

基于状态映射的跨语种自适应研究

夏菽兰¹, 赵 力²

(1 盐城工学院 电气工程学院, 江苏 盐城 224051; 2 东南大学 信息科学与工程学院, 江苏 南京 210096)

摘 要: 以基于状态映射的中英跨语种自适应方法为研究对象, 将状态映射应用于跨语种自适应中, 对自适应中两种使用状态映射信息的方式(数据映射和变换映射)进行了比较分析. 通过实验发现, 基于状态映射的自适应方法具有一定的效果, CSMAPLR 算法优于 CMLLR 算法, 而且两种方式各有优劣.

关键词: 状态映射; 跨语种自适应; 数据映射; 变换映射

中图分类号: TN912.33

文献标识码: A

文章编号: 1000-7180(2015)04-0134-04

Research on Cross-lingual Adaptation Based on State Mapping

XIA Shu-lan¹, ZHAO Li²

(1 College of Electrical Engineering, Yancheng Institute of Technology, Yancheng 224051, China;

2 College of Information Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: The Chinese and English cross-lingual adaptation algorithms are studied, which are based on state mapping method. Two modes of state mapping method applied in cross-lingual adaptation are compared and analyzed. Experiments show that the adaptation algorithms can make sense, and that CSMAPLR is better than CMLLR algorithms. Furthermore, voice quality is better in transform mode than in data mode. And it needs high consistency of speaker characteristics of two original language models. On the contrary, speaker similarity in data mode is better.

Key words: state mapping; cross-lingual adaptation; data mapping; transform mapping

1 引言

随着中英文的跨语种语音合成技术的发展, 合成固定说话人的中英文语音已不能满足应用的需求, 人们渐渐趋向于因人而异的语音的合成实现. 跨语种自适应是在同一语种的说话人自适应技术的基础上发展起来, 它在自适应说话人特征的同时, 保留目标说话人的语言特征. 早在 20 世纪末期, 就有研究者将最大后验概率^[1] (MAP) 和最大似然度^[2] (ML) 用于说话人自适应技术的研究中, 提出最大后验概率线性回归^[3] (MAPLR) 和最大似然度线性回归^[4] (MLLR) 的参数调整算法, 后来自适应算法的研究大都是在他们的基础上发展起来的, 如结构化的最大后验概率的参数变换方法^[5] (SMAPLR), 而结构化的最大似然度线性回归^[6-7] (CMLLR) 则

继承了 MLLR 中均值矢量的变换方法, 并且将变换的范围扩展至方差矩阵. 相类似的受限的结构化 MAPLR^[8] (CSMAPLR) 在 MAPLR 的基础上进一步对方差矢量进行了变换. 本文针对中英文跨语种自适应方法展开研究, 在分析比较各种自适应算法性能的基础上, 将状态映射应用于跨语种自适应过程中, 并对自适应的效果进行分析研究.

2 基于状态映射的跨语种自适应方法

基于状态映射的方法, 建立映射是在比音素更小的状态层次上, 该方法虽然发展时间较短, 但是具有极大的潜力.

2.1 基于 KLD 的状态映射

跨语种自适应是根据两种语言模型间的状态映射关系, 将针对输入语言模型估计出的线性变换应

收稿日期: 2014-07-23; 修回日期: 2014-09-10

基金项目: 国家自然科学基金项目 (60472058)

用于输出语言模型. 而状态映射是以最小 KL 散度^[9] (Kullback-Leibler divergence, KLD) 为依据, 为源语言模型和目标语言模型之间的状态建立配对关系. 为了得到更好的映射效果, 两种语言的模型应该用同一说话人的双语语料分别训练得到, 或者采用两种语言各自的平均语音模型^[10] (average voice model), 本文以同一说话人的中英 HMMs (Hidden Markov Models, 隐马尔可夫模型) 为例, 对状态映射的原理进行说明.

图 1 为中文 HMMs 决策树至英文 HMMs 决策树的状态映射示意图, 映射的两个状态之间具有最小的 KLD, 且可能出现多个中文 HMMs 的状态对应于同一个英文 HMMs 状态, 即一对多的情况. 英文 HMMs 至中文 HMMs 的状态映射与此类似, 实际操作时只需要从相反的方向进行.

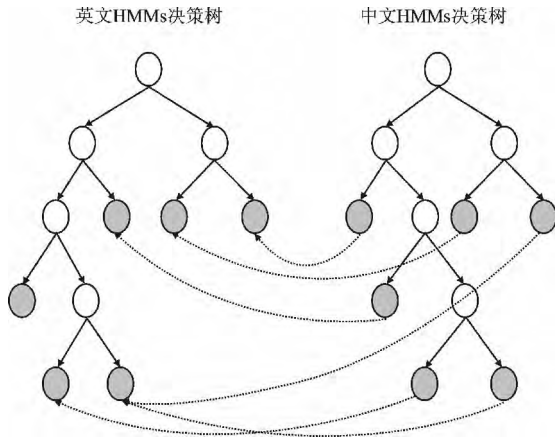


图 1 中文 HMMs 决策树至英文 HMMs 决策树的状态映射

KLD 是衡量两个概率分布之间的相似(异)性的一种信息度量. 对于给定连续随机变量的两个概率分布 P 和 Q , P 和 Q 的 KLD 计算公式为

$$D_{KL}(P, Q) = \int p(x) \log \frac{p(x)}{q(x)} dx + \int q(x) \log \frac{q(x)}{p(x)} dx \quad (1)$$

式(1)是对称的, p 和 q 分别表示 P 和 Q 的密度. 而对于两个多维高斯分布来说, 式(1)变为

$$D_{KL}(P, Q) = \frac{1}{2} \text{tr} \left\{ \left(\sum_p^{-1} + \sum_q^{-1} \right) (\mu_p - \mu_q) \cdot (\mu_p - \mu_q)^T \sum_p \sum_q^{-1} + \sum_q \sum_p^{-1} - 2I \right\} \quad (2)$$

式中, μ 和 $\sum$ 分别为均值矢量和方差矩阵.

在基于 HMM 的语音合成中, 基频、频谱和时长参数是分开建模的, 它们的 HMMs 分别被聚类成相互独立的模型决策树. 频谱和时长用 HMMs

建模, 可用式(1)来衡量 HMMs 的两个状态的相似性. 而基频是用 MSD-HMM 建模, MSD-HMM 包括两个概率空间, 即一个描述清/浊音比例的离散空间和一个描述浊音段基频的连续空间. MSD-HMMs 的两个状态的 KLD 上限为式(3).

$$D_{KL}(P, Q) \leq (\omega_0^p - \omega_0^q) \log \frac{\omega_0^p}{\omega_0^q} + (\omega_1^p - \omega_1^q) \log \frac{\omega_1^p}{\omega_1^q} + \frac{1}{2} \text{tr} \left\{ \left(\omega_1^p \sum_p^{-1} \mathbb{I} + \omega_1^q \sum_q^{-1} \right) (\mu_p - \mu_q) (\mu_p - \mu_q)^T + \left(\omega_1^p \sum_p \sum_q^{-1} - I + \omega_1^q \sum_q \sum_p^{-1} - I \right) \right\} + \frac{1}{2} (\omega_1^p - \omega_1^q) \log \left| \sum_p \sum_q^{-1} \right| \quad (3)$$

式中, ω_0 和 ω_1 分别为浊音和清音子空间的先验概率. 中文 HMMs 与英文 HMMs 的建模参数相同, 都包括基频、频谱和时长, 两个模型的状态映射具体表现在对应决策树中的叶节点的映射.

2.2 映射方式

基于状态映射的跨语种自适应方法的原理如图 2 所示, 其中有两种使用状态映射的方式: 数据映射方式和变换映射方式.

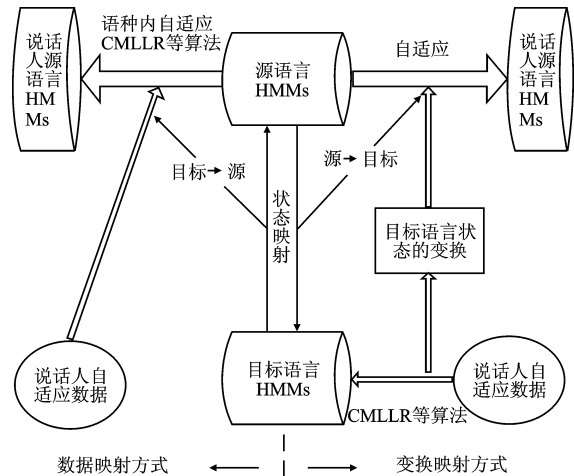


图 2 基于状态映射的跨语种自适应方法

数据映射方式中, 状态映射信息是在目标语言的自适应数据被附着于源语言的语音模型的过程中使用, 操作步骤如下所示.

- (1) 训练源语言和目标语言的语音模型, 两个模型具有同一说话人特征或者均为平均语音模型.
- (2) 在源状态模型和目标状态模型之间建立状态映射, 例如为目标语言的每个状态模型找到一个与它最接近(KLD 最小)的源语言状态模型.
- (3) 根据状态映射信息, 将目标语言的自适应数据附着于源语言的语音模型.
- (4) 把自适应数据当作源语言的数据, 对源语言的语音模型进行语种内说话人自适应.

数据映射的方式是音素映射方法的一种扩展,区别在于它在源语音模型和自适应数据之间建立联系是在状态层次,而不是在音素层次。需要注意的是,数据与语音模型之间的映射并不会对它们的真正关系产生影响,例如,该映射不包括与自适应数据每一帧相关联的状态模型的信息,它只反映状态序列和每一个自适应句子间的映射关系。

变换映射方式中,状态映射信息是在目标语言语音模型的变换被附着于源语言语音模型的过程中使用,操作步骤如下所示。

(1)训练源语言和目标语言的语音模型,两个模型具有同一说话人特征或者均为平均语音模型。

(2)在源状态模型和目标状态模型之间建立状态映射,例如为源语言的每个状态模型找到一个与它最接近(KLD 最小)的目标语言状态模型。

(3)用自适应数据训练目标语言语音模型的变换。

(4)基于状态映射信息,用与源语言状态模型相映射的目标语言状态模型的变换,对每一个源语言状态模型进行自适应。

变换映射的方式没有在源语言语音模型和目标语言自适应数据之间建立直接的联系,它用状态映射信息将针对目标语言的变换应用于源语言,然后用映射得到的变换对源语言状态模型进行自适应。需要注意的是,变换映射的方式是基于源语言模型与目标语言模型相似的假设的。如果两个模型是用同一说话人的双语语料库训练出来的,则满足该假设,但实际上很难获得同一说话人的双语语料库。虽然可以用说话人个数相近、性别比例相近的两个语料库来代替同一说话人的双语语料库,但是训练出来的语音模型仍然存在一定的差异。

2.3 优缺点分析

在跨语种说话人自适应中,自适应数据与源语言模型的说话人特征和语言类型都不相同,研究的关键在于如何在自适应目标说话人特征的同时保持源语言模型的语言类型,即要尽量避免自适应数据的语言类型对说话人自适应的影响。本文从使用状态映射信息的原理出发分析数据映射和变换映射的优缺点。

(1)数据映射方式。优点:直接将自适应数据用于说话人自适应,可实现源语言模型说话人特征至目标说话人特征的高效自适应。缺点:虽然经过映射之后,可以将自适应数据当作源语言的数据使用,但是数据中仍然存在目标语言的信息,以至于源语言模型的语言类型不能被完整地自适应为目标语言。

(2)变换映射方式。优点:因为以同语种自适应的方式来训练目标语言的变换,所以可以避免自适应数据中不同语言类型的影响。缺点:要求源语言模型和目标语言模型具有相同的说话人特征,如果两个模型的说话人特征不匹配,则自适应后合成出的语音与目标说话人的特征不相符,即说话人自适应的效果较差。

由于是在模型的状态层次上建立映射关系,状态映射可以避免音素映射中语言音素差异对自适应效果的影响,对不同语言的适用性较强,有比较好的应用前景。

3 自适应实验

3.1 实验条件

实验以中文、英文两种语言模型为对象,采用不同的自适应算法进行说话人自适应,并对自适应的效果进行分析。实验中,用 UEDIN 的双语语料库中的 3 个男性(MM1-3)、3 个女性(MF1-3)的双语语料(中、英文各 942 句)训练中文和英文模型,每个说话人具有 145 句英文语料和 169 句中文语料;英文说话人自适应数据选自 CMU-ARCTIC 英文库,中文说话人自适应数据选自 EMIME 的自适应库。语音信号采样率为 16 kHz,且采用 5 ms 帧移、25 ms Blackman 窗。建模语音参数包括对数基频 f_0 、25 阶的梅尔倒谱系数以及它们的一阶、二阶动态系数;用 5 状态的与文本相关的 HSMMs 对中文、英文音素进行建模;每个状态存在两个分布,一是状态输出的含对角协方差矩阵的单高斯 pdf,二是状态时长的含标量方差的单高斯 pdf。合成语音用合成引擎 `hts_engine` 实现,合成滤波器为 MLSA(Mel Log Spectrum Approximation)滤波器。

3.2 中英跨语种自适应

根据前人基于音素映射的方法研究的结果,时长模型的自适应对跨语种自适应技术的影响不大,因此,实验中只对频谱和 f_0 的模型进行自适应。由于本文研究以解决中文说话人的英文合成的语料短缺问题为目标,主要研究中文至英文的跨语种说话人自适应,故实验中采用中文说话人的中文语音作为自适应数据,通过跨语种自适应来实现该说话人的英文语音的合成。

从 EMIME 库中挑选 100 句目标说话人的中文句子,分别采用数据映射和变换映射两种方式,进行基于状态映射的中英跨语种自适应,得到说话人的英文模型,然后对 10 句英文句子文本用自适应后的英文模型合成。

由于缺乏目标说话人的英文语料,不能计算目标说话人原始的英文模型与跨语种自适应得到的英文模型之间的欧氏距离,为此本次实验主要从主观上评价跨语种自适应的效果.评价方法与语种内自适应相同,只是缺少目标说话人相似度评价的英文参考语音,实际评测时用目标说话人的中文语音来代替.另外,由于引擎的合成滤波器对音质存在一定的影响,为了排除合成阶段这一影响因素,参考语音通过400句语料训练出的该说话人的中文HMMs合成得到,而不采用语料库中的自然的中文语音.

由于方差对合成语音的韵律有较大的影响,考虑到中文和英文两种语言的韵律差异,实验中只采用变换模型方差的CMLLR和CSMAPLR自适应算法.主观评价结果如表1所示.

表1 跨语种自适应效果的主观MOS评分

映射方式	数据映射		变换映射	
算法	CMLLR	CSMAPLR	CMLLR	CSMAPLR
音质	3.125	3.25	3.25	3.5
说话人相似度	3.25	3.5	2.875	3.25

从MOS评分结果中看出,基于状态映射的跨语种自适应方法取得了一定的效果,而且变换映射方式的语音具有较好的音质,但是与目标说话人的相似度却低于变换映射的方式.也就是说,变换映射的方式能够获得更好的说话人相似度,而数据映射的方式可获得更好的音质.该结果验证了两种方式原理上的优缺点分析结果的正确性,由于用于自适应的原始模型是由四个说话人的双语数据训练出来的,而没有采用同一说话人的双语数据,两个语言模型的说话人特征存在一定的差异,故而降低了变换映射方式的说话人自适应性能,使合成语音的说话人特征与目标说话人之间的相似程度降低.另外,与CMLLR自适应算法相比,CSMAPLR算法能够产生更好的音质与说话人相似度效果.

综上所述,基于状态映射的跨语种自适应方法的中英自适应效果较好,且两种方式性能各有优劣,可根据具体情况具体选择.

4 结束语

以基于状态映射的跨语种自适应方法为研究对象,分析了基于数据映射和基于变换映射的两种方式,从原理上对它们的优缺点进行了分析.实验结果表明,中英跨语种自适应中基于状态映射的自适应方法具有一定的效果,CSMAPLR算法优于CM-

LLR算法,而且两种方式各有优劣.变换映射的方式能实现更好的语音音质,对原始的两个语言模型说话人特征的一致性有较高要求,而数据映射的方式可达到更高的说话人相似度.

参考文献:

- [1] Gauvain J L, Lee C H. Maximum a posteriori estimation for multi-variant Gaussian mixture observations of Markov chains[J]. IEEE Trans on Speech and Audio Process, 1994, 2(2): 291-298.
- [2] Cox S J. A speaker adaptation technique using linear regression[J]. ICASSP, 1995, 1(1): 700-703.
- [3] Ogata K, Tachibana M, Yamagishi J, et al. Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis [C]// Interspeech Pittsburah, PA, USA; IEEE, 2006: 1328-1331.
- [4] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. Computer Speech & Language, 1995, 9(2): 171-185.
- [5] Gales M. Maximum likelihood linear transformations for HMM-based speech recognition [J]. Computer Speech & Language, 1998, 12(2): 75-98.
- [6] Ferras M, Leung C C, Barras C, et al. Constrained MLLR for Speaker Recognition[J]. ICASSP, 2007, (4): 53-56.
- [7] 曹骥, 吴玲达. 一种改进的领域本体语义相似度计算方法[J]. 微电子学与计算机, 2014, 31(8): 109-114.
- [8] Nakano Y, Tachibana M, Yamagishi J, et al. Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis [C] // proceedings of the Ninth International Conference on Spoken Language, Pittsburgh, PA, USA; IEEE, 2006.
- [9] Tamura M, Masuko T, Tokuda K, et al. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR[C]// Proceeding of the International Conference on Acoustics, Speed, and Signal, Salt Lake City, UT; IEEE, 2001: 805-808.
- [10] Yamagishi J, Kobayashi T. Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training[J]. IEICE Trans Inf Syst, 2007, 90(2): 533-543.

作者简介:

夏菽兰 女, (1969-), 硕士, 副教授. 研究方向为电子技术与信息处理. E-mail: xslnj@126.com.