科学通报 2009 年 第 54 卷 第 8 期: 1133 ~ 1138
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

基于统计建模的可训练单元挑选语音合成方法

王仁华, 戴礼荣, 凌震华, 胡郁

中国科学技术大学电子工程与信息科学系讯飞语音实验室, 合肥 230027
E-mail: rhw@ustc.edu.cn

2008-11-18 收稿, 2009-03-25 接受
国家自然科学基金(批准号: 60475015, 60610298)和国家高技术研究发展计划(编号: 2006AA01Z137, 2006AA010104)资助项目

摘要 提出了一种基于统计建模的可训练单元挑选语音合成方法. 在模型训练阶段, 提取训练语料库中的多种声学参数并训练各自对应的统计模型; 在合成阶段, 基于统计模型的最大似然准则实现语料库中最优备选单元序列的挑选; 最终通过波形拼接输出合成语音. 实验结果表明, 该方法可以有效改善传统单元挑选与波形拼接语音合成方法在系统构建自动化程度低、对专家知识依赖性强、以及合成效果稳定性不足等方面的问题. 此外, 针对单元挑选语音合成的特点,提出了一种新的最小单元挑选错误准则, 采用区分性模型训练方法进行模型参数的更新, 实现了系统构建的全自动化, 并进一步提高了合成语音的自然度.

关键词
语音合成
单元挑选与波形拼接
统计模型
最大似然准则

基于大语料库的单元挑选与波形拼接技术是现今最为常见的一种语音合成方法[1,2]. 其基本思想就是基于对输入的待合成语句的文本分析结果, 从一个预先录制好的语料库中挑选合适的单元序列, 将其波形拼接得到最终的合成语句. 由于使用了自然的语音波形, 合成语音的音质可以得到保证; 并且随着语料库规模的不断增大, 合成语音的自然度也显著提高. 为了在合成过程中实现语音单元的合理选择, 需要计算各个备选单元相对于合成目标单元的目标代价, 以及前后备选单元之间的连接代价, 并通过动态规划算法来进行最优备选序列的搜索[1]. 在传统的单元挑选算法中, 目标代价与连接代价的计算往往是通过度量单元间的上下文属性的差异或者备选单元声学参数与预测目标参数之间的距离来实现的. 这样造成代价函数的设计需要语种相关的语音学专家的参与, 进行大量的手工调试, 这使得系统构建的自动化程度受到限制; 并且设计的代价函数难以保证普适性, 往往会带来合成效果不稳定的问题.

近年来, 隐马尔科夫模型(hidden markov model, HMM)这一在语音识别领域得到成功应用的声学建模方法, 也在语音合成系统中得到了使用. 一方面, HMM可以被用于实现波形拼接合成系统中音库的自动构建以及备选样本的聚类[3]; 另一方面, 一种基于HMM的参数语音合成方法得到了迅速的发展[4~6]. 该方法对频谱、基频等声学参数进行HMM建模, 并且通过最大似然参数生成算法[5]来实现合成时的参数预测, 最终经过参数合成器生成语音. 整个系统可以实现训练的自动化和语种的无关性, 并且合成语音的连续性、稳定性和韵律的自然度都相当高. 但是由于参数合成器的限制, 使得这种合成方法最终恢复语音的音质往往不够理想.

为此本文提出将对声学参数的统计建模引入到单元挑选与波形拼接合成中, 以综合统计建模方法在自动训练、灵活性和稳定性方面的长处, 以及波形拼接合成在输出语音音质上的优点, 最终在统计建模框架下实现整个单元挑选语音合成系统的可训练化. 该算法分为模型训练和语音合成两个阶段, 如图1 所示. 在模型训练阶段, 我们利用语料库中各音素单元包含的声学参数以及对应的音段、韵律等标注属性, 进行统计模型的训练. 在合成阶段, 首先对输入的文本进行分析, 得到目标合成句中各个音素的上下文属性描述, 并依此去训练好的模型集合中决策

引用格式: 王仁华, 戴礼荣, 凌震华, 等. 基于统计建模的可训练单元挑选语音合成方法. 科学通报, 2009, 54(8): 1133~1138
Wang R H, Dai L R, Ling Z H, et al. Trainable unit selection speech synthesis under statistical framework. Chinese Sci Bull, 2009, 54(11): 1963−1969, doi: 10.1007/s11434- 009-0267-3

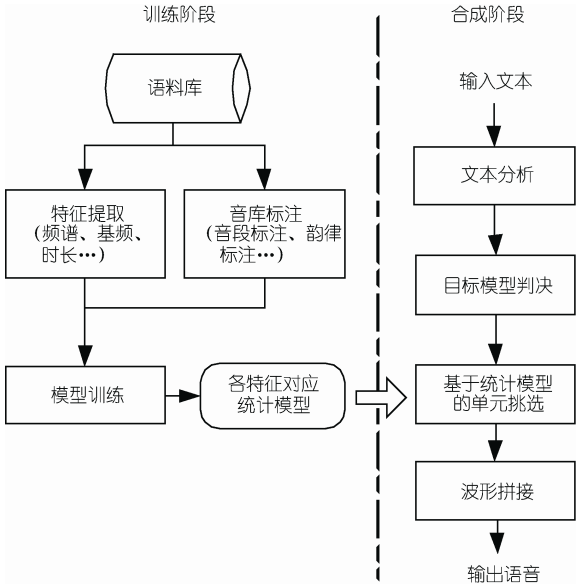

图 1 基于统计建模的可训练单元挑选语音合成算法流程

其对应的声学模型，然后经过基于最大似然准则的单元挑选，得到最优备选单元序列，经过波形拼接输出合成语音. 在统计模型的训练准则方面，本文还提出了最小单元挑选错误准则，采用区分性模型训练方法进行模型参数的更新，更好地满足了单元挑选语音合成的需求.

1 基于统计建模的单元挑选语音合成方法

1.1 模型训练

首先，定义一组能够反映语音合成系统性能的声学特征，例如各个音素单元对应的频谱、基频、时长特征等，令选择的特征种类数为 M. 模型训练阶段的目标就是利用语料库训练这 M 种声学特征对应的统计模型 $\{\Lambda_1,\cdots,\Lambda_M\}$. 设语料库中包含的句子数为 R; 其中第 r 句话对应的音素单元序列记为 $\boldsymbol{V}_r=\{v_{r1},\cdots,v_{rN_r}\}$; v_{rn} 为第 r 句话中的第 n 个音素; N_r 为该句话包含的音素数目. 同时设 C_r 表示单元序列 $\boldsymbol{V}_r$ 对应的上下文描述信息，这些上下文描述信息包括对于该句中各音素的音段信息以及韵律特征的描述，可以通过对语料库的自动或人工标注得到. 基于最大似然(maximum likelihood, ML)准则进行每一种声学特征对应统计模型 Λ_m 的训练，如(1)式所示.

$$\Lambda_m^*=\arg\max_{\Lambda_m}\sum_{r=1}^{R}\lg P\left(X(\boldsymbol{V}_r,m)\,|\,\Lambda_m,C_r\right),\qquad(1)$$

其中 $X(\boldsymbol{V}_r,m)$， $m=[1,\cdots,M]$ 表示提取得到的单元序列 $\boldsymbol{V}_r$ 的第 m 种声学特征.

对于不同的特征，需要使用不同的模型形式加以描述. 例如，对于频谱特征，可以借用语音识别中常用的隐马尔科夫模型来表示; 对于基频特征，基于多空间概率分布(multi-space probability distribution, MSD)[7]的HMM是一种更加合理的模型形式，它可以有效地对清音段的基频缺失现象进行描述; 对于音素的时长，则使用高斯混合模型(Gaussian mixture model, GMM)来建模. 同时，为了反映各种语音特征的分布随着不同上下文环境所发生的变化，这里训练的统计模型均是上下文相关的，即(1)式中计算的条件概率依赖于当前语句的上下文描述信息 C_r. 为了解决在上下文相关模型训练过程中存在的数据稀疏问题，引入基于决策树的模型聚类方法，来保证估计得到的模型参数的鲁棒性[8].

1.2 语音合成

在合成阶段，对于输入的待合成语句文本首先通过文本分析模块得到其对应的上下文描述信息 C. 在单元挑选时，保证挑选得到的最优备选单元序列 $\boldsymbol{U}^*$ 所对应的声学参数，相对训练得到的统计模型 $\{\Lambda_1,\cdots,\Lambda_M\}$ 有最大的似然值，即

$$\boldsymbol{U}^*=\arg\max_{\boldsymbol{U}}\sum_{m=1}^{M}w_m\lg P\left(X(\boldsymbol{U},m)\,|\,\Lambda_m,C\right),\qquad(2)$$

其中 w_m 为不同特征统计模型对应的权值，由手工设定. 针对具体的特征选择，(2)式可以转换为传统的目标代价与连接代价加和的形式，从而通过动态规划算法实现最优单元序列的搜索[9]. 相对于传统的单元挑选算法，这里的代价函数由统计模型自动导出，需要进行的手工调试很少，从而极大地减少了系统构建过程中对于语种相关专家知识的依赖性. 同时基于统计模型的单元挑选准则也更好的保证了合成效果的稳定性. 当合成使用的语料库规模较大时，为了提高单元挑选过程的运算效率，可以利用训练得到的统计模型，基于Kullback-Leibler距离进行备选单元的快速预选[9].

在传统的单元挑选算法中，往往会通过引入连接代价，以得到尽量长串的拼接单元，减少实际拼接点并保证最终合成语音的自然度. 而在这种基于统计建模的单元挑选算法中，虽然没有对于连接代价的显式定义，但是我们可以通过在特征提取时加入频谱、基频等声学参数的帧间动态特征，并且在依据

(2)式进行单元搜索时计算音素拼接处动态声学参数的似然值，来实现对于样本间连接代价的计算，从而达到挑选尽量连续的拼接单元的目的.

在单元挑选完成之后，对挑选得到的合成单元通过波形拼接的方法生成最终语音. 这里对于相邻音素边界处的波形拼接，采用了平移加窗叠加的方法. 首先通过对拼接处前后两帧进行平移以搜索波形相关系数最大时对应的平移位置，然后对平移后的波形进行时域的加窗叠加以实现拼接处的平滑过渡[10].

2 最小单元挑选错误模型训练

2.1 算法提出

在以上介绍的基于统计建模的可训练语音合成方法中，使用最大似然准则进行各声学特征对应概率模型的训练，如(1)式所示. 但是该模型训练准则存在以下两点不足：(ⅰ) 无法实现完全自动的系统训练. 在(2)式中，综合不同特征对应统计模型的似然值时使用的权重 w_m 无法通过最大似然准则进行自动训练，需要手工设定. (ⅱ) 不能保证模型训练准则和单元挑选合成目标的一致性. 后者是希望通过单元挑选，合成出尽量自然的合成语音，即希望合成语音与自然语音尽量接近；但在模型训练中使用的最大似然准则与此目标并没有直接的关系.

因此，本文提出一种新的模型训练准则来改进以上两点不足. 新的准则能够在训练集上对单元挑选合成语音的效果进行整体性的评估，从而实现在综合不同模型时对模型间权重进行合理估计，以及保证模型训练准则与单元挑选合成目标的一致性.

令 $\Phi=\{\Lambda_1,\cdots,\Lambda_M,w_1,\cdots,w_M\}$ 表示所有需要估计的模型参数集合，同时将(2)式简写为

$$\boldsymbol{U}^*=f_{syn}\left(\Phi,C\right), \tag{3}$$

则此时的模型训练准则可以描述为

$$\Phi^*=\arg\max_{\Phi}\sum_{r=1}^{R}f_{eva}\left(f_{syn}\left(\Phi,C_r\right),\boldsymbol{V}_r\right), \tag{4}$$

其中函数 $f_{eva}(\cdot)$ 用以评估挑选得到的最优备选单元序列 $\boldsymbol{U}^*$ 相对音库中的真实样本序列 $\boldsymbol{V}_r$ 的接近程度. $f_{eva}(\cdot)$ 可以有多种不同的实现形式，这里考虑其一种最为简单的实现形式，即

$$f_{eva}\left(\boldsymbol{U},\boldsymbol{V}\right)=\begin{cases}1, & \boldsymbol{U}=\boldsymbol{V},\\ 0, & \boldsymbol{U}\neq\boldsymbol{V}.\end{cases} \tag{5}$$

该准则希望在合成训练语料库中的语句时，挑选和真实单元序列尽可能一致的备选序列，我们称这种准则为最小单元挑选错误 (Minimum Unit Selection Error, MUSE) 准则.

2.2 算法实现

由于MUSE准则与语音识别中使用的最小分类误差(MCE)准则[11]在定义上较为类似，可以使用相似的区分性训练方法来实现模型参数的估计. 首先，对于给定的模型集合 Φ，定义使用备选单元序列 $\boldsymbol{U}$ 合成训练语料库中上下文属性为 C_r 的一句话时的区分函数(Discriminant Function)为

$$g(C_r,\boldsymbol{U};\Phi)=\sum_{m=1}^{M}w_m\lg P\left(X(\boldsymbol{U},m)\,|\,\Lambda_m,C_r\right). \tag{6}$$

此时(2)式表示的单元挑选过程可以改写为

$$\boldsymbol{U}^*=\arg\max_{\boldsymbol{U}}g(C_r,\boldsymbol{U};\Phi). \tag{7}$$

为了描述上式中的最大化决策过程，引入错分度量函数(Misclassification Measure)，如下

$$d(C_r;\Phi)=-g(C_r,\boldsymbol{V}_r;\Phi)+g(C_r,\overline{\boldsymbol{U}};\Phi), \tag{8}$$

其中

$$\overline{\boldsymbol{U}}=\arg\max_{\boldsymbol{U}\neq\boldsymbol{V}_r}g(C_r,\boldsymbol{U};\Phi). \tag{9}$$

$d(C_r;\Phi)<0$ 表示挑选得到的最优备选单元序列 $\boldsymbol{U}^*$ 与自然单元序列 $\boldsymbol{V}_r$ 一致，没有挑选错误；$d(C_r;\Phi)>0$ 则表示存在单元挑选错误. 进一步，将 $d(C_r;\Phi)$ 转换为[0,1]间平滑的损失函数(Loss Function)，如下式所示，其中 γ 控制S形映射曲线的平滑程度，

$$l(C_r;\Phi)=\frac{1}{1+e^{-\gamma d(C_r;\Phi)}}. \tag{10}$$

最终，MUSE 准则可以表示为最小化(11)式所示的训练语料库上的全局经验损失(Empirical Loss)

$$L\left(\Phi\right)=\frac{1}{R}\sum_{r=1}^{R}l(C_r;\Phi). \tag{11}$$

我们使用广义概率下降(generalized probabilistic descent, GPD)[12]算法来实现以最小化(11)式为准则的模型参数优化. GPD通过迭代更新实现，对于训练语料库中的每句话，参数更新公式如下

$$\Phi(r+1)=\Phi(r)-\varepsilon_r\nabla l(C_r;\Phi)|_{\Phi=\Phi(r)}, \tag{12}$$

其中 ε_r 为迭代更新的步长. 具体的参数更新公式依据具体的特征选择而不同，在文献[13]中给出了一个示例. 通过这样的迭代更新，不仅可以实现对于模型分布参数的优化，也可以实现对于模型权值 w_m 的估计，从而达到系统构建完全自动化的目的.

3 实验与评测

3.1 系统构建

实验使用的合成语料库为一个专业播音员录制的中文女声音库，共包含 13000 句，约 20 h的语音数据(16 kHz采样，16 bit量化)，同时具有对应的音段与韵律标注信息．我们使用的韵律标注主要包括对于中文的韵律词、韵律短语和主短语的划分．这些标注信息通过基于文本的自动韵律预测得到，并进行了手工修正．我们选择三种声学特征进行统计模型的训练(M = 3)，分别为频谱、基频与音素时长特征．针对中文语音的特点，这里的音素以声韵母来表示．我们提取训练语音数据的 13 阶mel倒谱[14]和对数基频数值作为频谱和基频参数，参数分析的帧移为 5 ms．最终的频谱和基频特征不仅包含静态参数，还包含一阶和二阶的时域差分参数，以更好的描述语音的动态特征．对于频谱和基频特征，使用的HMM为 5 状态从左至右无跳转的模型结构，对于音素时长采用单高斯分布来表示．在各特征对应的上下文相关统计模型的训练过程中，我们使用决策树来进行模型聚类以解决数据稀疏的问题．这里，基于中文语音的特点来进行决策树问题集的设计[6]．

在模型训练准则方面，本文分别尝试了最大似然(ML)准则和第 3 节中介绍的最小单元挑选错误(MUSE)准则．在进行MUSE模型训练时，我们以ML训练得到的模型参数作为初始值，依据(12)式在训练数据集上进行了 10 遍模型参数的更新．在更新过程中，分别尝试了只更新模型权值 $\{w_1,\cdots,w_M\}$ 和同时更新模型权值与模型分布参数两种情况．最终，我们训练得到 3 个合成系统，分别记为：

(ⅰ) ML：基于最大似然准则训练；

(ⅱ) MUSE-W：基于 MUSE 准则训练，只更新模型权值；

(ⅲ) MUSE-ALL：基于 MUSE 准则训练，同时更新模型权值与模型分布参数．

图 2 给出了 MUSE-W 和 MUSE-ALL 训练过程中，训练集上以音素为单位统计的单元挑选错误率随迭代次数的变化情况．从中可以看出，通过 MUSE 训练，可以有效降低训练数据集上的单元挑选错误率；同时更新模型权值和模型参数能够取得比只更新模型权值更大的错误率下降．

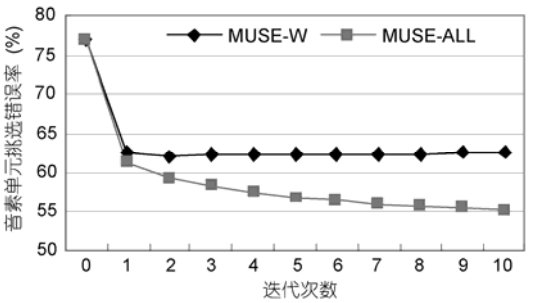

图 2 MUSE 训练中音素单元挑选错误率变化情况

3.2 最大似然模型训练系统性能

用作对比的基线系统为一个使用同样音库的基于传统代价函数构建的单元挑选与拼接合成系统，并且该系统的相关代价表已经较多的专家手工调试．我们选择了集外的 44 句合成语音，分别由基线系统和 ML 系统进行合成．每句合成语音均由 5 名测听人员进行针对自然度的 5 分制评分，分数越高表示越自然．统计两个系统的平均得分(mean opinion score, MOS)如图 3 所示．由图中可以看出，在使用了基于统计建模的可训练语音合成方法，并采用最大似然准则进行模型训练后，合成语音的自然度相对基线系统得到了显著的提升．同时，ML 系统相对基线系统，在构建过程中需要的人工调试也大大减少．

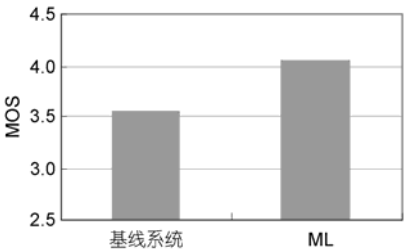

图 3 最大似然准则训练系统的自然度评测结果

3.3 最小单元挑选错误模型训练系统性能

为了测试提出的最小单元挑选错误准则应用于可训练单元挑选语音合成系统的有效性，进行了对比最大似然模型训练的倾向性主观测听实验．测试过程同样由 5 名测听人员进行．使用 ML，MUSE-W 和 MUSE-ALL 系统分别合成集外的 44 句测试文本．对于每一句测试文本，由测听人员在 ML 与 MUSE-W 以及 ML 与 MUSE-ALL 中各选择一句认为较为自然的合成语音．最后计算得到 MUSE-W 相对 ML 系统以及 MUSE-ALL 相对 ML 系统的平均倾向性百分

比，如图 4 和 5 所示，图中误差线表示每个系统的倾向性百分比均值的 95%置信区间.

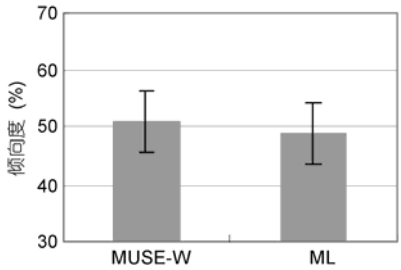

图 4　MUSE-W 与 ML 系统间自然度倾向性测试结果

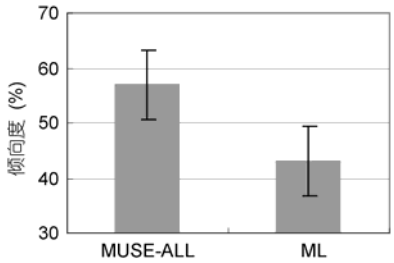

图 5　MUSE-ALL 与 ML 系统间自然度倾向性测试结果

从这两张图中可以看出，MUSE 准则用于模型训练相对 ML 准则可以进一步提升合成语音的自然度. 在只对模型权值进行 MUSE 训练时，这种提升并不明显；如果同时进行模型权值和模型分布参数的更新，则可以显著提升合成语音的效果.

3.4　Blizzard Challenge 2008 国际语音合成评测

本文所提方法的有效性，在 Blizzard Challenge 国际语音合成评测活动中得到了进一步的确认. Blizzard Challenge 旨在通过对使用同一数据集、不同技术构建的语音合成系统进行统一的评测，来推动语音合成技术的发展. 我们使用上述方法构建英文合成系统参加了 2007 年和 2008 年的 Blizzard Challenge 评测. 在这两次评测活动中，我们提交的参测系统均有良好的性能表现 [1,2].

以 2008 年的Blizzard Challenge为例，共有来自全球的 19 家科研机构与公司参加了此次的评测活动. 活动要求各个参测单位在约 8 周的时间内使用一个发布的 15 h的英文音库构建合成系统并合成要求的测试文本. 在各单位提交测试句后，评测工作由活动组织者统一进行. 测试的项目包括合成语音相对原始语音的相似度、合成语音的自然度以及可懂度. 其中相似度和自然度使用 5 分制的评分，分数越高表示系统性能越好；可懂度使用单词听写错误率表示，错误率越低表示可懂度越高. 整个测试活动基于网络进行，测试人员包括语音专家、英语母语学生以及网络上的志愿者. 图 6~8 给出了使用英文 15 h音库时，所有系统合成语音相似度、自然度与可懂度的评测结果，其中我们提交的系统编号为J. 系统A为自然语音；系统B为单元挑选与波形拼接合成方法的基准系统，使用Festival[3]构建；系统C为基于HMM的参数语音合成方法的基准系统，使用HTS[15]构建. 从这几张图中可以看出，在使用 15 h英文音库时，我们提交的参测系统在合成语音相似度和自然度指标方面是所有系统中最好的；进一步的Wilcoxon符号秩检验(α = 0.01)结果表明这种性能上的优势是显著的 [4]；而在可懂度方面该系统也有良好的表现，单词听写

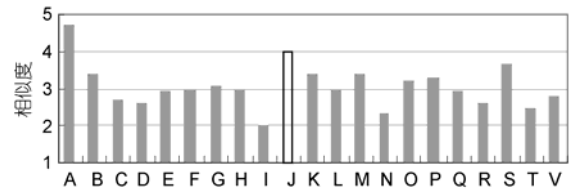

图 6　Blizzard Challenge 2008 相似度评测结果

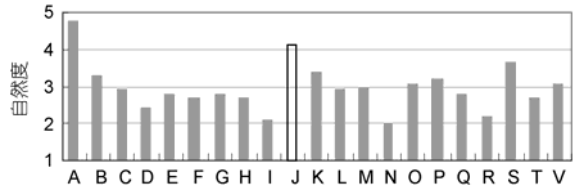

图 7　Blizzard Challenge 2008 自然度评测结果

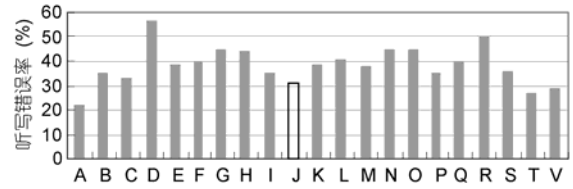

图 8　Blizzard Challenge 2008 可懂度评测结果

1) Ling Z H, Qin L, Lu H, et al. The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007. In: Proceedings of Blizzard Challenge workshop, 2007

2) Ling Z H, Lu H, Hu G P, et al. The USTC system for Blizzard Challenge 2008. In: Proceedings of Blizzard Challenge workshop, 2008

3) Richmond K, Strom V, Clark R, et al. Festival Multisyn voices for the 2007 Blizzard Challenge. In: Proceedings of Blizzard Challenge workshop, 2007

4) Karaiskos V, King S, Clark R, et al. The Blizzard Challenge 2008. In: Proceedings of Blizzard Challenge workshop, 2008

错误率均低于两个基准系统.

4 结论

本文提出了一种基于统计建模的可训练单元挑选语音合成算法. 通过选择合理的声学特征、利用训练语料库进行各特征上下文相关统计模型的训练、以及合成时基于目标模型最大似然准则的单元挑选, 实现了整个单元挑选语音合成系统的自动构建. 此外, 针对单元挑选语音合成应用的特殊性, 我们又在最大似然准则基础上, 提出了新的最小单元挑选错误准则, 采用区分性模型训练方法进行模型参数的更新. 在中文音库上的实验结果表明, 使用最大似然准则构建的可训练单元挑选语音合成方法相对传统的基于代价函数的单元挑选算法, 可以提高合成语音自然度约 0.5 MOS 分. 在引入最小单元挑选错误准则进行模型训练后, 合成语音的自然度得到了进一步的提升. 此外, 在这种可训练的单元挑选语音合成算法中, 只有语料库的上下文属性标注以及模型聚类决策树的问题集设计部分是语种相关的, 其他部分均为语种无关并且对专家经验和人工调试的依赖较小, 这也使得这种方法非常适合在多语种语音合成上的应用. 在 2007 年和 2008 年的 Blizzard Challenge 国际语音合成评测活动中, 基于此项技术构建的英文参测系统均取得了良好的整体性能表现.

然而, 本文只是以在统计框架下实现单元挑选语音合成的可训练化为目标进行了初步的探索工作, 在该研究方向上仍然存在许多值得深入探究的课题. 例如在语音生成过程中将单元挑选与参数合成相结合、在单元尺度上尝试比音素更小的拼接单元(如状态和帧)、对模型训练准则的进一步优化等. 这些都将是我们后续研究的重点内容.

参考文献

1 Hunt A, Black A. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996, Atlanta. 373—376

2 Wang R H, Ma Z K, Li W, et al. A corpus-based Chinese speech synthesis with contextual dependent unit selection. In: Proceedings of International Conference on Spoken Language Processing, 2000, Beijing. 391—394

3 Donovan R. Trainable speech synthesis. Doctoral Dissertation. Cambridge: Cambridge University, 1996

4 Yoshimura T, Tokuda K, Masuko T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: Proceedings of Eurospeech, 1999, Budapest. 2347—2350

5 Tokuda K, Yoshimura T, Masuko T, et al. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, Istanbul. 1315—1318

6 吴义坚, 王仁华. 基于 HMM 的可训练中文语音合成. 中文信息学报, 2006, 20(4): 75—81

7 Tokuda K, Masuko T, Miyazaki N, et al. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999, Phoenix. 229—232

8 Shinoda K, Watanabe T. MDL-based context-dependent subword modeling for speech recognition. J Acoust Soc Jpn (E), 2000, 21(2): 79—86[doi]

9 Ling Z H, Wang R H. HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, Honolulu. 1245—1248

10 Hirai T, Tenpaku S. Using 5 ms segments in concatenative speech synthesis. In: Proceedings of 5th ISCA Speech Synthesis Workshop, 2004, Pittsburgh. 37—42

11 Juang B, Chou W, Lee C. Minimum classification error rate methods for speech recognition. IEEE T Speech Audi P, 1997, 5: 257—265[doi]

12 Blum J. Multidimensional stochastic approximation method. Ann Math Stat, 1954, 25: 737—744[doi]

13 Ling Z H, Wang R H. Minimum unit selection error training for HMM-based unit selection speech synthesis system. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008, Las Vegas. 3949—3952

14 Fukada T, Tokuda K, Kobayashi T, et al. An adaptive algorithm for mel-cepstral analysis of speech. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992, San Francisco. 137—140

15 Zen H, Toda T, Nakamura M, et al. Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE T Inf Syst, 2007, E90-D(1): 325—333[doi]