文章编号: 1003- 0077(2006)04- 0075- 07

基于 HMM 的可训练中文语音合成*

吴义坚, 王仁华

(中国科学技术大学, 安徽 合肥 230026)

摘要: 本文将基于 HMM 的可训练语音合成方法应用到中文语音合成。通过对 HMM 建模参数的合理选择和优化, 并基于中文语音特性设计上下文属性集以及用于模型聚类的问题集, 提高其建模和训练效果。从对比评测实验结果来看, 98.5% 的合成语音在改进后其音质得到改善。此外, 针对合成语音节奏感不强的问题, 提出了一种基于状态和声韵母单元的两层模型用于时长建模和预测, 集外时长预测 RMSE 由 29.56ms降为 27.01ms。从最终的合成系统效果来看, 合成语音整体稳定流畅, 而且节奏感也比较强。由于合成系统所需的存贮量非常小, 特别适合嵌入式应用。

关键词: 计算机应用; 中文信息处理; 语音合成; HMM; 可训练语音合成; 时长模型

中图分类号: TP391 文献标识码: A

HMM-based Trainable Speech Synthesis for Chinese

WU Yi-jian, Wang Ren-hua

(University of Science and Technology of China (USTC), Hefei, Anhui 230026, China)

Abstract In this paper, the HMM-based trainable speech synthesis was applied for Chinese application. The appropriate HMM parameters are selected and optimized, and the contextual features and corresponding question set for tree-based HMM clustering are designed by considering the characteristics of Chinese, to improve the effect of HMM modeling and training. From the evaluation results, the preference score of the synthetic speech after the above improvement is 98.5%. Furthermore, in order to improve the rhythm of synthetic speech, a two-level based model is introduced for duration modeling and prediction, and the duration prediction RMSE was improved from 29.56ms to 27.01ms. From the evaluation results of the final system, the synthetic speech is stable, fluent and rhythmed. As the speech synthesis system only requires very small storage, it is specially fit for embedded application.

Key words: computer application; Chinese information processing; speech synthesis; HMM; trainable TTS; duration modeling

1 引言

基于大语料库的拼接合成方法是近年来语音合成中主流方法[1, 2]。其基本原理就是根据输入文本分析得到的信息, 从预先录制和标注好的语音库中挑选合适的单元, 然后拼接得到最终的合成语音。由于最终合成语音中的单元都是直接从音库中复制过来的, 其最大的优势就是在于保持了原始发音人的音质。

在现在合成音质和自然度都不错的情况下, 人们对合成系统提出了更多的需求))) 多样

* 收稿日期: 2005- 07- 15 定稿日期: 2006- 06- 02
基金项目: 国家自然科学基金资助项目 (60475015)
作者简介: 吴义坚 (1981-), 男, 博士研究生, 主要研究方向为语音合成.

化的语音合成, 包括多个发音人、多种发音风格、多种情感表达等。虽然大语料库拼接合成系统的效果不错, 但是也存在不少缺陷, 比如: 合成语音的效果不够稳定, 音库构建周期太长以及合成系统的可扩展性较差等, 这些缺陷明显限制了它在多样化语音合成方面的应用, 因此, 近年来基于隐马尔可夫模型[3] (HMM)的可训练语音合成方法被提出并逐渐得到应用。

针对基于 HMM 的可训练语音合成应用, 包括 IBM[4]、Microsoft[5]和 NII[6]等不同的研究机构提出了几种不同的实现技术和方法, 它们的共同点就是都是基于 HMM 对语音参数进行建模, 然后利用音库数据进行自动训练, 并最终形成一个相应的合成系统。与现在大语料库拼接合成相比, 其优势就在于可以在短时间内, 基本不需要人工干预的情况下自动构建一个新的系统, 而且整个训练过程基本上是不依赖于发音人、发音风格以及情感等因素。

这里我们所采用的方法主要是借鉴 NII[6, 7], 其特点是通过 HMM 对语音参数进行建模和训练, 然后在合成过程中利用训练好的 HMM 进行语音参数的生成, 并通过合成器得到最终的合成语音。在本文中, 我们将该方法应用到中文语音合成, 主要工作与改进包括: 选择合理的 HMM 建模参数, 并基于中文语音特性设计上下文属性集以及相应的属性问题集用于模型聚类, 从而提高了建模和训练效果; 为了解决合成语音节奏感不强的问题, 我们提出了一种基于状态和声韵母的两层时长模型, 通过在训练过程中分别对状态和声韵母时长进行建模, 并结合两者进行最终时长的生成, 从而明显提高合成语音的节奏感。本文的内容组织如下: 第二部分对基于 HMM 的可训练语音合成方法进行整体介绍; 第三部分介绍其在中文语音合成中的应用以及相应的一些改进; 第四部分是对合成系统的效果改进评测; 第五部分进行小结。

2 基于 HMM 的可训练语音合成概述

图 1 是基于 HMM 的可训练语音合成系统的基本流程图, 它可以分为两个阶段: 训练阶段和合成阶段。

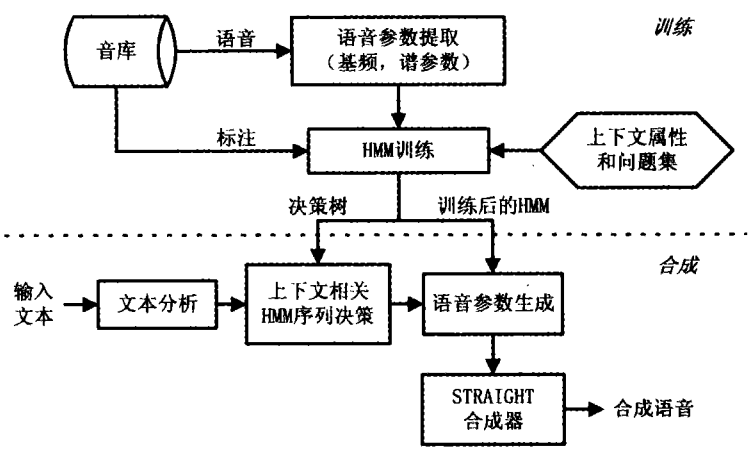


图 1 基于 HMM 的可训练语音合成系统

训练阶段主要包括预处理和 HMM 训练。在预处理阶段, 首先对音库中的语音数据进行分析, 提取相应的语音参数 (基频和谱参数)。根据提取的语音参数, HMM 的观测向量可分为谱和基频两个部分, 其中谱参数部分采用连续概率分布 HMM 进行建模, 而基频部分采用多空间概率分布 HMM (MSD2HMM)[8]进行建模。除此以外, 模型训练前还有一个重要的部分就是对上下文属性集和用于决策树聚类的问题集进行设计, 即根据先验知识来选择一些对声学参数 (谱、基频和时长)有一定影响的上下文属性并设计相应的问题集以用于上下文相关模型

聚类。

在完成预处理流程后, 接着就是整个可训练语音合成系统的核心部分))) HMM 训练, 其训练步骤依次为模型初始化、声韵母 HMM 训练、扩展上下文相关模型训练、聚类后的模型训练以及时长模型训练, 最后得到的训练结果包括谱、基频和时长参数的聚类 HMM 以及各自的决策树。

在合成阶段, 首先输入文本经过文本分析后转换为上下文相关的单元序列, 然后利用训练中得到的决策树对每一个单元进行决策, 得到对应的聚类状态模型, 并形成聚类状态模型序列。最后, 根据文献 [6]中的参数生成算法, 利用参数的动态特性来生成目标声学参数序列, 并通过 STRAIGHT[10]合成器得到最终的合成语音。

3 中文可训练语音合成应用和改进

这里我们将可训练语音合成方法应用到中文语音合成上, 通过选择合理的 HMM 建模参数, 设计合理的上下文属性集和问题集, 以及加入音素时长模型来对合成效果进行改善。

311 建模参数选择

针对中文语音识别应用, HMM建模参数的选择已经比较成熟, 在初期建模中, 我们直接借鉴语音识别中采用的一些参数, 然后通过一些测试试验对其进行优化。

在基于 HMM 的可训练语音合成系统中, HMM 的观测向量包括谱参数和基频参数两部分。对于谱参数部分的选择, 在语音识别中大多都采用 MFCC 参数, 但是由于 MFCC 参数不能完全对频谱进行还原, 因此它并不适合语音合成应用。而从其他基于 HMM 的可训练语音合成系统来看[6 7], 它们采用的是 MCEP(mel2cepstral)参数[11], 因此在最初的系统构建中我们也采用该参数。通过对一些其他谱参数化方法进行分析和对比试验, 我们发现采用 LSP(Line Spectral Pair)参数[12]更适合可训练语音合成的 HMM 建模。图 (2)为一个 MCEP 和 LSP 参数序列的对比示例。

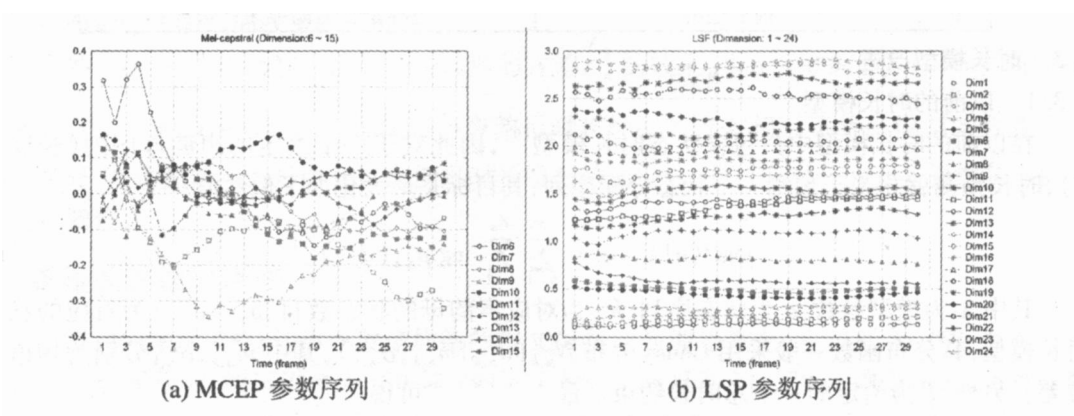


图 2 MCEP和 LSP参数序列对比

对比图 2(a)和图 2(b), 可以看出 LSP参数在高阶和低阶都具有更好的内插特性, 而MCEP参数只在低阶有比较好的内插特性, 在高阶则非常差; 另外, LSP参数与共振峰有直接的对应关系, 当 LSP参数在某个频率周围分布密集时, 其对应为频谱上 (相应频率)的一个峰, 而分布稀疏时, 则对应为频谱上的一个谷, 也就是 LSP参数在频率轴上分布的疏密程度直接对应到频谱的峰和谷; 从另外一个角度来看, LSP参数也具有很好的局部效应, 即某一阶 LSP参数的误差只会影响局部的谱结构, 而不会扩散到整个频谱上。上述这些 LSP参数的特点,

使得其更适合可训练语音合成的 HMM 建模; 而且从测试试验的结果来看 (411 节), 采用 LSP 参数后的 HMM 建模以及合成效果有明显的改善。

在最终 HMM 建模中采用的观测向量为: 谱参数部分由 LSP 参数以及相应的一阶和二阶差分系数构成, 基频部分则由基频对数值以及相应的一阶二阶差分系数构成。

312 上下文属性和问题集设计

由于在实际语流中, 同一个声韵母单元在不同上下文环境下 (比如前后声韵母、调型等), 会产生不同的发音变体, 这里我们采用上下文相关模型对这些发音变体进行建模, 并通过基于决策树的模型聚类方法来提高模型的鲁棒性。因此, 设计合理的上下文属性集和用于模型聚类的问题集, 对建模效果的好坏起到非常关键的作用。

需要注意的是, 由于设计的问题集是用于各种声学参数的模型聚类, 因此需要针对不同的参数变化特性设计对应的问题集, 而上下文属性设计则需要综合所有参数的变化特性来进行设计。比如针对谱参数变化特性, 需要添加前后接声韵母属性, 同时根据不同的前后接声韵母分类设计对应的属性问题集; 而针对基频参数变化特性, 则需要添加前后调类型, 同样也需要根据前后调分类来设计对应的属性问题集。

针对中文语音的特性, 我们采用如表 1所示的上下文属性集, 并设计相应的问题集用于模型聚类。

表 1 上下文属性集和对应的问题集示例

	上下文属性	对应的上下文属性问题示例
1	{前接, 当前, 后接 }声韵母	前接声母是否为擦音?
2	{前接, 当前, 后接 }声调	后接声调是否为 3调?
3	{前接, 当前, 后接 }词性	当前词性是否为虚词?
4	绝对韵律位置信息	当前音节在句中的位置?
5	相对韵律位置信息	当前音节为句头, 句中还是句尾?
6	前后边界	前边界是否为韵律短语边界?

313 时长模型改进

31311 以前的时长模型

在以前的时长模型中只包含状态时长模型[9], 因此对于一个给定的声韵母序列 (长度为 N), 时长预测就相当于预测一个状态分配序列, 其目标是最大化如下似然值:

$$\log P(q|K,T)=\sum_{n=1}^{N}\sum_{k=1}^{K_n}\log p_{n,k}(d_{n,k}) \tag{1}$$

其中 K为模型参数, q为状态序列, K_n 为对应声韵母的状态数目, $p_{n,k}(d_{n,k})$为对应的状态时长模型, 其分布函数一般采用 Gauss分布 $N_{n,k}(d_{n,k}|m_{n,k}, R_{n,k})$, 其中 $m_{n,k}$, $R_{n,k}$分别为均值和方差。另外, T为给定的一个总时长约束。最大化 (1)式可得:

$$d_{n,k}=m_{n,k}+Q\cdot R_{n,k}^2 \tag{2}$$

$$Q=\left[T-\sum_{n=1}^{N}\sum_{k=1}^{K_n}m_{n,k}\right]\Big/\sum_{n=1}^{N}\sum_{k=1}^{K_n}R_{n,k}^2 \tag{3}$$

如果没有总的时长约束 (Q= 0), 则各个状态的时长就是对应状态模型的均值。

31312 改进的时长模型

由于在以前的时长模型中, 过多的考虑了状态时长分配, 而对于声韵母的时长考虑不够, 导致合成语音中, 有些单元的时长过于 / 平均0, 听起来很平淡, 节奏感不强。对此, 我们在状

态时长模型的基础上, 加入一个声韵母时长的决策树模型, 其模型初始化和决策树聚类过程与状态时长模型类似。在合成过程中, 同时对状态时长和声韵母时长模型进行决策, 最后综合两者进行最终的时长生成。

改进后的时长模型可以表述为: 对于一个给定的声韵母序列 (长度为 N), 时长预测模型就相当于预测一个状态分配序列, 其目标是最大化如下似然值:

$$\log P(q \mid K, T) = \sum_{n=1}^{N} \left[\sum_{k=1}^{K_n} \log p_{n,k}(d_{n,k}) + w \log p_n(d_n) \right] \tag{4}$$

其中 $p_n(d_n)$为对应的声韵母时长模型, w 为权重因子, 声韵母时长 d_n 满足:

$$d_n = \sum_{k=1}^{K_n} d_{n,k}, \quad n = 1, 2, \ldots, N \tag{5}$$

声韵母时长模型的分布函数也采用 Gauss分布 $N_n(d_n \mid m_n, R_n)$, 其中 m_n, R_n 分别为均值和方差。

状态时长模型预测相当于最大化 (4)式, 通过 (4)式相对于 $d_{n,k}$ 求偏导可得:

$$\frac{d_{n,k} - m_{n,k}}{R_{n,k}^2} + w \# \frac{d_n - m_n}{R_n^2} = 0, \quad n = 1, 2, \ldots, N \quad k = 1, 2, \ldots, K_n \tag{6}$$

结合 (5)式可得:

$$d_{n,k} = m_{n,k} + Q \# R_{n,k}^2, \tag{7}$$

$$Q = w \# \left[m_n - \sum_{k=1}^{K_n} m_{n,k} \right] / \left[R_n^2 + w \# \sum_{k=1}^{K_n} R_{n,k}^2 \right] \tag{8}$$

当权重因子 w y 0时, 则 Q y 0, d_n y $\sum_{k=1}^{K_n} m_{n,k}$, 当权重因子 w y +] 时, 则 d_n y m_n, 可以明显看出声韵母时长模型的权重因子的效果。

另外, 我们还可以考察声韵母时长模型的方差 R_n^2 的效果, 当 R_n^2 y 0时, 则 d_n y m_n, 而当 R_n^2 y] 时, 则 p_n y 0, d_n y $\sum_{k=1}^{K_n} m_{n,k}$。也就是当声韵母时长模型的方差相对于状态模型的方差而言非常小时, 则预测的声韵母时长以声韵母时长模型的均值为主, 状态时长的分配根据各自的方差进行调整。

4 系统改进效果测试

这里我们采用 1000句声学覆盖均衡的中文语料作为训练数据, 它包括 25, 096个声母和 29, 942个韵母单元, 测试语料为 800句, 包括 17, 860个声母和 21, 389个韵母单元, 所有的数据都经过手工标注。下面我们给出建模参数优化和时长改进的一些对比测试试验结果。

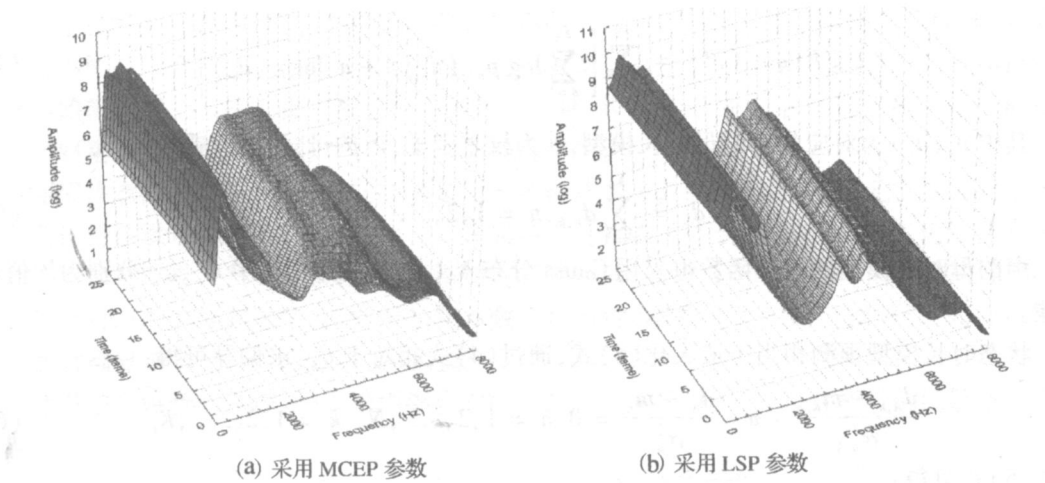
411 采用 MCEP和 LSP 参数的效果对比

我们分别采用 24阶 MCEP和 24阶 LSP参数进行建模训练, 并构建相应的合成系统, 针对以下几个方面进行对比评测:

a1从聚类模型后模型的个数来看, 采用 LSP参数的聚类模型为 2832个, 而采用 MCEP参数的聚类模型个数为 2080个, 这说明采用 LSP参数后的模型对谱参数描述更为精细;

b1通过对比分析一些合成语音的语谱图 (图 3为一个示例), 可以发现采用 MCEP参数的合成共振峰被过度平滑, 而 LSP参数的合成语音的共振峰则更清晰;

c1从主观听感上来看, 采用 LSP参数的合成语音显得更为清晰。为了确认其效果, 我们进行了一个系统的对比评测试验。首先从测试数据中随机挑选了 50句, 分别采用这两个系统合成, 然后给 8个评测人员进行两两对比测听, 主观感知效果更好的一句打分为 1, 而另一句则打分为 0。最终的评测结果是采用 LSP参数的合成系统平均得分为 01985, 远远高于采用 MCEP参数构建的系统平均得分 01015。



(a) 采用 MCEP 参数　　(b) 采用 LSP 参数

图 3　合成语谱图对比

从上述对比结果可以看出, 在采用 LSP参数后, HMM 建模以及合成效果有明显的改善, 因此在后面的试验以及最终系统的构建我们都采用 LSP参数。

412　时长模型改进效果

从时长模型改进的公式 (7)和 (8)可以看出, 通过设置不同的声韵母时长模型权重因子可以产生不同的时长预测效果。通过一些初期的权重因子测试试验, 我们最终采用的声韵母时长权重因子为 5。我们从以下两个方面对时长模型改进效果进行评测:

表 2　时长和基频预测的集外测试结果

	时长预测 RMSE(ms)	基频预测 RMSE(Hz)
改进前	29156	31118
改进后	27101	30139

a1从时长模型改进前后对集外数据的时长和基频预测效果来看 (见表 2), 在经过时长模型改进后, 不仅时长预测的 RMSE误差由 29156ms降到 27101ms, 而且基频预测的效果也得到一定的改善。

b1从主观听感来看, 加入声韵母时长模型后, 合成语音的节奏感有明显的改善。同样我们进行了一个系统的对比评测, 随机选取 50句集外测试语料, 分别采用改进前后改进后的系统进行合成, 给 8个评测人员进行两两对比测听。最终的评测结果是时长模型改进后的系统平均得分为 01807, 远高于改进前的系统平均得分 01193。

由此看出, 经过对时长模型的改进后, 合成系统的效果有比较明显的改善, 合成语音的节奏感明显增强。

5　小结

本文将基于 HMM的可训练语音合成方法应用到中文语音合成。通过对 HMM 建模和声学参数的优化, 并基于中文语音特性设计上下文属性集以及用于模型聚类的问题集, 来提高其建模和训练效果。另外, 针对合成语音节奏感不强的问题, 我们提出了一种基于状态和声韵母的两层模型用于时长建模和预测。根据设计好的上下文属性和问题集, 并结合优化的建模和

声学参数,以及改进后的时长模型,我们实现了模型训练以及合成系统的构建。从最终合成语音的效果来看,自然度比较高,整体感觉非常流畅,而且韵律节奏感也比较强。

本文介绍的方法是采用参数预测和参数合成器生成合成语音,与大语料库拼接合成系统相比,虽然其音质上要差一些,主要是因为拼接合成中的拼接单元基本上都是采用原始语音;但是它的优势在于合成语音的更为流畅和平滑,而且对不同句子的合成效果也更为稳定。而且可训练语音合成所特有的可训练性、易扩展性、灵活性等等都使得它在今后多样化语音合成中更具有竞争力。从应用角度来看,两种方法也各有所长:大语料库拼接合成系统需要存储大量的语音库资源,而且需要大量的运算进行单元搜索,因此适合服务器级的应用,虽然可以对音库进行裁减以及采用语音参数化编码的方式进行音库压缩,以达到嵌入式语音合成的需求,但是语音合成的音质和自然度都会有所下降;而可训练语音合成系统只是在训练阶段需要访问音库数据进行模型训练,而在合成阶段中只需访问训练好的模型数据,因此该系统所需的存储量非常小,比较适合嵌入式方面的应用。

参 考 文 献:

[1] R. H. Wang, Qingfeng Liu, Deyu Xia. Towards A Chinese Text2To2Speech System With Higher Naturalness [A]. In: Proc of ICSLP [C]. Sydney, 1998, p2047- 2050.

[2] R. H. Wang, Zhongke Ma, Wei Li, Donglai Zhu. A Corpus2Based Chinese Speech Synthesis with Contextual2Dependent Unit Selection[A]. In: Proc of ICSLP [C]. Beijing, 2000, p391- 394.

[3] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proc of IEEE, 1989 [J]. vol 77, pp 257- 286.

[4] R. E. Donovan and E. M. Eide. The IBM trainable speech synthesis system[A]. In: Proc of ICSLP [C]. Sydney, 1998, vol 5, pp. 1703- 1706.

[5] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe. Recent improvements on Microsoft's trainable text2to2speech system – Whistler[A]. In: Proc of ICASSP [C]. Munich, 1997, pp. 9592962.

[6] T. masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis from HMMs using dynamic features[A]. In: Proc of ICASSP[C]. Atlanta, 1996, pp 389- 392.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM2based speech synthesis[A]. In: Proc of Eurospeech [C]. Budapest, 1999, vol 5, pp 2347- 2350.

[8] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi2space probability distribution for pitch pattern modeling. In: Proc of ICASSP [C]. Arizona, 1999, pp. 229- 232.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura. Duration modeling in HMM2based speech synthesis system[A]. In: Proc of ICSLP [C]. Sydney, 1998, vol 2, pp. 29- 32.

[10] H. Kawahara, I. Masuda2Katsuse and A. deCheveigne. Restructuring speech representations using pitch2adaptive time2frequency smoothing and an instantaneous2frequency2based F0 extraction: possible role of a repetitive structure in sounds. Speech Communication [J]. 1999, vol 27, pp. 187- 207.

[11] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai. An adaptive algorithm for mel2cepstral analysis of speech [A]. In: Proc of ICASSP [C]. 1992, vol 1, pp. 137- 140, 1992.

[12] F. Itakura. Line spectral representation of linear predictive coefficients. Journal of Acoustic Society of America [J]. 1990, vol 87(4), pp. 1738- 1752.