ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 385－392
DOI:10.16337/j.1004-9037.2016.02.020
© 2016 by Journal of Data Acquisition and Processing
http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

融合自动检错的单元挑选语音合成方法

孙晓辉　凌震华　戴礼荣

(中国科学技术大学语音及语言信息处理国家工程实验室,合肥,230027)

摘　要:提出了一种融合自动检错的单元挑选语音合成方法。本文方法旨在设计与主观听感更加一致的单元挑选准则,以提高合成语音的自然度。首先利用众包网络平台快速大量地收集测听人对于合成语音的主观评价数据,取代了传统的利用具备语言学知识的专家收集主观评价数据的方法;然后基于这些主观评价数据,提取对应语音的音节时长、单元代价以及声学参数距离等特征,构建基于支持向量机的合成错误检测器;在合成阶段,该检测器被用来对传统单元挑选输出的 N 条路径行重打分,以确定最优的单元挑选序列。倾向性测听结果表明本文方法可以有效地提高合成语音的自然度。

关键词:语音合成;单元挑选;支持向量机;众包;合成错误检测

中图分类号:TN912.33　　**文献标志码**:A

Unit Selection Speech Synthesis Integrating Automatic Error Detection

Sun Xiaohui, Ling Zhenhua, Dai Lirong

(National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China)

Abstract: A unit selection speech synthesis method is presented using an automatic error detection. It aims to design a unit selection criterion consistent with the subjective perception of listeners so as to improve the naturalness of synthetic speech. Firstly, crowdsourcing platform, instead of linguistics experts in the traditional approach is facilitated to collect mass perceptual data efficiently. Then, a synthetic error detector based on a support vector machine (SVM) classifier is constructed based on speech features such as syllable duration, unit cost and acoustic parameters distance extracted from subjective evaluations. During speech synthesis, N-best unit selection results given by conventional unit selection algorithms are rescored by the trained synthetic error detector in order to select the optimal one. Preference test results show that the proposed method can effectively improve the naturalness of synthetic speech.

Key words: speech synthesis; unit selection; SVM; crowdsourcing; synthetic error detection

引　　言

基于大语料库的单元挑选与波形拼接技术[1-2]是目前最为主流的语音合成方法之一。它的基本原理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

基金项目:国家自然科学基金(61273032)资助项目。
收稿日期:2014-12-18;**修订日期**:2015-01-30

终的合成语音。基于代价函数的单元挑选是其技术核心,代价函数往往通过计算备选单元与目标单元的上下文属性差异和备选单元与预测目标的声学距离来实现。基于隐马尔可夫模型(Hidden Markov model, HMM)的单元挑选[3]是一种有效的实现代价函数的方式。语音合成最终目的是依据输入文本合成尽可能自然的语音,而语音自然与否主要通过人的主观测听来评估,如倾向性测听、平均意见得分(Mean opinion score, MOS)等。然而,在传统的单元挑选语音合成方法中,用于指导最优备选单元序列挑选的准则往往基于声学距离、上下文属性差异等客观度量进行设计,缺乏对于人的主观感知特性的直接考虑。近些年,将人对合成语音的主观评价融入到单元挑选语音合成系统中的一些方法相继被提出,这些方法主要分为两类,即:基于主观评价数据设计代价函数的方法和基于主观评价数据重打分的方法。

基于主观评价数据设计代价函数的方法包括重新设计单元挑选代价函数和优化现有代价函数两种形式。文献[4]从主观评价数据中自动学习一个分类器,并将该分类器的输出映射为 3 个预先设置的离散值之一,用作目标代价代替传统方法中的连续目标代价值。文献[5]首先依据经验选择部分模型权值组合进行单元挑选语音合成,并对合成结果进行人工测听;然后对合成语音的客观度量和主观 MOS 分之间的关系建模,构造一个合成语音 MOS 分预测器;最后利用该 MOS 分预测器在权值空间自动搜索最优权值。文献[6]对合成的自然语音和不自然语音分别构建声学模型,在合成阶段,将根据这两个声学模型计算出的对数似然比整合到目标代价中,指导最优备选序列的挑选。

基于主观评价数据重打分的方法首先根据人对合成语音的评价构建一个合成错误检测器,然后利用该检测器对传统单元挑选方法输出的代价最小的 N 条路径(简称 N 最优路径)进行重打分,以确定最优的单元挑选序列。文献[7]收集主观评价数据时,主要考虑日文中的音调错误,然后利用上下文特征和基频特征构建合成错误检测器。文献[8]以韵律词为单位收集主观评价数据,将合成语音划分为自然空间和不自然空间,分别构建声学模型,然后利用这两个模型计算的后验概率作为输入特征构建合成错误检测器。

本文采用基于主观评价数据重打分的方法改善单元挑选语音合成系统,提出了融合自动检错的单元挑选语音合成方法。该方法相对于文献[7~8]的改进主要体现在两个方面:(1) 基于众包网络平台进行主观评价数据的收集,降低了数据收集的难度。现有方法的一个共同问题在于需要具备一定语言学知识的专家对合成语音以音节或韵律词为单位进行主观评估,以标定每个音节或韵律词的合成结果是否存在错误,这个过程十分耗时,并且代价较高。本文在收集主观评价数据时,降低了对测听者语言学专业能力的要求。合成语句中只有目标韵律词为合成结果,其余为自然语音,这样测听者测听时就会关注目标韵律词,整句话的 MOS 分就可以作为目标韵律词的评分,此测听任务可以由一般母语发音人较好地完成。本文利用 AMT(Amazon mchanical Turk)[9]平台,通过众包的模式,使得收集工作可以快速大量地进行。(2) 改善了现有方法中合成检错器的普适性不足。在现有方法中,文献[7]仅仅关注日文中的音调错误;文献[8]对于含有不同音节数目的韵律词需要构建不同的检错器,否则会出现特征维数不匹配的问题。本文考虑多种可能造成合成语音不自然的原因,综合使用声学特征、声学参数距离特征和单元代价特征训练合成错误检测器;而且特征的提取以韵律词为单位,不存在检错器只能处理包含特定音节数目的韵律词的问题。通过将输入语句看作韵律词的集合,可以处理任意输入文本。本文还探讨了在重打分时使用的 N 最优路径数目大小对系统性能的影响。

1 基于 HMM 的单元挑选语音合成

基于 HMM 的单元挑选语音合成方法[3]分为训练与合成两个阶段。

1.1 训练阶段

首先依据先验知识,从训练数据库的语音波形中提取 M 种可以反应单元挑选与波形拼接系统自然

度的特征参数,例如基频、频谱和音素的时长等。然后针对每一种特征,训练其上下文相关的音素的HMM模型[10],采用基于决策树的模型聚类方法解决训练过程中数据稀疏的问题。最终训练得到的模型集合为 $\Lambda=\{\Lambda_1,\cdots,\Lambda_M\}$。

1.2 合成阶段

首先对于输入的待合成语句文本进行文本分析得到其上下文描述信息 C,然后基于目标函数采用动态规划搜索算法得到最优音素单元序列

$$\boldsymbol{U}^{*}=\arg\max_{\boldsymbol{U}}\sum_{m=1}^{M}w_m\left[\log P_{\Lambda_m}(\boldsymbol{X}(\boldsymbol{U},m)\mid C)-w_{\mathrm{KLD}}D_{\Lambda_m}(C(\boldsymbol{U}),C)\right] \tag{1}$$

式中:M为统计模型的数目;$\boldsymbol{U}=\{u_1,u_2,\cdots u_N\}$为备选单元序列;$\boldsymbol{X}(\boldsymbol{U},m)$为备选单元序列$\boldsymbol{U}$的第$m$种声学特征;$P_{\Lambda_m}(X\mid C)$为在给定模型$\Lambda_m$和上下文信息$C$的情况下观测特征$\boldsymbol{X}$的输出概率;$C(\boldsymbol{U})$为备选单元序列$\boldsymbol{U}$对应的上下文信息;$D_{\Lambda_m}(C(\boldsymbol{U}),C)$为$C(\boldsymbol{U})$和$C$分别对应的第$m$种声学特征HMM模型之间的KLD(Kullback-Leibler divergence)距离,ω_m和ω_{KLD}分别代表声学模型的权值和KLD的权值。

本文使用的声学特征数目为$M=6$,分别为基频、频谱、时长、拼接基频[11]、拼接频谱和长时基频[12]。训练得到的6个统计模型均用于计算输出概率,而在计算KLD距离时,仅使用基频、频谱和时长3个统计模型。出于降低计算量的考虑,本文首先通过KLD对备选单元进行快速预选,减少进入动态规划搜索的单元数目。而拼接基频、拼接频谱和长时基频模型的KLD无法针对单个音素直接计算从而无法用于单元预选,出于算法简便的考虑,不再计算其KLD距离。在完成了最优备选样本的挑选后,通过对前后两帧搜索波形最大的相关位置并进行加窗叠加来实现波形的拼接,合成最终语音。

2 融合自动检错的单元挑选合成

本文提出的融合自动检错的单元挑选语音合成系统(简称检错系统)流程如图1所示。主要分为3步:(1) 主观评价数据的收集;(2) 合成错误检测器的构建;(3) 检错系统构建。

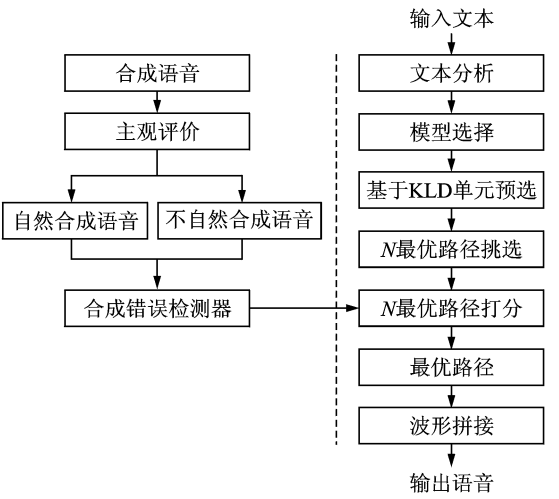


图1 融合自动检错的单元挑选合成系统流程图

Fig. 1 Flowchart of unit selection speech synthesis using an automatic error detector

2.1 主观评价数据

众包模式(即把工作任务以自由自愿的形式外包给非特定的大众网络的做法)近年来成为人工智能领域获取主观数据的一种常用方法[13-16]。众包的优势在于可以快速完成大量的任务,但是缺点也很明显,众包平台上的用户一般不是领域内的专家,如果任务过于困难,完成质量无法保证,本文中通过降低众包任务的难度弥补了这一缺陷。文献[15]表明,采用众包平台收集合成语音主观评价数据可行。

本文采用众包的方式,对合成语音中每个韵律词是否自然进行主观评价数据的收集。假设用于生成待测听数据的开发集有R句话$\{S_1,S_2,\cdots,S_R\}$,每句话包含的韵律词个数分别为$\{T_1,T_2,\cdots,T_R\}$,并且该开发集有对应的语音数据。为了降低测听的专业性要求,使得不具备语言学专业知识的母语发音人都可以进行韵律词合成质量的判断,采用部分合成的方法进行测听语音的生成。在这种部分合成方法中,为了评估开发集中第r句话的第t个目标韵律词的合成质量,在合成第r句话时,只对该韵律词$w_{r,t}$包含的音素采用单元挑选方式进行合成,而对于该句话中其他音素采用开发集中S_r自身的波形进行合成。也就是说在这种方法下合成的语句中,只有

第 t 个韵律词可能存在合成错误，而其他韵律词都是自然录音。这使得本文可以通过收集测听人对于这句话的 MOS 分作为对目标韵律词 $w_{r,t}$ 合成质量的评价。由于合成单个韵律词与合成整句话采用的方法相同，对合成单个韵律词的研究可以自然地推广到整句话的合成方法中。

按上述方法一共合成了 $\sum_{i=1}^{R} T_i$ 句话，从中选取一部分，采用 AMT 众包平台进行主观评价数据的收集。AMT 平台通过互联网发布任务，并且多个任务可以并行发布，一般只需 2～3 天即可收集到结果。该平台还允许自定义众包任务交互界面，并且提供了一系列控制参与者的方法，如限制参与者国籍，限制参与者已完成任务数量等，确保其可以较好地完成任务。本文中限制参与者为美国国籍，并且在交互界面通过听写方式确定参与者具有一定的英文水平。每个语句保证有 10 个测听者的 MOS 得分，取平均值作为其最终得分，以保证得分的可靠性。此外，这种简单的测听任务花费也较低，平均收集 1 个样本的主观评价得分(需要 10 个测听者的得分)花费约 \$ 0.2。

2.2 合成错误检测器

合成错误检测器的构建流程如图 1 左半部分所示。收集到主观评价数据后，依据设定的门限，本文将合成的韵律词划分为自然和不自然两类，并采用基于支持向量机的方法构造合成语音发音错误检测器。本文采用的特征以韵律词为单位提取，包括单元代价、声学参数距离、长时基频和音节时长。具体的特征定义如下：

(1)使用的单元代价共 9 维，即式(1)中 6 个声学模型分别对应的模型输出概率与其中 3 个声学模型分别对应的 KLD 部分。

(2)使用的声学参数距离共 11 维，包括：(a) 频谱距离：合成语音与参考语音每帧 Mel 倒谱的平均距离，采用动态时间规整算法做帧间对齐；(b) 基频距离：合成语音与参考语音每帧基频参数的平均距离，计算方法同频谱距离；(c) 能量距离：合成语音与参考语音每帧能量参数的平均距离，计算方法同频谱距离；(d) 时长差异：合成语音与参考语音音素时长差异的平均值和极大值；(e) 拼接处频谱跳变：合成语音与参考语音在音素拼接处前后两帧倒谱参数的跳变差异的平均值和极大值；(f) 拼接处基频跳变：合成语音与参考语音在音素拼接处前后两帧基频参数的跳变差异的平均值和极大值；(g) 拼接处能量跳变：合成语音与参考语音在音素拼接处前后两帧能量参数的跳变差异的平均值和极大值。

(3)使用的长时基频有 13 维。先按照文献[8]采用的方法，对韵律词单元内的各个音节分别提取 13 维基频相关特征。然后对每一维取各个音节对应维的极大值作为韵律词单元的特征。

(4)使用的音节时长共 2 维：韵律词所包含音节的时长的均值和极大值。

提取的各组特征之间有一定的冗余性，全部叠加在一起并不能取得最优的效果。因此，本文采用如下特征选择算法从上述 4 组特征中选择出最有效的特征组合，其特征算法如下：

(1) 初始化 S 为一个空集。

(2) 令 P 为所有特征的集合。

(3) 令 ScoreBest=0。

(4) for $i=1:k$ (k 是特征总数目)

for 特征 j 不属于 S

令特征集合 $S_j=\{S,j\}$，然后利用 S_j 中的特征训练 SVM 分类器，将该 SVM 分类器的性能指标(例如正确率)作为特征集合 S_j 的得分 ScoreS_j，得分均为正值，且越高越好

end for

$j^* = \arg\max_j \{\text{ScoreS}_j\}$

if ($\text{ScoreS}_{j^*} \leqslant$ ScoreBest)

终止循环

else

$S=\{S,j^*\}$, ScoreBest=ScoreS_{j^*}

end if

end for。

(5) S 即为挑选出的特征集合。

2.3 检测系统

本文提出的检错系统分为训练和合成两个阶段。训练阶段与基于 HMM 的单元挑选语音合成系统[3]相同,合成阶段的流程如图 1 右半部分所示。不同于传统单元挑选方法采用的动态规划算法挑选代价最小的一条路径合成语音,本文采用的动态规划算法在每一个节点保留到当前节点路径中代价最小的 N 个结果,这样在搜索终点处就可以得到备选序列中代价最小的 N 条路径。然后利用合成错误检测器对每一条路径内部各个韵律词分类,输出值是将其判定为不自然类别的概率的估计值,可以作为该韵律词的得分,整条路径的得分为其内部各个韵律词得分之和。整条路径的分数越高,就代表其合成的语音越有可能不自然。因此选取得分最低的一条路径作为最终挑选结果。

3 实验验证

3.1 实验条件

实验采用 6 576 句英文男声音库,约 9 h 的语音数据(16 kHz 采样率,16 b 量化),同时还有对应的音段与韵律标注信息。随机选取其中 3 945 句作为训练集,1 973 句作为开发集,657 句作为测试集。采用基于 HMM 的单元挑选语音合成方法作为基线系统,利用 2.1 节介绍的方法合成语音用于主观评价数据收集。模型训练使用的声学特征数目为 $M=6$,分别为基频、频谱、时长、拼接基频、拼接频谱和长时基频。

3.2 主观评价数据的验证

开发集 1 973 句话一共含有 23 939 个韵律词,考虑到收集代价较高,挑选其中一部分进行收集。为了尽量减少句子长度对测听结果的影响,工作只针对包含 6～13 个韵律词的语句结果,一共有 1 073 句话,包含 9 969 个韵律词。它们所包含的音节数目分布如图 2 所示。

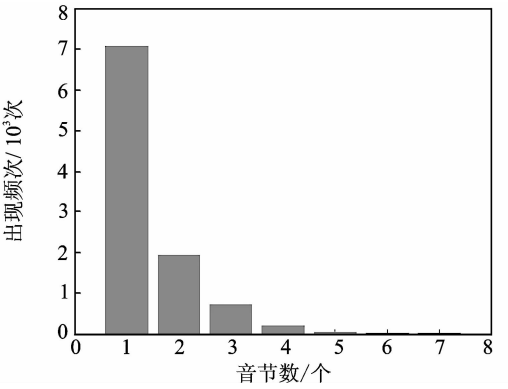


图 2 韵律词包含音节数目直方图

Fig. 2 Histogram of numbers of syllables in a word

本文从包含 1,2 个音节的韵律词中各自随机抽取 1 000 个样本,对于包含 3～5 个音节的韵律词取全部样本,合计有 2 952 个韵律词样本,然后利用 2.1 节介绍的方法进行主观测听数据的收集。

将得到的主观评价 MOS 分根据经验性门限 3.8 分划分为自然空间和不自然空间,收集结果如表 1 所示。可以看出对于包含 1～4 音节的韵律词样本,随着音节数的增多,样本不自然的比例也在增加。说明随着合成音节数目增多,出现合成不自然的可能性也相应增加。对于包含 5 音节的韵律词样本,没有延续上述规律,是因为样本数目较少,个别样本对计算不自然样本比例有较大影响。

表 1 主观评价结果

Tab. 1 Results of subjective evaluation

韵律词音节数目	自然合成语音/个	不自然合成语音/个
1 音节	586	414
2 音节	497	503
3 音节	323	390
4 音节	80	126
5 音节	16	17
合计	1 502	1 450

3.3 合成错误检测器的验证

本文将收集到的主观评价数据 90%作为训练集,10%作为测试集,则训练集有 1 352 个自然的结果和 1 305 个不自然的结果,测试集有 150 个自然的结果和 145 个不自然的结果。实验中 SVM 分类器的训练采用一个公开的工具包 LibSVM[17],采用径向基核函数(Radial basis functional,RBF)对于参数 c 和 γ,利用网格搜索和交叉验证的方法确定。

基于 2.2 节介绍的方法提取出每一个合成韵律词的各组特征,然后以组为单位进行特征挑选。在特征挑选过程中采用不同特征组合得到的分类器在测试集上的性能如表 2 所示。可以看到采用音节时长、单元代价和声学参数距离的特征组合时性能最好,再加上长时基频特征性能反而出现了下降。这是因为在计算韵律词 13 维长时基频特征时挑选了其内部来自不同音节的不同维的数据,然后简单把它们组合在一起,这种处理方式不合理。因此,本文最终选取音节时长,单元代价和声学参数距离共 22 维特征构建合成错误检测器。

表 2 合成错误检测器性能

Tab. 2 Performance of synthetic error detector %

特征	准确率	召回率	F 值
单元代价	60.16	53.10	56.41
单元代价+声学参数距离	60.12	65.51	62.71
单元代价+声学参数距离+音节时长	61.73	68.97	65.15
单元代价+声学参数距离+音节时长+长时基频特征	54.39	68.28	60.55

3.4 检错系统的验证

根据 2.3 节,将以上得到的合成错误检测器融入到单元挑选语音合成系统中,就得到检错系统。考虑重打分时使用的 N 最优路径数目大小对系统性能可能有影响,将 N 值分别设置为 50,100,200,1 000,对应构建了 4 个检错系统,分别称为 P50,P100,P200,P1 000。

本文组织了 4 组倾向性测听,分别对比基线系统(Baseline,BS)和 4 个检错系统的自然度。从测试集中随机挑选 20 句话作为测试语句,分别用上述 5 个系统合成,然后利用 AMT 平台收集了 10 名母语发音人的测听结果,如图 3 所示。可以看出 P50 和 BS 性能相当,一个可能的原因是,当 N 值较小时,不同路径使用的备选单元之间的差异非常小,这时基于合成语音检错器的重打分方法所能发挥的作用有限。P100 相对 BS 性能有显著提高,表明当 N 值适中时,检错系统可以有效提升合成语音的自然度。P200 和 P1 000 相对 BS 的提升有限,这表明当 N 值较大时,忽略原有单元挑选准则而单纯依靠检错器选择最优路径有其局限性。

4 结束语

本文针对传统单元挑选语音合成方法中指导最优备选单元序列挑选的客观度量准则与人的主观听

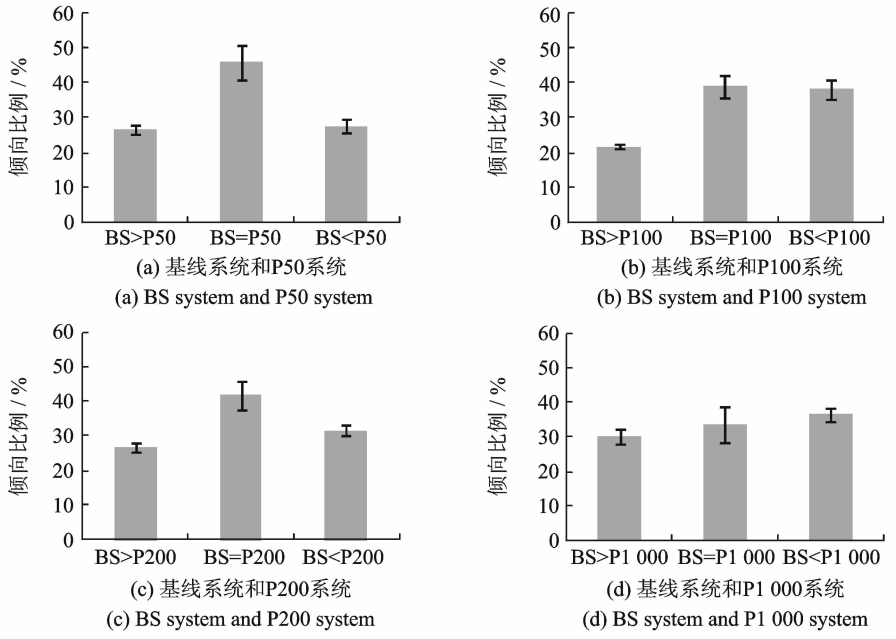


图 3 基线系统和本文提出的 4 个系统的倾向性比例(95%置信区间)

Fig. 3 Preference percentage between BS system and each of the proposed systems with 95% confidence intervals

感不一致的问题,提出了一种融合自动检错单元挑选语音合成方法。该方法首先利用众包网络平台快速大量地收集主观评价数据,然后根据人对合成语音的主观评价训练得到一个合成错误检测器。在合成阶段,该检测器被用来对于传统单元挑选方法输出的 N 最优路径进行重打分,以确定最优的单元挑选序列。实验结果表明,本文提出的方法可以有效提升合成语音的自然度。将本文中所提出的方法与现有的基于主观评价数据的代价函数设计方法[6]进行实验对比,是本文今后的研究任务之一。

参考文献:

[1] Mizutani T, Kagoshima T. Concatenative speech synthesis based on the plural unit selection and fusion method [J]. IEICE Trans, 2005, 88(11): 2565-2572.

[2] Gros J Z, Zganec M. An efficient unit-selection method for concatenative text-to-speech synthesis systems [J]. Journal of Computing and Information Technology, 2008, 16(1): 69-78.

[3] Ling Zhenhua, Wang Renhua. HMM-based unit selection combining Kullback-Leibler divergence with likelihood criterion[C]// Proc of International Conference on Spoken Language Processing. Honolulu, USA: IEEE, 2007: 1245-1248.

[4] Strom V, King S. A classifier-based target cost for unit selection speech synthesis trained on perceptual data [C]// Interspeech. Makuhari, Japan: IEEE, 2010: 150-153.

[5] 宋阳,凌震华,戴礼荣. 基于合成质量预测的单元挑选语音合成优化方法[J]. 清华大学学报:自然科学版, 2013, 53 (6): 762-766.
Song Yang, Ling Zhenhua, Dai Lirong. Optimization method for unit selection speech synthesis based on synthesis quality prediction [J]. Journal of Tsinghua University: Sci & Tech, 2013, 53 (6): 762-766.

[6] Xia Xianjun, Ling Zhenhua, Jiang Yuan, et al. HMM-based unit selection speech synthesis using log likelihood ratios derived from perceptual data [J]. Speech Communication, 2014, 63: 27-37.

[7] Yoshida A, Mizuno H, Mano K. Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis [C] // ICASSP. Las Vegas, USA: IEEE, 2008: 4617-4620.

[8] Lu Heng, Ling Zhenhua, Dai Lirong, et al. Building HMM-based unit selection speech synthesis system using synthetic speech

naturalness evaluation score [C] // ICASSP. Vancouver, Canada: IEEE, 2011: 5352-5355.

[9] Jurcícek F, Keizer S, Gasic M, et al. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk[C]// Proc of Interspeech. Florence: IEEE Press, 2011: 3061-3064.

[10] 戴礼荣，张仕良. 深度语音信号与信息处理：研究进展与展望[J]. 数据采集与处理，2014，29(2)：171-179.
Dai Lirong, Zhang Shiliang. Deep speech signal and information processing: Research progress and prospect [J]. Journal of Data Acquisition and Processing, 2014, 29(2): 171-179.

[11] Ling Zhenhua, Lu Heng, Hu Guoping, et al. The USTC system for Blizzard challenge 2008 [EB/OL]. http://festvox.org/blizzard/bc2008/ustc_Blizzard2008.pdf, 2008-07-24/2015-01-29.

[12] Ling Zhenhua, Wang Zhiguo, Dai Lirong, et al. Statistical modeling of syllable-level F0 features for HMM-based unit selection speech synthesis [C]// Proc of ISCSLP. Tainan, China: IEEE Press, 2010: 144-147.

[13] Alonso O, Rose D E, Stewart B. Crowdsourcing for relevance evaluation[C]// ACM SigIR Forum. Singapore: ACM, 2008, 42(2): 9-15.

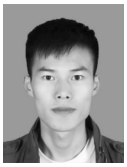
[14] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]// Computer Vision and Pattern Recognition, 2009. Miami, USA: IEEE, 2009: 248-255.

[15] Callison B C, Dredze M. Creating speech and language data with Amazon's Mechanical Turk[C]// Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg, USA: Association for Computational Linguistics, 2010: 1-12.

[16] Wolters M K, Isaac K B, Renals S. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk [C]// Proceedings of the 7th ISCA Speech Synthesis Workshop(SSW7). Kyoto, Japan: The International Speech Communication Association, 2010: 136-141.

[17] Chang C C, Lin C J. LIBSVM : A library for support vector machines [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 1-27.

作者简介：



孙晓辉(1991-)，男，硕士研究生，研究方向：语音合成，E-mail：sunxh06@mail.ustc.edu.cn。

凌震华(1979-)，男，副教授，研究方向：语音合成和说话人转换。

戴礼荣(1962-)，男，教授，研究方向：数字信号处理和人机语音通信。