基于声韵母基元声学特征的中文 TTS 系统中音库的量化压缩策略

张皖志，陶建华

中科院自动化所模式识别国家重点实验室，北京 100080

Email: wzzhang@nlpr.ia.ac.cn　jhtao@nlpr.ia.ac.cn

摘要

本文引入声韵母作为中文语音合成系统的基本单元，从根本上提升了音库的可压缩性。并提出了一种基于声韵母基元声学特征的音库量化压缩方法。整个量化压缩过程分为粗选和聚类两个阶段，分别基于声韵母基元的不同的声学特征。粗选阶段中通过筛选准则自动删除音库中声学特征异常的单元，从而大幅提升系统的稳定性。聚类阶段中从音段特征及超音段特征两个角度进行分析，并结合高层音韵学信息，对分类后的声母和韵母分别进行聚类压缩。主观听辩实验及统计分析结果表明，采用小音库的系统合成结果的可懂度和自然度均接近于桌面系统。

关键词：声韵母基元；量化压缩；ISODATA 聚类；协同发音

1．引言

近年来基于大规模语料库的合成方法渐渐成为语音合成（TTS）领域的主流技术[1][2][3]，其基本思想是从大量自然语流中依照一定的规则选择语音单元进行拼接，从而得到高自然度的合成语音。为保证合成结果具备丰富的韵律表现，其音库规模往往达到数百兆字节，无法应用到存储空间受限的小型嵌入式设备上，从而严重地制约了语音合成技术的发展空间。为实现向嵌入式平台下的移植，须解决的关键问题是：如何在保持合成结果的自然度及可懂度的前提下，尽量降低音库中的声学冗余度，从而实现高效率的压缩。

Sanghun Kim[4]等人提取单元的韵律及频谱参数组成特征向量，采用加权矢量量化的方法对音库容量进行压缩。双志伟[5]等人基于韵律特征及语音学特征来定义单元间的距离测度，用于对音库聚类压缩。孙金城[6]等人以基频作为特征采用 k 均值聚类算法裁减音库。上述方法均采用音节作为拼接合成的基本单元。相对于声韵母来说，汉语中的音节在声学意义上更为独立，相邻音节间的协同发音效应较弱，因此在进行基元拼接时音质的损失较小[7]。但基于音节的音库压缩能力有限，因为汉语中音节的个数远大于声韵母，当音库容量下降到一定程度时，每个音节保留下来的样本数将显著减少，从而导致合成语音的自然度及音质显著下降。

为此，本文引入声韵母作为合成系统的基本单元，并在此基础之上提出了一套完整的音库量化压缩策略。整个量化压缩过程分为粗选和聚类两个阶段，分别基于来自声韵母基元的不同的声学特征。粗选阶段中通过事先设定的筛选准则，自动删除音库中声学特征异常的单元，从而大幅提升系统的稳定性。聚类阶段中首先采用分类与回归树（CART）结合高层音韵学信息对音库中的声韵母单元进行预分类，并提出了一种改进的 ISODATA 聚类方法，分别以 Mel 频标倒谱参数（MFCC）和基频包络作为特征，对 CART 叶子结点声母和韵母进行聚类压缩。主观听辩实验及统计分析结果表明，在音库容量大幅压缩的情况下，该系统合成结果的可懂度和自然度接近于桌面系统，可较好地应用到嵌入式平台上。

文章内容安排如下：第二部分介绍声韵母音库的创建过程；第三部分给出完整的音库量化压缩策略；第四部分简单介绍整个系统的结构框架；第五部分从主观听测实验及客观统计分析两个角度对系统的性能进行评价；最后一部分总结文章并介绍未来的工作。

2. 声韵母音库的构建

声韵母组合结构是汉语特有的语音现象，采用声韵母作为合成基元可有效减少音库中样本的数目（汉语普通话中有 21 个声母和 43 个韵母，而有调音节约有 1300 个），从而从根本上提升音库的可压缩性。

本文采用的声韵母音库在已有的基于音节的大音库的基础上构建。利用语音识别工具包 HTK 对原始音库进行声韵母一级的切分，然后再对边界进行手工校正。

汉语普通话中音节内部的声韵母协同发音现象非常严重，有必要对声/韵母依与其共处同一音节内部的韵/声母类型进一步细分。此处借用语音学中对声韵母的分类

方法[8]，声母分为四类：后接开口呼，后接齐齿呼，后接合口呼，后接撮口呼；韵母分为九类：前接不送气塞音，前接送气塞音，前接不送气塞擦音，前接送气塞擦音，前接不发音擦音，前接发音擦音，前接鼻音，前接边音，零声母韵母。分类后的环境相关的声韵母共同组成音库的基本单元。

3. 音库量化压缩策略

3.1 粗选阶段

受录音人、录音设备及音库标注等人为音素的影响，音库中存在大量从声学特征上来看较为反常的样本。当音库规模较大时，这些音被选出的概率较小，对合成结果影响较小。可当对音库进行压缩后，残留下来的畸变的样本则很容易被选出用于合成语音，从而大幅降低合成结果的稳定性，同时还将占用宝贵的存储空间。本文采用下述三种筛选准则，自动对音库进行预筛选，剔除掉音库中的不稳定因素。

准则一：韵律异常度准则

此处考虑的韵律因素包括样本的音长、基频曲线和能量。定义第 i 个样本的韵律异常度（Prosodic Salience）为：

$$PS(i)=\frac{\omega_1 D_d(i)+\omega_2 D_p(i)+\omega_3 D_e(i)}{\omega_1+\omega_2+\omega_3} \tag{1}$$

其中各子异常度为：

$$D_d(i)=(\frac{d(i)-\bar{d}}{\bar{d}})^2 \tag{2}$$

$$D_p(i)=(\frac{p(i)-\bar{p}}{\bar{p}})^2 \tag{3}$$

$$D_e(i)=(\frac{e(i)-\bar{e}}{\bar{e}})^2 \tag{4}$$

$d(i)$、$p(i)$ 和 $e(i)$ 分别为第 i 个样本的音长、基频均值和平均能量，$\bar{d}$、$\bar{p}$ 和 $\bar{e}$ 分别为该基元所有样本相应特征的均值。各子异常度的权值 ω_1，ω_2 和 ω_3 根据实验得出。对任一样本 i，对 $\forall x, x\in\{d,p,e\}$，若有

$$D_x(i)>T_x \tag{5}$$

或

$$PS(i)>T \tag{6}$$

则删除该样本。其中 T_x 和 T 分别为各子韵律异常度和总韵律异常度的阈值。该准则可剔除音长或峰值点标注出错的样本，以及录音过程中人为因素导致的能量过弱或过强的样本。

准则二：粘连度准则

该准则考察音库中的样本在原始语流中与相邻单元协同发音的程度。对基于小音库的系统来说，拼接处由谱不连续导致的音质损失尤为严重，在建库阶段尽量剔除粘连度较强的音是一种可行的方案。定义第 i 个样本的粘连度（Context Dependency）为：

$$CD(i)=\frac{\omega_l \bar{e}_l(i)+\omega_r \bar{e}_r(i)}{\omega_l+\omega_r} \tag{7}$$

其中 $\bar{e}_l(i)$ 和 $\bar{e}_r(i)$ 分别为样本左、右边界处的平均能量，可根据单元的声学特征决定其权值。例如，对塞音和塞擦音可令 ω_l 为 0。类似的，若样本 i 的 $CD(i)$ 大于某个阈值 T，则剔除该样本。

准则三：音质异常度准则

录音人在长期录音的过程中由于疲劳或其它心理因素可能导致录制的某些样本音质出现异常，表现为气声、耳语或掺杂明显的情感。这些音往往出现在句子结尾处，能量偏弱，且元音的周期性较差。对样本 i, 定义其音质异常度（Quality Distortion）为：

$$QD(i)=\frac{n_{peak}(i)}{\bar{e}(i)\cdot dur(i)} \tag{8}$$

其中 $n_{peak}(i)$ 为该样本峰值点的数目，$\bar{e}(i)$ 为平均能量，$dur(i)$ 为该样本的音长。若样本 i 的 $CD(i)$ 大于某个阈值 T，则剔除该样本。

实验表明，采用上述三准则对原始音库进行压缩，可压缩掉 10%的样本，且合成结果可懂度及自然度没有明显损失。

3.2 聚类阶段

语音合成系统的性能可分别采用自然度与可懂度进行评价，其中自然度基于超音段特征，可懂度基于音段特征[9]。下面就在粗选后的音库基础之上，从上述两个角度进一步压缩。

首先对声韵母进行预分类，使得每类内部的单元具备相似的声学特征。然后分别对预分类后的声韵母基于各自的声学特征进一步聚类压缩。从而保持原始音库的音段特征多样性及超音段特征多样性。

- 基于音韵学环境属性的预分类

在连续语流中，可认为在相似的音韵学环境属性下，样本的声学特征也较为相似，即音韵学环境属性和声学特征之间存在某种映射，即利用环境属性对声学单元进行分类。鉴于环境属性均为离散特征，所以采用分类与回归树（CART）方法。

在 CART 中，如何设计决策属性是一个关键问题。考虑到声韵母之间的强协同发音效应，只有选取与语音学

现象密切相关的决策属性才能起到较好的分类效果。本文选取的决策属性建立在对上下文的描述之上，包括：

- 与当前声/韵母同音节的韵/声母类型及 ID。
- 前音节韵母类型及 ID
- 后音节声母类型及 ID
- 声韵母所在音节的调形，前音节调 形，后音节调形（包括阴平，阳平，上声，去声，轻声五种）。
- 低层次韵律层次相对高层次韵律层次 的相对位置，韵律层次包括韵律词、韵律短语、语句。相对位置包括在层次的首、中、尾。
- 声韵母所属音节的韵律词长度，韵律短语长度，以音节个数为单位。
- 声韵母所属音节的前后静音段的长度。

选用 12 阶 Mel 频标倒谱参数（MFCC）作为声韵母单元的特征参数，选用 mahalanobis 距离来计算单元间的距离。单元 M，N 的距离定义如 Eq.(9)。

$$dis(M,N)=\sum_{i=1}^{|M|}\sum_{j=1}^{12}\left[P_{ij}(M)-P_{(i\cdot\frac{|M|}{|N|})j}(N)\right]^2 \qquad (9)$$

其中 $P_{ij}(M)$ 为第 i 帧的第 j 个 MFCC 参数，$|M|$ 为 M 的帧数。实际计算时，把音节内部声韵母间过渡段的 MFCC 也包含在声韵母的参数向量之内，目的是更好的对声韵母间的协同发音进行建模，使得分类结果对与其相邻的声韵母更加敏感。

利用 CART 训练工具 wagon 为每个声韵母生成一颗 CART 树，叶子结点上样本数目控制在 50-100 之间。统计结果表明，分类结果的集外测试正确率在 85%-95%之间。CART 不仅有效的把样本依声学特征的相似度分开，同时还提供了一个由音韵学信息到样本聚类的映射关系，可用于声学模块的基元选取。

- 基于超音段特征对韵母进行压缩

韵母的听感上的差异主要来自于超音段特征，具体表现为基频曲线的变化。韵母决策树叶子结点上的样本基频曲线存在较大冗余性，须对其进行聚类，从而保持超音段特征的多样性。

K 均值聚类算法实现简单，但需要手动指定类别数，通常并不能反映真实的情况。ISODATA 算法[10]可通过调整样本所属类别完成聚类分析，并能自动地进行类的“合并”和“分裂”，从而得到类数较为合理的各个聚类，但其需要手工指定的参数多达 7 个，当对待分类样本缺乏先验知识时，很难平衡参数之间的关系，往往导致聚类结果不收敛。

为解决这一问题，设计了一种改进的 ISODATA 方法。设类别数为 C，一个聚类中的最小类内平均距离为 θ_S，一个聚类中的最少样本数为 θ_N，两类的合并参数为 θ_C。算法如下：

1．令 C=1。

2．基于某种距离测度，用 K 均值方法把全部样本分成 C 类，得到 C 个类中心点 c_n，每类的样本数为 n_n $(n=1,2,\mathrm{L}\ C)$。

3．计算每一类的平均类内距离 d_n，若 $d_n>\theta_S$ 且 $n_n>2\theta_N$，则进行分裂；分裂方法如下：对 $\forall i,j\quad 1\le i,j\le n_n$ 且 $i\ne j$，假设以样本 m_{in} 和 m_{jn} 作为分裂后新的中心，将该类所有其他样本按到这两个中心的距离分为两类，每类个数为 n_{in} 和 n_{jn}，每类内部平均距离为 d_{in} 和 d_{jn}，以使

$$\overline{d}=\frac{n_{in}d_{in}+n_{jn}d_{jn}}{n_{in}+n_{jn}} \qquad (10)$$

最小的 m_{in} 和 m_{jn} 作为新的类中心。记分裂的类数为 p，则全部分裂完成后 $C=C+p$。若 p 等于 0，转到 4，否则转到 2。

4．对 $\forall i,j\quad 1\le i,j\le C$ 且 $i\ne j$，设将第 i 类和第 j 类合并后其类内平均距离为 $\overline{d}$，若 $\overline{d}<\theta_C$，则将这两类合并。记合并类数为 q，则 $C=C-q$。

与传统 ISODATA 算法不同，该方法先尽可能的细分样本，待分裂全部结束之后再进行合并，从而可保证收敛，可通过调节 θ_S 和 θ_C 来控制类别数。

实验中以韵母的基频曲线作为特征参数，采用欧氏距离作为聚类测度。设样本间的平均距离为 $\overline{dis}$，令

$$\theta_s=\overline{dis}\cdot r_{final} \qquad (11)$$

且

$$\theta_s<T_s \qquad (12)$$

其中 r_{final} 为期望中的韵母的压缩比，T_s 为分类精度控制参数。

- 基于音段特征对声母进行压缩

声母的多样性来自于音段特征。实验中以 MFCC 作为特征参数，仍采用 Eq.(9)作为聚类测度，采用上述聚类算法。同样的可通过调节声母压缩比 r_{ini} 来控制声母压缩的程度。

3.3 音库压缩结果

原始音库在 16K 采样率下所占空间为 479MB。当 $T_s=10Hz$ 时，调节声韵母的压缩比 r_{ini} 、r_{final} 生成不同的音库并进行比较，结果如表一所示。

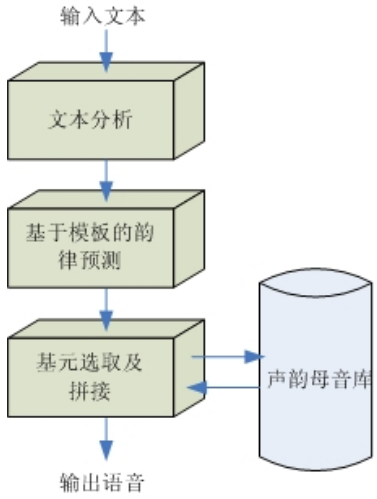
音库 / 参数	1	2	3	4
r_{ini}	0.7	1.0	0.7	1.0
r_{final}	0.3	0.3	0.5	0.5
所占空间 (MB)	210	188	205	184
MOS	3.90	3.87	3.90	3.85

表一 $T_s=10Hz$ 时，音库间性能比较

增大声母压缩比，音库将得到大幅压缩且性能无明显下降，说明声母的可替换性较强；增大韵母压缩比，音库容量下降不明显，说明在指定的压缩精度下，压缩能力已趋于饱和。将 T_s 上调至 15Hz， 令 $r_{ini}=1.0$ ， $r_{final}=0.3$ ，生成的音库作为最终的音库。

4．系统结构

基于上述小音库，构建了以声韵母为基本单元的 TTS 系统，结构框图如图一所示。



图一 嵌入式 TTS 系统框图

系统中韵律预测模块采用基于模板的预测方法[11]，并针对声韵母单元作了优化。声学模块分为选音和拼接两部分，选音时首先基于 3.2 中生成的决策树进行预选，选出与目标基元声学特征较为相似的候选基元，然后再基于目标损失和拼接损失利用 Viterbi 算法进一步搜索。目标损失考虑了候选单元与目标单元间基频曲线及音长的差距，拼接损失考虑了候选基元间拼接处频谱的不连续性。

声韵母间协同发音效应的强弱与单元类型相关，对音库分析后发现，在塞音和塞擦音前协同发音较弱。于是引入发音组块模型，即以最近的两个塞音或塞擦音内部的声韵母序列作为声学平滑单元，在组块内部进行基元的选取和拼接。以边界处谱的 KL 距离来表征谱不连续性[12]，作为衡量协同发音强弱的客观标准。基于压缩后的音库合成了 100 句文本并统计平均 KL 距离。表二表明，采用该模型后声谱得到了平滑。拼接时采用 PSOLA 算法对韵律进行调整。

	平均 KL 距离
使用前	1.35
使用后	1.21

表二 采用发音组块前后的谱不连续性

5．性能评价

记原始的基于音节的音库为 *LibA* ，转成声韵母后的音库为 *LibB* ，基于前述压缩策略，将 *LibB* 压缩至 98MB，然后在 16K 采样率下用 GSM 编码，生成大小为 7MB 左右的音库，记为 *LibC* 。

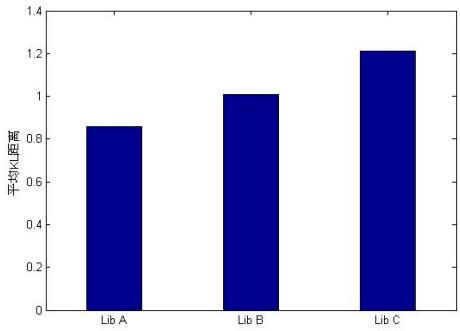
基于上述三个音库各合成一百句文本。统计句子内部基元拼接处的平均 KL 距离作为客观评价标准，实验结果如图二所示。同时邀请五位专业人士参与主观评测实验（采用 5 分制，1 为最差，5 为最好），实验结果如图三所示。

实验结果表明，以声韵母作为系统合成的基本单元与音节相比，须付出更大的拼接代价，但从听感上并没有明显的区别。当压缩比达到 4：1 时，基于 *LibC* 的系统从主客观评价两个方面来看性能都有所下降，但仍与原系统较为接近。

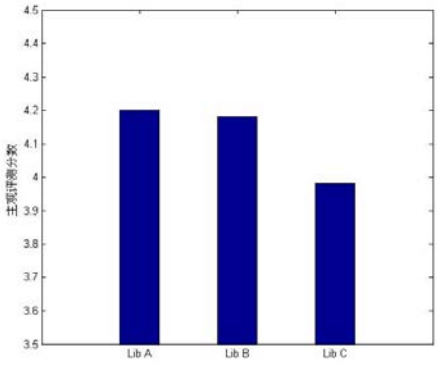
6．结论及展望

本文提出了一种的基于声韵母的声学特征对中文 TTS 系统中的音库进行量化压缩的方法。采用声韵母作为基元可大幅提升系统的可压缩性，而且在同等音库规模下与基于音节的合成系统的性能几乎没有差别。主客观评价实验均表明，压缩后系统的性能与桌面系统较为接近。

汉语中的声韵母基元集可与英语中的音素集在某种程度上实现共享，今后将在本文对汉语声韵母合成研究的基础之上，进一步探讨两种语言在音库及声学模型上的有机融合问题，最终实现中英文混合的双语合成系统。



图二 不同音库性能的客观评价



图三 不同音库性能的主观评价

参考文献

[1]A.Hunt and A. Black., “Unit selection in a concatenative speech synthesis system using a large speech database”, In ICASSP-96, volume 1, pages 373-376, Atlanta, Georgia, 1996.

[2]Renhua Wang, Zhongke Ma, “A corpus-based Chinese Speech Synthesis with Contextual-Dependent Unit Selection”, ICSLP2000

[3]Min Chu, Hu Peng, Hong-yun Yang,Eric Chang, “Selecting non-uniform units from a very large corpus for concatenative speech synthesizer”, Proceedings of ICASSP2001

[4]Sanghun Kim, Youngjik Lee, and Keikichi Hirose, “Pruning of Reduntant Synthesis Instances Based on Weighted Vector Quantization”, Proceedings of Eurospeech 2001, pp.2231-2234,2001

[5]Zhiwei Shuang, etc, “A miniature Chinese TTS system based on tailored corpus”, ICSLP2002

[6]孙金城，易立夫，分层语音合成数据库设计与分析，全国声学学术会议，2002，377～378

[7]Chilin Shih, Richard Sproat, “Issues in Text-to-Speech Conversion for Mandarin”, Computational Linguistics & Chinese Lang. Processing, vol. 1, 1996

[8]罗常培 王均，普通语音学纲要（修订本），商务印书馆，2002

[9]Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, “Selection of waveform units for corpus-based mandarin speech synthesis based on decision trees and prosodic modification costs”, in Proc. Eurospeech,1999

[10]边肇祺 张学工，模式识别（第二版），清华大学出版社，2000

[11]Tao Jianhua, “Trainable prosodic model for standard Chinese Text-to-Speech system”, Chinese Journal of Acoustic, Vol.20, 2001, P257-265

[12]E.Klabbers and R.Veldhuis, “Reducing audible spectral discontinuities”, IEEE Trans. Speech and Audio Processing, vol.9, no.1, pp.39-51, 2001