基于自适应训练的疑问句语音合成*

方硕，温正棋 ，王洋，陶建华
中国科学院自动化研究所 模式识别国家重点实验室 北京 100190

文 摘： 针对目前合成语音缺乏表现力的现状，本文提出了一种基于自适应训练的疑问句语音合成方法。采用基于统计参数语音合成技术，用大规模的陈述句语料训练初始声学模型，在此基础上，采用小规模的疑问句语料进行自适应训练，得到疑问句的声学模型，从而合成出具有疑问语气的语音。实验结果表明，在疑问句训练语料较少的情况下，该方法的合成语音在自然度以及客观评价指标上明显优于以相同疑问句语料直接训练的方法，并且能用较少的语料达到直接采用大语料训练的效果。

关键词： 疑问句生成；自适应；语音合成
中图分类号： H116.4；TN912.3

语气合成是富有表现力的语音合成方法研究中的一个重要方面。语气大体可以分为陈述句、疑问句、祈使句和感叹句等。目前的语音合成系统大多针对陈述句设计，在合成其它语气时，语气的表达效果不明显。如果语音合成系统能够在语气的合成上有较大的突破，那么合成语音的表现力将会进一步提高，人机交互将会更加和谐自然。

疑问句是自然口语中常见的语言现象。本文以疑问语气为例，重点研究疑问句的语音合成方法。文献[1]在分析了带有情态标记的疑问句的韵律特点之后，通过构建新的韵律模板库和构建新的目标代价函数，在波形拼接合成系统框架[2]下，实现疑问句的生成。从实现方法上来说，文献[1]中的方法有以下几点不足。首先该方法需要在具有文本的情感标记的基础上进行疑问句的韵律分析；其次，需要把语调的变化归结为在关键音节的前后几个位置的变化，不具有一般性；最后，系统采用波形拼接的方法实现语气的合成，会保留这种方法的不足。文献[3]在统计参数语音合成的框架下[4][5]，用一定的疑问句语料进行训练，实现疑问句的生成。该方法不需要进行疑问句的韵律分析，采用机器学习的方式来学习疑问语气中的韵律，实现语气的合成，方法更为一般化。但是该方法对疑问句的训练语料量要求较大。而大量的疑问句的训练语料是比较难以获取的。因此，如何用少量的疑问句语料，快速地构建一个疑问句合成系统是问题的关键所在。

本文提出了一种基于自适应训练的疑问句生成方法。在统计参数语音合成方法的框架下，用大规模的陈述句语料进行训练，得到陈述句的声学模型，将此模型作为自适应训练的初始模型，采用小规模的疑问句语料对其进行自适应训练，从而得到疑问句的声学模型，最终合成出具有疑问语气的语音。实验结果表明在少量疑问句语料的情况下，本文提出的方法能够合成较高质量的具有疑问语气的语句，在客观指标上明显优于采用相同数量语料的直接训练的合成方法，甚至达到了采用大语料直接训练的合成结果。

1 系统框架

疑问句语音合成系统以统计参数语音合成方法为基础，如图 1 所示，系统主要由以下三部分组成，分别为陈述语气模型训练，疑问语气自适应训练以及语音生成部分。

在第一部分的陈述语气模型训练中，以大规模的陈述语料作为训练语料，提取基频、谱参数进行 MSD-HSMM（Multi-Space Distribution-Hidden Semi Markov Model）训练，训练所得模型作为自适应训练的初始模型。值得注意的是，针对不同自适应语料的实验，这一部分只需进行一次训练，不同实验只需将该部分的模型作为输入的初始模型即可。

在第二部分的疑问语气自适应训练中，以小规模的疑问句语料作为训练语料，采用自适应训练算法，对上一步所得的陈述语气 MSD-HSMM 进行疑问语气的自适应训练，得到疑问语气的 MSD-HSMM。

*基金项目： 青年科学基金项目 (61403386) 个性化语音合成的自适应方法研究
作者简介：方硕（1991），女（汉），浙江，硕士研究生。
通讯联系人：陶建华，研究院，E-mail:jhtao@nlpr.ia.ac

第三部分即为语音生成部分，输入的文本经过文本分析转换为上下文相关的音素标记序列。根据该序列，得到句子级别的 MSD-HSMM 序列，运用基于极大似然准则的参数生成算法[6][7]得到相应各音素的基频、谱和时长参数，将语音参数输入声码器得到疑问语气的合成语音。

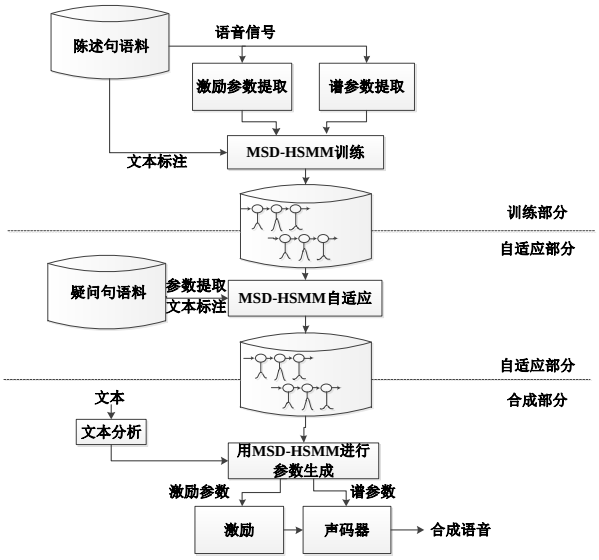

图 1　基于自适应方法的疑问语气合成系统框图

2　声学模型及自适应算法

疑问句与陈述句的区别主要体现在韵律上，在语音建模过程中，韵律的不同主要体现在基频以及音素时长的建模上。在基于隐马尔科夫模型（HMM）的统计参数语音合成系统框架下，可以采用自适应的方法进行谱和基频分布的自适应训练。然而 HMM 没有显示的时长分布函数，较难对时长进行自适应训练。隐半马尔科夫模型（HSMM）是一种具有确定时长分布的隐马尔科夫模型，能同时对基频、谱和时长进行自适应训练。HSMM 训练中的相关数学推导与 HMM 相似，可以参考[8]。

自适应训练中，常用的自适应算法有 MAP（Maximum A Posterior）[9]、 SMAP（Structural MAP）[10]、 MLLR（Maximum Likelihood Linear Regression）[11]、CMLLR（Constrained MLLR）[8]或多种算法相结合的方法[12]。根据疑问句生成的需求，本文采用 CMLLR 以及 SMAP 相结合的自适应算法。HSMM-CMLLR 采用同一个变换矩阵同时对状态输出分布以及时长分布的均值和方差进行变换，如公式(1)和(2)：

$$\begin{aligned} b_i(\boldsymbol{o}) &= N(\boldsymbol{o};\boldsymbol{\zeta}'\boldsymbol{\mu}_i-\boldsymbol{\varepsilon}';\boldsymbol{\zeta}'\boldsymbol{\Sigma}_i\boldsymbol{\zeta}'^{\boldsymbol{T}}) \\ &= |\boldsymbol{\zeta}|\,N(\boldsymbol{\zeta o}+\boldsymbol{\varepsilon};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \\ &= |\boldsymbol{\zeta}|\,N(\boldsymbol{W}\xi;\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i) \end{aligned} \tag{1}$$

$$\begin{aligned} p_i(d) &= N(d;\chi' m_i-\upsilon',\chi'\sigma_i^2\chi') \\ &= |\chi|\,N(\chi d+\upsilon;m_i,\sigma_i^2) \\ &= |\chi|\,N(\boldsymbol{X}\phi;m_i,\sigma_i^2) \end{aligned} \tag{2}$$

其中，$\boldsymbol{o}\in R^L$ 是一个 L 维的观测向量，d 为时长，矩阵 $\boldsymbol{\zeta}'\in R^{L\times L}$ 是对状态 i 的输出高斯分布的均值及方差进行变换的变换矩阵，标量 χ' 用来对时长分布的均值及方差进行变换。$\boldsymbol{\varepsilon}'\in R^L$ 以及 υ' 为偏差。公式（1）、（2）的第二个和第三个等号的推导表示对模型参数的变换，等价于对观测向量 $\boldsymbol{o}$ 和时长 d 作变换。其中 $\boldsymbol{\zeta}=\boldsymbol{\zeta}'^{-1}$，$\boldsymbol{\varepsilon}=\boldsymbol{\zeta}'^{-1}\boldsymbol{\varepsilon}'$，$\boldsymbol{\chi}=\boldsymbol{\chi}'^{-1}$，$\xi=[\boldsymbol{o}^T,1]^T$，$\phi=[d,1]^T$。$\boldsymbol{W}=[\boldsymbol{\zeta},\boldsymbol{\varepsilon}]\in R^{L\times(L+1)}$，$\boldsymbol{X}=[\chi,\upsilon]\in R^{1\times 2}$ 为状态输出分布及时长分布的变换矩阵。

估计变换参数 $\Lambda=(\boldsymbol{W},\boldsymbol{X})$ 即为最大化自适应数据 $\boldsymbol{O}$ 的似然值，如公式(3)所示：

$$\Lambda=(\boldsymbol{W},\boldsymbol{X})=\arg\max_{\Lambda} P(\boldsymbol{O}\,|\,\lambda,\Lambda) \tag{3}$$

λ 为原始模型参数。运用 EM（Expectation Maximization）算法[13]对变换参数 Λ 进行求解。

MAP 算法的基本思想可由公式(4)进行说明，其中 g、f 均为概率密度函数。相较于 ML（Maximum Likelihood）算法，MAP 引入了模型参数的先验分布，在数据量较少的情况下，对模型参数的估计更加可靠。

$$\begin{aligned} \theta_{MAP} &= \arg\max_{\theta} g(\theta\,|\,x) \\ &= \arg\max_{\theta} f(x\,|\,\theta)g(\theta) \end{aligned} \tag{4}$$

但传统的 MAP 只能根据数据对局部模型进行自适应，SMAP 主要针对 MAP 局部估计的这个问题，通过构建模型参数空间的层级结构，用少量的自适应数据对所有的模型参数进行调整。

3　实验

3.1　语料

实验训练所需的疑问句语料 2600 句男声汉语疑问句语料，其中包含数量相当的特指疑问句、是非疑问句、一般疑问句以及选择疑问句。陈述句语料为与疑问句语料同一男声录制的 6700 句。所有的录音都为单通道、16kHz 采样率、16 位的 wav 文件格式。采用 7 状态的上下文相关的一阶 MSD-HMM 作为声学模型，以中文的声韵母作为建模单元。HSMM 各个状态的输出分布为单高斯分布。每一帧语音的特征向量由谱、能量、基频以及它们各自的一阶差分及二阶差分构成。其中谱参数为采用 STRAIGHT（Speech Transformation and Representation based on Adaptive Interpolation of weighted spectrogram）[14]提取的 40 阶的线谱对参数（Line Spectral Pair, LSP）[15]。

实验过程中，6700 句陈述句作为初始模型的训练。从 2600 句疑问句中随机选取 100 句作为测试语句。从剩余的 2500 句疑问句中选择 100、300、500、1500、2500 句分别进行自适应训练，得到疑问句模型，在下文中将该方法称为自适应方法。测试语句以及训练语句中都包含数量相当的特指疑问句、是非疑问句、选择疑问句以及一般疑问句。为了进行对比实验，又采用传统的基于 HMM 的统计参数语音合成方法，选择与自适应方法中相同的 100、300、500、1500、2500 句疑问句进行训练，得到疑问句声学模型，在下文中将该方法称为直接训练方法。各组的实验数据如表 1 所示。其中 dlr 表示陈述句语料，qst 表示疑问句语料，M1 为自适应方法各组实验的训练语料，M2 为直接训练方法各组实验的训练语料，共进行五组实验。

表 1　各组实验使用的训练语料说明

	EXP.1	EXP.2	EXP.3	EXP.4	EXP.5
M1	dlr:6700 qst:100	dlr:6700 qst:300	dlr:6700 qst:500	dlr:6700 qst:1500	dlr:6700 qst:2500
M2	qst:100	qst:300	qst:500	qst:1500	qst:2500

3. 2 实验结果分析

各组实验均合成了 100 句疑问句测试语音，从中各随机挑选 20 句作为测听实验语句。选择 20 个专业测听人员，使用 MOS 对测听语句进行评价。MOS（Mean Opinion Score）是目前使用得比较广泛的一种主观评定方法，评分范围是 1 到 5 分。MOS 得分的表现体现在合成语音的自然度上，即合成语音在节奏感和清晰度等方面与自然语音的相近程度。

图 2 为不同疑问句训练语料下两种方法得到的测听语音平均得分。首先，随着语料的逐渐增加，两种方法的 MOS 都有一定的提高。从 MOS 分上分析，直接训练的方法随语料的增长，合成音质有明显的提升。自适应的方法的音质较为稳定，随着自适应语料的增加略有增长，但是没有直接训练的方法提升得明显。并且在 100、300、500 句的疑问句训练量上，自适应方法也能保持较高的音质水平。直接训练的方法在 100、300 句的疑问句训练量下，合成语音的可懂度较差，甚至不能听清语音的内容。主要原因是在直接训练的方法中，模型参数完全依赖于疑问句的训练语料，在数据量较小的情况下，得到的模型不够精确。但是在自适应的方法下，以大规模的陈述句训练得到初始模型，该初始模型本身已经具有一定的精确度，再以该模型作为疑问句自适应训练的输入模型，即使在小规模的疑问句训练语料情况下，得到的模型也具有较高的精确度。

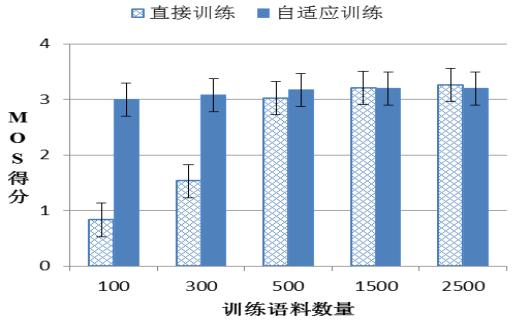

图 2　不同疑问句训练语料下两种方法的 MOS 得分

除了用 MOS 分这一主观评价指标来评价合成语音的音质，实验还设计了一些客观评价指标，根据合成语音声学参数与原始语音声学参数之间的距离，对合成语音的音质进行评价。

图 3 为在不同疑问句训练语料下，谱、基频、时长和清浊判别错误四个客观评价指标的结果。其中（a）图表示的是原始语音的 LSP 与合成语音 LSP 的均方误差（Mean Square Error ,MSE）。从图（a）中可以得到，自适应方法的谱误差小于直接训练方法的谱误差。原因在于自适应训练有大规模陈述句的谱参数作为训练数据，对频谱的分布有更为精确的描述，在此基础上进行自适应训练，生成的谱参数更为精确，与原始语音的谱参数更为接近。（b）（c）分别表示合成语音的基频、时长与原始语音基频、时长的均方根误差（Root Mean Square Error, RMSE），（d）表示生成语音中的清浊音判别错误率，同（a），这三张图也反映了自适应的方法在疑问句语料规模较小的情况下，客观评价指标要优于直接训练的方法。并且随着疑问句训练语料数量的增大，两种方法的客观误差相差不大。与图（a）（c）不同的是，在图（b）（c）中的基频和时长误差在训练语料较大的情况下，直接训练的方法比自适应训练的方法反而会有略微的优势。这是因为，陈述句与疑问句的主要差别体现在韵律上，在声学参数中，主要体现在谱和时长上，自适应训练中的陈述句模型反而会对疑问句的时长和基频有一定的影响。

基频是体现疑问句韵律的一个重要方面。为了分析生成疑问句的基频情况，随机挑选一句以 300 句疑问句为训练语料生成的测试语音，绘制其基频曲线进行分析，如图 4 所示。图(a)为直接训练方法得到的基频曲线与原始语音基频曲线的

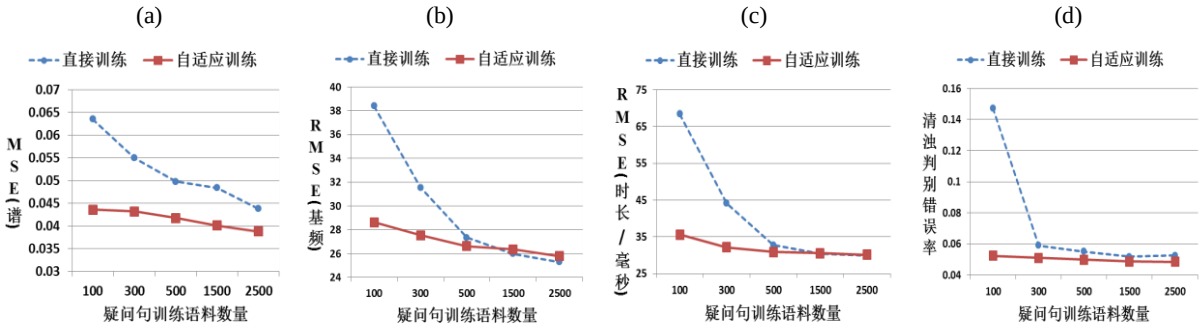

图 3　在不同疑问句训练语料下，谱（a）、基频（b）、时长（c）、清浊判别错误率（d）四个客观评价指标的值

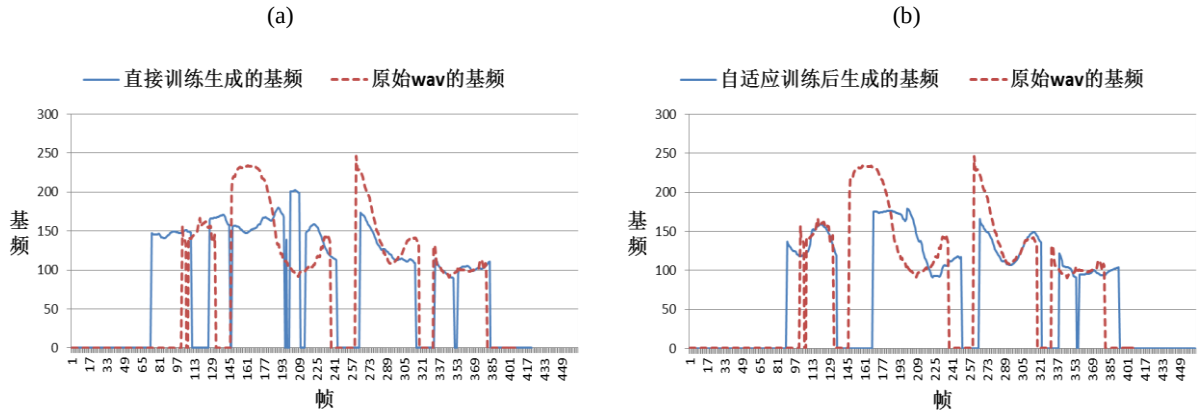

图 4　两种方法得到的基频曲线与原始语音基频曲线对比图

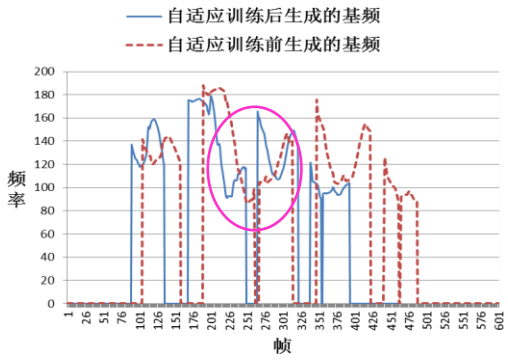

图 5　自适应训练前后基频曲线对比图

对比，图(b)为自适应方法得到的基频曲线与原始语音基频曲线的对比。从两组图对比可以得出，在 300 句疑问句训练语料的情况下，自适应方法得到的基频曲线更好地复现了原始基频曲线，与原始基频曲线更为相近，这也再次验证了主观 MOS 分以及客观评价指标的结果。图 5 为自适应前的初始模型和自适应后的疑问句模型各自生成的基频曲线对比图。图中椭圆部分是自适应前后基频改变最为明显的地方，与图 4 中的原始基频曲线对比，发现椭圆部分的基频改变趋势与原始基频曲线一致，说明了自适应的有效性。此外，从基频的帧数上也可以发现，自适应后的时长变短。主观上来说，疑问句的速度往往会比陈述句快一些，符合人的听觉感知。

从上述的分析中得出，客观评价指标和主观的 MOS 分评价结果是一致的，两者同时说明了在小规模疑问句训练语料下，自适应的方法优于直接训练的方法。在只有 100 句疑问句训练语料的情况下，相比于直接训练的方法，自适应的方法在 LSP 的均方误差，基频和时长的均方根误差上分别有 31.1%、25.4%、47.8%的提升。随着训练语料的逐渐增加，提升程度会变小。其它训练语料下各指标的相对提升情况如下表所示。

表 2 自适应方法在各客观评价指标上的相对提升

语料量	谱	基频	时长	清浊
100	31.45%	25.45%	47.87%	64.47%
300	21.45%	12.70%	27.00%	13.39%
500	16.06%	2.67%	5.96%	9.64%
1500	17.15%	-1.37%	-0.56%	6.13%
2500	11.42%	-1.91%	-0.52%	7.80%

4 总结

本文提出了一种基于自适应训练的疑问句语音合成方法。在统计参数语音合成方法的框架下，以大规模的陈述句语料训练得到一个初始模型，

将其作为自适应训练的输入模型，以小规模的疑问句语料做自适应训练，得到疑问句模型，实现疑问句语音的合成。实验结果表明，在疑问句训练语料较少的情况下，该方法较好地实现了疑问句的语音合成，并且在自然度和客观评价指标方面该方法都优于采用相同疑问句语料的直接训练合成方法，并且能够达到用大语料疑问句直接训练的效果。因此，该方法可以解决在疑问句训练语料不充足的情况下，传统方法得到的合成语音效果不佳的问题。同时也可将该方法应用于感叹、祈使语气等其他语气的合成，对实现有表现力的语音合成具有重大意义。未来的工作将研究设计针对疑问句语料的标注问题，在上下文信息中突出疑问句与陈述句的不同，使自适应训练能够训练出疑问句更加鲜明的特征，使疑问句的合成语音更加自然。

参 考 文 献

[1] 贾惠彬.汉语感叹句和疑问句的生成方研究[D].北京：中国科学院自动化研究所，2009.
Huibin Jia. Research on Generating Exclamatory and Question Speech in Mandarin[D]. Beijing: Institute of Automation, Chinese Academy of Sciences. (in Chinese)

[2] Hunt A J, Black A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. IEEE, 1996, 1: 373-376.

[3] 吴义坚. 基于隐马尔科夫模型的语音合成技术研究[D].安徽：中国科学技术大学，2006.
Yijian Wu. Research on the HMM-based Trainable TTS[D].Anhui: University of Science and Technology of China.(in chinese)

[4] Takayoshi Yoshimura. Simultaneous Modeling of Phonetic and Prosodic Parameters, and Characteristic Conversion for HMM-Based Text-to-Speech System[D]. Japan: Nagoya Institute of Technology, 2002.

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

[6] K. Tokuda, T.K. , and S. Imai. Speech parameter generation from HMM using dynamic features[C], ICASSP95, 1995.

[7] T. Toda and K. Tokuda. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis[C], Proc. of Euro speech, 2005.

[8] Yamagishi J. Average-voice-based speech synthesis[J]. Tokyo Institute of Technology, 2006.

[9] Gauvain J, Lee C. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains[J]. Speech and Audio Processing, IEEE Transactions on, 1994, 2(2):291 - 298.

[10] Shinoda K, Lee C. A structural Bayes approach to speaker adaptation[J]. Speech and Audio Processing, IEEE Transactions on, 2001, 9(3):276 - 287.

[11] J. Leggetter C, C. Woodland P. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. Computer Speech & Language, 1995, 9(2):171–185.

[12] Yamagishi J, Kobayashi T, Nakano Y, et al. Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2009, 17(1):66 - 83.

[13] P. Dempster A, M. Laird N, B. Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 1977, (1):1-38.

[14] Kawahara H. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited[C]. //Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. IEEE, 1997:1303 - 1306.

[15] Soong F K, Juang B. Line spectrum pair (LSP) and speech data compression[C]. //Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84. IEEE, 1984:37 - 40.

Speech Synthesis of Questions Based on Adaptive Training

FANG Shuo, WEN Zhengqi, WANG Yang, TAO Jianhua

Institute of Automation, Chinese Academy of Science, Beijing 100190, China

Abstract: In order to improve the expressiveness of synthetic speech, a question speech synthesis method based on adaptive training is proposed in this paper. By statistical parametric speech synthesis based on MSD-HSMM, a large corpus of exclamatory sentences is used to train the initial acoustic model. Based on the initial acoustic model, a small corpus of question sentences is adaptively trained to gain the question acoustic model and then synthesize the question speech. Conclusions can drew from the experiments that the subjective and objective evaluations of the synthetic speech by this method is superior to the traditional method, where the same size corpus of question sentences is trained directly. Besides, the results of a relatively small training corpus by this method are comparable with the larger training corpus by the traditional method.

Key words: generation of question sentences; adaptation; speech synthesis