cross-lingual adaptation of prosodic features as well.

Thirdly, the data-driven and phonological knowledge-guided approach did not considerably alleviate the negative effect of language mismatch. As it has been concluded earlier, language information needs to be separated from transforms which are meant for speaker adaptation only. A two-layer, hierarchical adaptation framework that captures language information and speaker characteristics by separate sets of linear transforms deserves to be investigated. Currently how such a hierarchy should be established and trained has been determined based on a limited number of good bilingual speakers. The estimation of a speaker-independent language layer needs to be further investigated in the future when a larger bilingual corpus containing more speakers with natural-sounding accents in both languages is available.

Fourthly, the state-of-the-art techniques for HMM state mapping construction has been always based on the assumption that the two sets of average voice synthesis models in the input and output languages respectively have identical voice characteristics and overlapping model space. This assumption is scarcely true, since the training procedure of average voice synthesis models in the EM fashion cannot guarantee such consistency, which highly depends on the method of model initialization and training corpora themselves. Research that addresses the inconsistency between two sets of average voice synthesis models deserves to be undertaken.

Lastly, the experiments on human perception of speaker identity in this thesis were mainly focused on listeners' perception of other speakers' voices. It would be interesting to take into account listeners' perception of their own voices.

A Appendix – Phonemes and Their Categories for Question Design

Each phoneme of all the four languages (or accents) was considered to belong to one of the seven categories: silence, vowel, plosive, fricative, nasal, affricate and approximant. Questions for HMM state mapping construction and regression class tree growth using the minimum generation error criterion were designed according to the phoneme-category relationship.

A.1 American English

57 phonemes were employed for American English and they cover all the seven categories [Fitt, 2000].

Table A.1 – Phonemes in American English and their categories

Unilex Symbol	Word Example	IPA Symbol	Category
@	**a**bout	/ə/	vowel
#	(a period of silence)	—	silence
a	m**a**p	/æ/	vowel
aa1	c**o**ck	/ɑ/	vowel
aer1	requ**i**re	/arˠ/	vowel
ai	l**i**ne	/aɪ/	vowel
ar	p**a**rty	/ɑˠ/	vowel
b	**b**oat	/b/	plosive
ch	**ch**eese	/ʧ/	affricate
d	**d**oes	/d/	plosive
dh	**th**is	/ð/	fricative
e	dr**e**ss	/e/	vowel
eh	m**a**n	/æ/	vowel
ei1	m**a**ke	/eɪ/	vowel
eir1	wh**e**re	/eˠ/	vowel
f	**f**ont	/f/	fricative

Table A.1 – Phonemes in American English and their categories (continued)

Unilex Symbol	Word Example	IPA Symbol	Category
g	**g**un	/ɡ/	plosive
h	**h**air	/h/	fricative
hw	**wh**ite	/ʍ/	fricative
i	k**i**d	/ɪ/	vowel
ir	n**ear**	/ɪˠ/	vowel
iy	cit**y**	/i/	vowel
jh	en**g**ine	/ʤ/	affricate
k	**c**at	/k/	plosive
l	enc**l**ose	/l/	approximant
l!	ab**l**e	/l̩/	approximant
lw	fee**l**	/ɬ/	approximant
m	**m**ark	/m/	nasal
m!	multilingualis**m**	/m̩/	nasal
n	**n**ot	/n/	nasal
n!	heave**n**	/n̩/	nasal
ng	si**ng**	/ŋ/	nasal
oi	b**oy**	/ɔɪ/	vowel
oo1	w**a**ter	/ɒː/	vowel
or	h**o**rse	/ɔˠ/	vowel
ou1	g**oa**t	/oʊ/	vowel
ow	h**ou**se	/aʊ/	vowel
owr1	h**our**	/aʊˠ/	vowel
p	**p**urr	/p/	plosive
pau	(a short pause)	—	silence
r	**r**oad	/ɹ/	approximant
@r	wat**e**r	/əˠ/	vowel
@@r1	n**u**rse	/ɜˠ/	vowel
s	**s**et	/s/	fricative
sh	**sh**oe	/ʃ/	fricative
t	**t**ooth	/t/	plosive
tˆ	be**tt**er	/ɾ/	plosive
th	**th**ank	/θ/	fricative
u	p**u**t	/ʊ/	vowel
uh	l**o**ve	/ʌ/	vowel
ur1	j**u**ry	/ʊˠ/	vowel
uw	f**oo**d	/uː/	vowel
v	**v**ote	/v/	fricative
w	**w**et	/w/	approximant
y	**y**es	/j/	approximant

Table A.1 – Phonemes in American English and their categories (continued)

Unilex Symbol	Word Example	IPA Symbol	Category
z	**z**oo	/z/	fricative
zh	u**s**ually	/ʒ/	fricative

A.2 Mandarin

51 phonemes were employed for Mandarin and they cover all the seven categories. This Mandarin phoneme set was kindly provided by Nokia, a partner of the EMIME project.

Table A.2 – Phonemes in Mandarin and their categories

Symbol	Pinyin/Character Example	IPA Symbol	Category
A	ji**a**ng (江)	/ɑ/	vowel
a	l**a**i (來)	/a/	vowel
a2	f**a** (發)	/aː/	vowel
a3	ku**a** (跨)	/ɑː/	vowel
ae	qu**a**n (圈)	/œ/	vowel
e	n**e**i (内)	/e/	vowel
E	qi**a**n (錢)	/ɛ/	vowel
E_r	bi**e** (別)	/ɛː/	vowel
f	**f**ang (放)	/f/	fricative
I	za**i** (在)	/ɪ/	vowel
i	m**i**n (民)	/i/	vowel
i:	l**i** (李)	/iː/	vowel
i2	s**i** (四)	/ɨː/	vowel
i3	sh**i** (是)	/ɨˬː/	vowel
j	**y**un (雲)	/j/	approximant
kh	**k**e (可)	/k^h/	plosive
Mk	**g**uo (國)	/k/	plosive
Ml	**l**i (李)	/l/	approximant
Mm	**m**in (民)	/m/	nasal
Mn	**n**ei (内)	/n/	nasal
Mp	**b**u (不)	/p/	plosive
Mt	**d**ui (對)	/t/	plosive
N	ji**ng** (江)	/ŋ/	nasal
n2	mi**n** (民)	/n̩/	nasal
o	b**o** (剝)	/ɔː/ [1]	vowel

1. right after an initial (mainly /p/, /p^h/, /m/ and /f/)

Table A.2 – Phonemes in Mandarin and their categories (continued)

Symbol	Pinyin/Character Example	IPA Symbol	Category
o2	tu**o** (脱)	/ɔː/ [2]	vowel
ph	**p**eng (彭)	/pʰ/	plosive
s	**s**uo (所)	/s/	fricative
s@	zh**e**n (真)	/ɤ/	vowel
s1	(when no initial exists) [3]	—	silence
s2	**x**ing (型)	/ɕ/	fricative
s3	**sh**uo (説)	/ʂ/	fricative
s7	z**e** (则)	/ɤː/	vowel
sil	(a period of silence)	—	silence
sp	(a short pause)	—	silence
s@r	**er** (爾)	/ɤ˞ː/	vowel
th	**t**ong (同)	/tʰ/	plosive
ts	**z**e (则)	/t͡s/	affricate
ts2	**j**ia (加)	/t͡ɕ/	affricate
ts3	**zh**ong (中)	/t͡ʂ/	affricate
tsh	**c**e (側)	/t͡sʰ/	affricate
tsh2	**q**ia (恰)	/t͡ɕʰ/	affricate
tsh3	**ch**ong (衝)	/t͡ʂʰ/	affricate
U	l**o**ng (龍)	/ʊ/	vowel
u	li**u** (劉)	/u/	vowel
u:	b**u** (不)	/uː/	vowel
w	**w**o (我)	/w/	approximant
x	**h**ao (好)	/x/	fricative
y	y**u**n (雲)	/y/	vowel
y:	j**u** (據)	/yː/	vowel
z2	**r**en (人)	/ʐ/	fricative

A.3 British English

52 phonemes were employed for UK English and they cover all the seven categories [Fitt, 2000].

Table A.3 – Phonemes in British English and their categories

Unilex Symbol	Word Example	IPA Symbol	Category
@	**a**bout	/ə/	vowel

2. after the glide /w/
3. This happens when a Pinyin transcription begins with “a”, “o” or “e”.

Table A.3 – Phonemes in British English and their categories (continued)

Unilex Symbol	Word Example	IPA Symbol	Category
#	(a period of silence)	—	silence
a	m**a**p	/æ/	vowel
aa	b**ar**	/ɑː/	vowel
ai	l**i**ne	/aɪ/	vowel
b	**b**oat	/b/	plosive
ch	**ch**eese	/ʧ/	affricate
d	**d**oes	/d/	plosive
dh	**th**is	/ð/	fricative
e	dr**e**ss	/e/	vowel
ei	m**a**ke	/eɪ/	vowel
eir	wh**er**e	/eə/	vowel
f	**f**ont	/f/	fricative
g	**g**un	/g/	plosive
h	**h**air	/h/	fricative
i	k**i**d	/ɪ/	vowel
i@	n**ear**	/ɪə/	vowel
ii	b**ee**	/iː/	vowel
iy	cit**y**	/i/	vowel
jh	en**g**ine	/ʤ/	affricate
k	**c**at	/k/	plosive
l	enc**l**ose	/l/	approximant
l!	ab**l**e	/l̩/	approximant
lw	fee**l**	/ɬ/	approximant
m	**m**ark	/m/	nasal
m!	multilingualis**m**	/m̩/	nasal
n	**n**ot	/n/	nasal
n!	heave**n**	/n̩/	nasal
ng	si**ng**	/ŋ/	nasal
o	l**o**t	/ɒ/	vowel
oi	b**oy**	/ɔɪ/	vowel
oo	h**or**se	/ɔː/	vowel
ou	g**oa**t	/əʊ/	vowel
ow	h**ou**se	/aʊ/	vowel
p	**p**urr	/p/	plosive
pau	(a short pause)	—	silence
r	**r**oad	/ɹ/	approximant
@@r	n**ur**se	/ɜː/	vowel
s	**s**et	/s/	fricative
sh	**sh**oe	/ʃ/	fricative

Table A.3 – Phonemes in British English and their categories (continued)

Unilex Symbol	Word Example	IPA Symbol	Category
t	**t**ooth	/t/	plosive
th	**th**ank	/θ/	fricative
u	p**u**t	/ʊ/	vowel
uh	l**o**ve	/ʌ/	vowel
ur	j**u**ry	/ʊə/	vowel
uu	f**oo**d	/uː/	vowel
uw	act**u**al	/u/	vowel
v	**v**ote	/v/	fricative
w	**w**et	/w/	approximant
y	**y**es	/j/	approximant
z	**z**oo	/z/	fricative
zh	u**s**ually	/ʒ/	fricative

A.4 German

58 phonemes were employed for German and they cover six categories (except “affricate”). This German phoneme set [Pucher et al., 2010], in which an affricate is split into a plosive and a fricative, was kindly provided by the Telecommunications Research Center Vienna (Forschungszentrum Telekommunikation Wien, FTW), Austria.

Table A.4 – Phonemes in German and their categories

Symbol	Word Example	IPA Symbol	Category
a	N**a**cht	/a/	vowel
a6	schw**ar**z	/a$\underset{\frown}{ɐ}$/	vowel
ah	Flugh**a**fen	/aː/	vowel
ah6	J**ahr**	/aː$\underset{\frown}{ɐ}$/	vowel
Ahn	Appartem**en**t	/ɑ̃ː/	vowel
aI	obgl**ei**ch	/aɪ/	vowel
aU	St**au**	/aʊ/	vowel
b	**B**ein	/b/	plosive
C	natürli**ch**	/ç/	fricative
ch	Na**ch**t	/x/	fricative
d	**D**eich	/d/	plosive
E	st**e**chen	/ɛ/	vowel
E6	St**er**n	/ɛ$\underset{\frown}{ɐ}$/	vowel
eh	Ch**e**mie	/eː/	vowel
Eh6	Entw**er**tung	/eː$\underset{\frown}{ɐ}$/	vowel

Table A.4 – Phonemes in German and their categories (continued)

Symbol	Word Example	IPA Symbol	Category
f	**f**ast	/f/	fricative
g	**G**unst	/ɡ/	plosive
GS	(a glottal stop) [4]	—	silence
h	**H**and	/h/	fricative
I	ep**i**sch	/ɪ/	vowel
I6	Geh**ir**n	/ɪɐ̯/	vowel
ih	gel**ie**bt	/iː/	vowel
ih6	N**ie**re	/iːɐ̯/	vowel
j	**J**ahr	/j/	approximant
k	**K**unst	/k/	plosive
l	**L**öwe	/l/	approximant
m	**m**ein	/m/	nasal
N	Di**ng**	/ŋ/	nasal
n	**n**ein	/n/	nasal
O	n**o**ch	/ɔ/	vowel
O6	N**or**d	/ɔɐ̯/	vowel
oh	Angeb**o**t	/oː/	vowel
oh6	temp**or**är	/oːɐ̯/	vowel
OY	d**eu**tsch	/ɔʏ/	vowel
p	**P**ein	/p/	plosive
P2h	D**ö**schen	/øː/	vowel
P2h6	St**ör**ung	/øːɐ̯/	vowel
P6	Säng**er**	/ɐ/	vowel
P9	T**ö**chter	/œ/	vowel
P96	B**ör**se	/œɐ̯/	vowel
pau	(a short pause)	—	silence
r	Demok**r**at	/ʁ/ [5]	approximant
s	Kun**s**t	/s/	fricative
S	wa**sch**en	/ʃ/	fricative
schwa	wasch**e**n	/ə/	vowel
sil	(a period of silence)	—	silence
t	Kuns**t**	/t/	plosive
U	K**u**nst	/ʊ/	vowel
U6	K**ur**den	/ʊɐ̯/	vowel
uh	B**u**ch	/uː/	vowel

4. It is the glottal stop before a vowel with which a word begins (e.g., “Ost”), or the glottal stop before such a word when it comprises a part of a compound word (e.g., “Nordost”).

5. This IPA symbol itself represents a fricative. This phoneme was considered an approximant because of the question set employed during the German average voice training.

Table A.4 – Phonemes in German and their categories (continued)

Symbol	Word Example	IPA Symbol	Category
uh6	Geb**ur**t	/uːɐ̯/	vowel
v	**w**as	/v/	fricative
Y	Geb**ü**sch	/ʏ/	vowel
Y6	Gew**ür**z	/ʏɐ̯/	vowel
yh	gr**ü**n	/yː/	vowel
yh6	verschn**ü**ren	/yːɐ̯/	vowel
z	Ha**s**e	/z/	fricative
Z	**G**enie	/ʒ/	fricative

B Appendix – Vowel Quadrilateral

The vowel quadrilateral is a part of the IPA chart [1] revised to 2005.

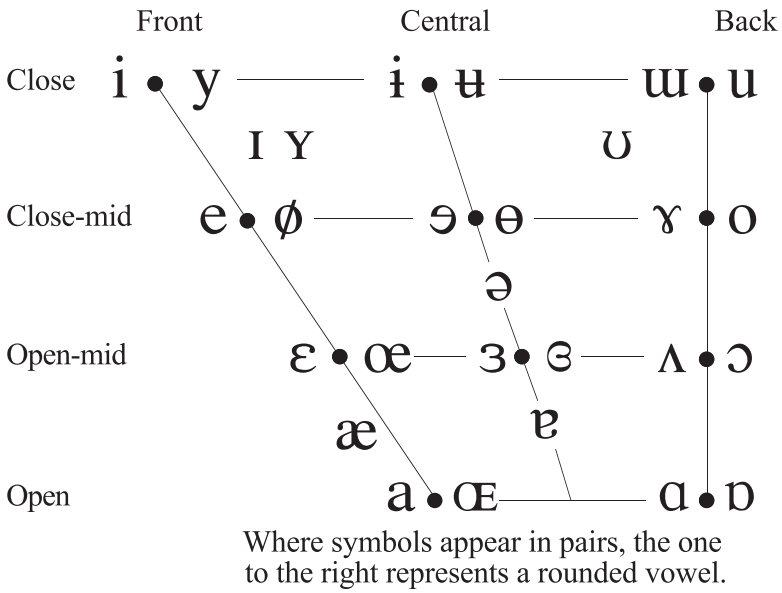


Figure B.1 – Vowel quadrilateral

1. http://www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf

Bibliography

Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *Proc. of ICSLP*, pages 1137–1140, October 1996.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum Likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, March 1983.

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. Real-time incremental speech-to-speech translation of dialogs. In *Proc. of NAACL-HLT*, pages 437–445, June 2012.

Thomas P. Barnwell III. Correlation analysis of subjective and objective measures for speech quality. In *Proc. of ICASSP*, volume 5, pages 706–709, April 1980.

Leonard E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *the Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *the Annals of Mathematical Statistics*, 41(1):164–171, February 1970.

Jeff A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report TR-97-021, International Coputer Science Institute, April 1998.

Alan W. Black and Keiichi Tokuda. The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. of Interspeech*, pages 77–80, Lisbon, Portugal, September 2005.

Michael Byram, editor. *Routledge Encyclopedia of Language Teaching and Learning*. Routledge, 2004. ISBN 0-415-33286-9.

William Byrne, Peter Beyerlein, Juan M. Huerta, Sanjeev Khudanpur, Bhaskara Marthi, John Morgan, Nino Peterek, Joseph Picone, Dimitra Vergyri, and Wei Wang. Towards language independent acoustic modeling. In *Proc. of ICASSP*, volume 2, pages 1029–1032, June 2000.

Yi-Ning Chen, Yang Jiao, Yao Qian, and Frank K. Soong. State mapping for cross-language speaker adaptation in TTS. In *Proc. of ICASSP*, pages 4273–4276, April 2009.

Arthur Pentland Dempster, Nan McKenzie Laird, and Donald Bruce Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistics Society*, 39(1): 1–38, 1977.

John Dines, Lakshmi Saheer, and Hui Liang. Speech recognition with speech synthesis models by marginalising over decision tree leaves. In *Proc. of Interspeech*, pages 1395–1398, September 2009.

Matthias Eichner, Matthias Wolff, and Rüdiger Hoffmann. A unified approach for speech synthesis and speech recognition using stochastic Markov graphs. In *Proc. of ICSLP*, pages 701–704, October 2000.

Matthias Eichner, Matthias Wolff, Sebastian Ohnewald, and Rüdiger Hoffmann. Speech synthesis using stochastic Markov graphs. In *Proc. of ICASSP*, volume 2, pages 829–832, May 2001.

Susan Fitt. Documentation and user guide to Unisyn Lexicon and post-lexical rules. Technical report, Centre for Speech Technology Research, University of Edinburgh, 2000.

Mark Fraser and Simon King. The Blizzard Challenge 2007. In *Proc. of the Blizzard Challenge*, 25 August 2007.

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

Mark J. F. Gales. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417–428, July 2000.

Arnab Ghoshal, Daniel Povey, Mohit Agarwal, Pinar Akyazi, Lukáš Burget, Kai Feng, Ondřej Glembek, Nagendra Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. A novel estimation of feature-space MLLR for full-covariance models. In *Proc. of ICASSP*, pages 4310–4313, March 2010.

Matthew Gibson, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, and William Byrne. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. In *Proc. of ICASSP*, pages 4642–4645, March 2010.

Augustine H. Gray Jr. and John D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5):380–391, October 1976.

Kei Hashimoto, Shinji Takaki, Keiichiro Oura, and Keiichi Tokuda. Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2011. In *Proc. of the Blizzard Challenge*, Turin, Italy, September 2011.

Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP*, pages 373–376, May 1996.

David Imseng, Hervé Bourlard, and Philip N. Garner. Boosting under-resourced speech recognizers by exploiting out of language data - Case study on Afrikaans. In *Proc. of the 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape Town, South Africa, May 2012.

International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, July 1999.

Dorota Iskra, Beate Grosskopf, Krzysztof Marasek, Henk van den Heuvel, Frank Diehl, and Andreas Kiessling. SPEECON - Speech databases for consumer devices: Database specification and validation. In *Proc. of LREC*, pages 329–333, May 2002.

Alexander Kain and Michael W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Proc. of ICASSP*, pages 285–288, May 1998.

Reima Karhila, Doddipatla Rama Sanand, Mikko Kurimo, and Peter Smit. Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN. In *Proc. of ICASSP*, pages 4501–4504, March 2012.

Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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*, 27(3-4):187–207, April 1999.

Hideki Kawahara, Jo Estill, and Osamu Fujimura. 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. of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, September 2001.

Simon King, Keiichi Tokuda, Heiga Zen, and Junichi Yamagishi. Unsupervised adaptation for HMM-based speech synthesis. In *Proc. of Interspeech*, pages 1869–1872, September 2008.

Dennis H. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793, September 1987.

Joachim Köhler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proc. of ICSLP*, pages 2195–2198, October 1996.

Joachim Köhler. Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication*, 35:21–30, 2001.

John Kominek, Tanja Schultz, and Alan W. Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proc. of the Workshop on Spoken Language Technologies for Under-Resourced Languages*, May 2008.

Robert F. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 125–128, May 1993.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Mikko Kurimo, William Byrne, John Dines, Philip Neil Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichiro Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, and Junichi Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *Proc. of the ACL 2010 System Demonstrations*, pages 48–53, July 2010.

Javier Latorre. *A study on speaker-adaptable multilingual synthesis*. PhD thesis, Tokyo Institute of Technology, July 2006.

Javier Latorre, Koji Iwano, and Sadaoki Furui. Polyglot synthesis using a mixture of monolingual corpora. In *Proc. of ICASSP*, pages 1–4, March 2005.

Javier Latorre, Koji Iwano, and Sadaoki Furui. New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. *Speech Communication*, 48(10): 1227–1242, October 2006.

Chin-Hui Lee, Frank K. Soong, and Biing-Hwang Juang. A segment model based approach to speech recognition. In *Proc. of ICASSP*, volume 1, pages 501–504, April 1988.

Kai-Fu Lee. *Automatic Speech Recognition: the Development of the SPHINX System*. Kluwer Academic Publishers, 1989. ISBN 0-89838-296-3.

Lori Levin, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Gavalda, Detlef Koll, and Alex Waibel. the Janus-III Translation System: Speech-to-speech translation in multiple domains. *Machine Translation*, 15:3–25, 2000.

Hui Lin, Li Deng, Dong Yu, Yi-Fan Gong, Alex Acero, and Chin-Hui Lee. A study on multilingual acoustic modeling for large vocabulary ASR. In *Proc. of ICASSP*, pages 4333–4336, April 2009.

Richard P. Lippmann, Edward A. Martin, and Douglas B. Paul. Multi-style training for robust isolated-word speech recognition. In *Proc. of ICASSP*, pages 705–708, April 1987.

Mikiko Mashimo, Tomoki Toda, Kiyohiro Shikano, and Nick Campbell. Evaluation of cross-language voice conversion based on GMM and STRAIGHT. In *Proc. of Eurospeech*, pages 361–364, September 2001.

Marko Moberg, Kimmo Pärssinen, and Juha Iso-Sipilä. Cross-lingual phoneme mapping for multilingual synthesis systems. In *Proc. of Interspeech*, pages 1029–1032, October 2004.

Tor André Myrvoll and Frank K. Soong. Optimal clustering of multivariate normal distributions using divergence and its application to HMM adaptation. In *Proc. of ICASSP*, volume 1, pages 552–555, April 2003.

Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, and Takao Kobayashi. Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis. In *Proc. of Interspeech*, pages 2286–2289, September 2006.

Keiichiro Oura, Yi-Jian Wu, and Keiichi Tokuda. Overview of NIT HMM-based speech synthesis system for Blizzard Challenge 2009. In *Proc. of the Blizzard Challenge*, 4 September 2009.

Keiichiro Oura, Keiichi Tokuda, Junichi Yamagishi, Simon King, and Mirjam Wester. Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 4594–4597, March 2010.

Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. of the Workshop on Speech and Natural Language*, pages 357–362, 1992.

Xianglin Peng, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda. Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices. In *Proc. of ICSP*, pages 605–608, October 2010.

Alan B. Poritz and Alan G. Richter. On hidden Markov models in isolated word recognition. In *Proc. of ICASSP*, pages 705–708, April 1986.

Michael Pucher, Friedrich Neubarth, Volker Strom, Sylvia Moosmüller, Gregor Hofer, Christian Kranzler, Gudrun Schuchmann, and Dietmar Schabus. Resources for speech synthesis of Viennese varieties. In *Proc. of the 7th International Conference on Language Resources and Evaluation*, May 2010.

Yao Qian, Hui Liang, and Frank K. Soong. A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1231–1239, August 2009.

Lawrence R. Rabiner, Jay G. Wilpon, and Frank K. Soong. High performance connected digit recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(8):1214–1225, August 1989.

Tony Robinson, Jeroen Fransen, David Pye, Jonathan Foote, and Steve Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. of ICASSP*, pages 81–84, May 1995.

Manfred R. Schroeder. A brief history of synthetic speech. *Speech Communication*, 13(1-2): 231–237, October 1993.

Tanja Schultz. GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Proc. of ICSLP*, pages 345–348, September 2002.

Tanja Schultz and Katrin Kirchhoff. *Multilingual Speech Processing*. Academic Press, 2006. ISBN 0-12-088501-8.

Tanja Schultz and Alex Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51, 2001.

Michael L. Seltzer and Alex Acero. Separating speaker and environmental variability using factored transforms. In *Proc. of Interspeech*, pages 1097–1100, August 2011.

Koichi Shinoda and Chin-Hui Lee. A structural Bayes approach to speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, 9(3):276–287, March 2001.

Koichi Shinoda and Takao Watanabe. MDL-based context-dependent subword modeling for speech recognition. *Journal of Acoustical Society of Japan (E)*, 21(2):79–86, 2000.

Peter Smit and Mikko Kurimo. Using stacked transformations for recognizing foreign accented speech. In *Proc. of ICASSP*, pages 5008–5011, May 2011.

Frank K. Soong and Biing-Hwang Juang. Line spectrum pair (LSP) and speech data compression. In *Proc. of ICASSP*, pages 37–40, March 1984.

David Sündermann, Harald Höge, Antonio Bonafonte, Hermann Ney, Alan Black, and Shri Narayanan. Text-independent voice conversion based on unit selection. In *Proc. of ICASSP*, pages 81–84, May 2006.

Tomoki Toda and Keiichi Tokuda. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Proc. of Interspeech*, pages 2801–2804, September 2005.

Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. In *Proc. the 5th ISCA Speech Synthesis Workshop*, pages 31–36, June 2004.

Keiichi Tokuda, Takao Kobayashi, Takashi Masuko, and Satoshi Imai. Mel-generalized cepstral analysis – A unified approach to speech spectral estimation. In *Proc. of ICSLP*, pages 1043–1046, September 1994.

Keiichi Tokuda, Takao Kobayashi, and Imai Satoshi. Speech parameter generation from HMM using dynamic features. In *Proc. of ICASSP*, volume 1, pages 660–663, May 1995a.

Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, and Satoshi Imai. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Proc. of Eurospeech*, pages 757–760, September 1995b.

Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 1315–1318, June 2000.

Keiichi Tokuda, Takashi Masuko, Noboru Miyazaki, and Takao Kobayashi. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems*, E85-D(3):455–464, March 2002a.

Keiichi Tokuda, Heiga Zen, and Alan W. Black. An HMM-based speech synthesis system applied to English. In *Proc. of IEEE Workshop on Speech Synthesis*, pages 227–230, September 2002b.

Oytun Türk and Levent M. Arslan. Subjective evaluations for perception of speaker identity through acoustic feature transplantations. In *Proc. of Eurospeech*, pages 2093–2096, September 2003.

Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz. Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training. In *Proc. of Interspeech*, pages 3145–3148, August 2011.

Oliver Watts, Junichi Yamagishi, Simon King, and Kay Berkling. HMM adaptation and voice conversion for the synthesis of child speech: A comparison. In *Proc. of Interspeech*, pages 2627–2630, September 2009.

Mirjam Wester. Cross-lingual talker discrimination. In *Proc. of Interspeech*, pages 1253–1256, September 2010a.

Mirjam Wester. The EMIME bilingual database. Technical Report EDI-INF-RR-1388, University of Edinburgh, 2010b.

Mirjam Wester and Reima Karhila. Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In *Proc. of ICASSP*, pages 5372–5375, May 2011.

Mirjam Wester and Hui Liang. The EMIME Mandarin bilingual database. Technical Report EDI-INF-RR-1396, University of Edinburgh, February 2011.

Stephen J. Winters, Susannah V. Levi, and David B. Pisoni. Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America*, 123(6): 4524–4538, June 2008.

Yi-Jian Wu and Ren-Hua Wang. Minimum generation error training for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 89–92, May 2006.

Yi-Jian Wu, Wu Guo, and Ren-Hua Wang. Minimum generation error criterion for tree-based clustering of context-dependent HMMs. In *Proc. of Interspeech*, pages 2046–2049, September 2006.

Yi-Jian Wu, Simon King, and Keiichi Tokuda. Cross-lingual speaker adaptation for HMM-based speech synthesis. In *Proc. of ISCSLP*, pages 1–4, December 2008.

Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proc. of Interspeech*, pages 528–531, September 2009.

Junichi Yamagishi. *Average-voice-based speech synthesis*. PhD thesis, Tokyo Institute of Technology, March 2006.

Junichi Yamagishi and Takao Kobayashi. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Transactions on Information and Systems*, E90-D(2):533–543, February 2007.

Junichi Yamagishi and Oliver Watts. The CSTR/EMIME HTS system for Blizzard Challenge 2010. In *Proc. of the Blizzard Challenge*, September 2010.

Junichi Yamagishi, Makoto Tachibana, Takashi Masuko, and Takao Kobayashi. Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 5–8, May 2004.

Junichi Yamagishi, Takao Kobayashi, Yuji Nakano, Katsumi Ogata, and Juri 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*, 17(1): 66–83, January 2009a.

Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhen-Hua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, and Steve Renals. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230, August 2009b.

Junichi Yamagishi, Bela Usabaev, Simon King, Oliver Watts, John Dines, Jilei Tian, Yong Guan, Rile Hu, Keiichiro Oura, Yi-Jian Wu, Keiichi Tokuda, Reima Karhila, and Mikko Kurimo. Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):984–1004, July 2010a.

Junichi Yamagishi, Oliver Watts, Simon King, and Bela Usabaev. Roles of the average voice in speaker-adaptive HMM-based speech synthesis. In *Proc. of Interspeech*, pages 418–421, September 2010b.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Duration modeling for HMM-based speech synthesis. In *Proc. of ICSLP*, pages 29–32, November 1998.

Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of Eurospeech*, pages 2347–2350, September 1999.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *the HTK Book*. March 2009.

Steve J. Young, Julian James Odell, and Philip C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. of the Workshop on Human Language Technology*, pages 307–312, 1994.

Heiga Zen and Norbert Braunschweiler. Context-dependent additive log F_0 model for HMM-based speech synthesis. In *Proc. of Interspeech*, pages 2091–2094, September 2009.

Heiga Zen and Tomoki Toda. An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proc. of Interspeech*, pages 93–96, September 2005.

Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Hidden semi-Markov model based speech synthesis. In *Proc. of Interspeech*, pages 1393–1396, October 2004.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, November 2009.

Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark J. F. Gales, Kate Knill, Sacha Krstulović, and Javier Latorre. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE Transactions on Audio, Speech and Language Processing*, 20(6): 1713–1724, August 2012.

Bowen Zhou, Yuqing Gao, Jeffrey Sorensen, Daniel Déchelotte, and Michael Picheny. A hand-held speech-to-speech translation system. In *Proc. of ASRU*, pages 664–669, November 2003.

Hui LIANG (梁晖)

CURRICULUM VITAE

PERSONAL INFORMATION

Nationality: Chinese, the People's Republic of

Languages: Mandarin native
English fluent, no Chinese accent
French basic
Japanese basic

E-mail: hui.ts.liang (AT) gmail.com

Address: Idiap Research Institute, Centre du Parc, Rue Marconi 19,
Case Postale 592, CH-1920 Martigny, Switzerland

RESEARCH INTERESTS

- ◇ text-to-speech synthesis, especially in the multilingual and cross-lingual settings
- ◇ speaker adaptation
- ◇ speech signal processing
- ◇ phonetics and linguistics

EDUCATION

École Polytechnique Fédérale de Lausanne (EPFL) Ph. D. Sep. 2008 ~ Dec. 2012

Location: Lausanne, Switzerland
Major: Electrical Engineering
GPA: 5.0 / 6.0
Supervisors: Prof. Hervé BOURLARD and Dr. John DINES
Thesis: Data-Driven Enhancement of State Mapping-Based Cross-Lingual Speaker Adaptation

Shanghai Jiao Tong University (SJTU) M. Eng. Sep. 2005 ~ Mar. 2008

Location: Shanghai, P. R. China
Major: Communication and Information Systems
GPA: 2.33 / 3.3
Supervisors: Dr. Yao QIAN, Dr. Frank K. SOONG, and Dr. Gongshen LIU
Thesis: A Bilingual (Mandarin-English) TTS System Based on Statistical Models

Huazhong University of Science & Technology (HUST) B. Eng. Sep. 2001 ~ Jun. 2005

Location: Wuhan, Hubei, P. R. China
Major: Communication Engineering
GPA: 86.33 / 100
Supervisors: Mr. Pingguo WAN and Prof. Yunyu ZHANG
Thesis: Topic Classification of BBS Posts Based on the Naive Bayes Algorithm

Hui LIANG (梁晖) CURRICULUM VITAE

WORK EXPERIENCES

Jan. 2013 ~ present	Postdoc Researcher	@ Computer Engineering & Networks Lab, ETHZ
Aug. 2008 ~ Dec. 2012	Research Assistant	@ Idiap Research Institute, Switzerland
May 2008 ~ Jul. 2008	Visiting Student	@ Speech Group of Microsoft Research Asia, China
Nov. 2006 ~ Nov. 2007	Visiting Student	@ Speech Group of Microsoft Research Asia, China
Mar. 2006 ~ Aug. 2006	Research Assistant	@ Content Security Laboratory of SJTU, China
Jan. 2005 ~ Jun. 2005	Intern	@ NetChina Co. Ltd., Beijing, China

RESEARCH/PROJECT EXPERIENCES

Sep. 2009 & Nov. 2010 **High-Quality Bilingual Database Recording**

- The research in EMIME required *bilingual* speech from multiple speakers. *Bilingual* means prompts in English and prompts in another language were read by the same speaker.
- I prepared prompts in Mandarin and English. I, with a colleague of mine, judged the accentedness of candidate speakers, recorded a database containing 18 speakers in an anechoic studio and segmented the recordings.
- I was one of the 18 voice talents reading the bilingual prompts due to my natural English accent.

Aug. 2008 ~ Feb. 2011 **Efficient Multilingual Interaction in Mobile Environments (EMIME)**

- This project was aimed at facilitating cross-lingual spoken communication by developing a mobile device that performed *personalized* speech-to-speech translation, i.e., the voice in synthesized speech of textual translation would sound like that of its user speaking his mother tongue.
- I participated in the research of speech recognition using HMMs trained for speech synthesis, which was proven to be feasible and helpful to bridge the gap between recognition and synthesis.
- I worked on HMM state mapping for cross-lingual speaker adaptation on the speech synthesis side. This is my PhD topic.
- I participated in the research of VTLN-based rapid cross-lingual speaker adaptation.

May 2008 ~ Jul. 2008 **Hybrid System Combining HMM-Based TTS and Unit Selection**

- The research was aimed at building a more natural-sounding, hybrid TTS system over the standard HMM-based framework, which suffered from the over-smoothing problem.
- Instead of generating speech parameters with clustered HMM states, it was investigated how to concatenate HMM states obtained before decision tree-based clustering for speech parameter generation. I designed a search algorithm following context-based pre-selection to determine an optimal unclustered HMM state sequence that could generate natural and crisp speech.

Aug. 2007 ~ Sep. 2007 **Improving F_0 Contour Prediction by Additive Tree Modeling**

- The research was aimed at improving overly smoothed pitch prediction and generating more natural prosody/intonation of spoken English, especially at the sentence level.
- I employed multiple additive regression trees (a gradient-based tree-boosting algorithm) to produce more natural F_0 trajectories. The trees were trained in successive stages to minimize the squared

error between natural and predicted F_0 trajectories. Experimental results of both Mandarin and English TTS trials showed that the proposed approach could increase not only the dynamic range of generated F_0 trajectories, but also improve the common objective and subjective measures.

Jun. 2007 ~ Jul. 2007 **Speech Synthesis in a Language without Prerecorded Data**

- § The research was aimed at building a synthesizer in language *A* on the basis of speech data in language *B*, i.e., to convey a voice identity across languages.
- The goal was achieved by cross-lingual state mapping and model parameter substitution. Comparing the original voice in language *A* and the synthetic voice in language *B*, we found close similarity. Perceptual tests confirmed high intelligibility.

Nov. 2006 ~ May 2007 **Bilingual (Mandarin-English) TTS System**

- § The research was aimed at building an HMM-based bilingual synthesizer which could generate Mandarin, English and code-switched utterances of high quality.
- I took charge of analyzing the phonologies of Mandarin and American English, designing items of bilingual context-dependent labels and a bilingual question set for tree-based clustering, etc.
- The goal was achieved by designing language specific and independent questions to facilitate HMM state sharing across the two languages. Due to shared states, the new system had a smaller footprint than the baseline. The synthesis quality was the same for mono-lingual (either Mandarin or English) utterances as that of the baseline, and much better for code-switched utterances.

Mar. 2006 ~ Aug. 2006 **Online System of Acquiring Public Sentiment**

- The project was aimed at analyzing information collected from the Internet (mainly BBS in Chinese) in order to help the government of Shanghai make policies and decisions. I took charge of designing a module of text clustering based on the K-Means algorithm.

Jan. 2005 ~ Jun. 2005 **Topic Classification of BBS Posts**

- This was the bachelor thesis project, which was aimed at applying topic classification to BBS posts. I took charge of designing a module for topic classification based on the naive Bayes algorithm.

PRACTICAL SKILLS

- ◇ *Programming*: C, C++ & STL, Perl, MatLab, Unix/Linux shell, Visual Basic
- ◇ *Speech processing toolkits*: HTK, HTS, SPTK, WaveSurfer
- ◇ *Operating systems*: Windows, Unix/Linux

AWARDS AND HONORS

June 2005

- Outstanding Graduate
- Hubei Province-Level Second Prize for Distinguished Bachelor's Thesis

The academic year of 2002 ~ 2003

- Excellent Student Leader
- Excellent Student in Study

The academic year of 2001 ~ 2002

- All-Round Good Student

HOBBIES

badminton, swimming, skating, travelling, reading, playing the harmonica

REFERENCES

available upon request

Hui LIANG (梁晖) PUBLICATIONS

JOURNAL PAPERS

1. **Hui LIANG** and John DINES, "Jointly Data-Driven and Phonological Knowledge-Guided Enhancement of State Mapping Based Cross-Lingual Speaker Adaptation", submitted to *IEEE Transactions on Audio, Speech and Language Processing*.

2. John DINES, **Hui LIANG**, Matthew GIBSON, Keiichiro OURA, Junichi YAMAGISHI, Simon KING, Mirjam WESTER, Lakshmi SAHEER, Teemu HIRSIMÄKI, Reima KARHILA, Keiichi TOKUDA, William BYRNE, Mikko KURIMO, "Personalising Speech-to-Speech Translation: Unsupervised Cross-Lingual Speaker Adaptation for HMM-Based Speech Synthesis", accepted by *Computer Speech & Language*.

3. Hervé BOURLARD, John DINES, Mathew MAGIMAI-DOSS, Philip N. GARNER, David IMSENG, Petr MOTLICEK, **Hui LIANG**, Lakshmi SAHEER, Fabio VALENTE, "Current Trends in Multilingual Speech Processing", *Sādhanā*, Volume 36, Part 5, pp. 885-915, October 2011. (an invited paper for the special issue on the topic of Speech Communication and Signal Processing)

4. Yao QIAN, **Hui LIANG** and Frank K. SOONG, "A Cross-Language State Sharing and Mapping Approach to Bilingual (Mandarin-English) TTS", *IEEE Transactions on Audio, Speech and Language Processing*, Volume 17, Issue 6, pp. 1231-1239, August 2009.

CONFERENCE PAPERS

1. **Hui LIANG** and John DINES, "Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation", *Proc. of Interspeech*, pp. 1825-1828, August 2011.

2. Mirjam WESTER and **Hui LIANG**, "Cross-Lingual Speaker Discrimination Using Natural and Synthetic Speech", *Proc. of Interspeech*, pp. 2481-2484, August 2011.

3. Mikko KURIMO, William BYRNE, John DINES, Philip N. GARNER, Matthew GIBSON, Yong GUAN, Teemu HIRSIMÄKI, Reima KARHILA, Simon KING, **Hui LIANG**, Keiichiro OURA, Lakshmi SAHEER, Matt SHANNON, Sayaka SHIOTA, Jilei TIAN, Keiichi TOKUDA, Mirjam WESTER, Yi-Jian WU and Junichi YAMAGISHI, "Personalising speech-to-speech translation in the EMIME project", *Proc. of the ACL 2010 System Demonstrations*, pages 48-53, July 2010.

4. Lakshmi SAHEER, John DINES, Philip N. GARNER and **Hui LIANG**, "Implementation of VTLN for Statistical Speech Synthesis", *Proc. of the 7th ISCA Speech Synthesis Workshop*, pp. 224-229, September 2010.

5. Mirjam WESTER, John DINES, Matthew GIBSON, **Hui LIANG**, Yi-Jian WU, Lakshmi SAHEER, Simon KING, Keiichiro OURA, Philip N. GARNER, William BYRNE, Yong GUAN, Teemu HIRSIMÄKI, Reima KARHILA,

Mikko KURIMO, Matt SHANNON, Sayaka SHIOTA, Jilei TIAN, Keiichi TOKUDA and Junichi YAMAGISHI, "Speaker Adaptation and the Evaluation of Speaker Similarity in the EMIME Speech-to-Speech Translation Project", *Proc. of the 7th ISCA Speech Synthesis Workshop*, pp. 192-197, September 2010.

6. **Hui LIANG** and John DINES, "An Analysis of Language Mismatch in HMM State Mapping-Based Cross-Lingual Speaker Adaptation", *Proc. of Interspeech*, pp. 622-625, September 2010.

7. **Hui LIANG**, John DINES and Lakshmi SAHEER, "A Comparison of Supervised and Unsupervised Cross-Lingual Speaker Adaptation Approaches for HMM-Based Speech Synthesis", *Proc. of ICASSP*, pp. 4598-4601, March 2010.

8. Lakshmi SAHEER, Philip N. GARNER, John DINES and **Hui LIANG**, "VTLN Adaptation for Statistical Speech Synthesis", *Proc. of ICASSP*, pp. 4838-4841, March 2010.

9. John DINES, Lakshmi SAHEER and **Hui LIANG**, "Speech Recognition with Speech Synthesis Models by Marginalising over Decision Tree Leaves", *Proc. of Interspeech*, pp. 1395-1398, September 2009.

10. Yao QIAN, **Hui LIANG** and Frank K. SOONG, "Generating Natural F0 Trajectory with Additive Trees", *Proc. of Interspeech*, pp. 2126-2129, September 2008.

11. **Hui LIANG**, Yao QIAN, Frank K. SOONG and Gongshen LIU, "A Cross-Language State Mapping Approach to Bilingual (Mandarin-English) TTS", *Proc. of ICASSP*, pp. 4641-4644, March 2008.

12. **Hui LIANG**, Yao QIAN and Frank K. SOONG, "An HMM-Based Bilingual (Mandarin-English) TTS", *Proc. of the 6th ISCA Workshop on Speech Synthesis*, pp. 137-142, August 2007.

TECHNICAL REPORTS

1. Lakshmi SAHEER, **Hui LIANG**, John DINES, Philip N. GARNER, "VTLN-Based Rapid Cross-Lingual Adaptation for Statistical Parametric Speech Synthesis", Idiap Research Institute, Switzerland, Tech. Rep. Idiap-RR-12-2012, April 2012.

2. Mirjam WESTER and **Hui LIANG**, "The EMIME Mandarin Bilingual Database", University of Edinburgh, United Kingdom, Tech. Rep. EDI-INF-RR1396, February 2011.