第28卷 第4期 计 算 机 工 程 2002年4月
Vol.28 №4 Computer Engineering April 2002
· 多媒体技术及应用 · 文章编号：1000—3428(2002)04—0171—02 文献标识码：A 中图分类号： TP 391.42

语音库辅助建立工具的开发

张文军，谢剑英，李 聪
（上海交通大学自动化系，上海200030）

摘 要：叙述了连续数字语音库的建立及相关辅助工具的开发。

关键词：语音库；标注；切分

Development of a Tool for Assisting Speech Corpora Production

ZHANG Wenjun，XIE Jianying，LI Cong
(Department of Automation ,Shanghai Jiaotong University,Shanghai 200030)

【Abstract】 This paper describes the process of producting continuous digital corpora and the development of a tool for assisting speech corpora production.

【Key words】 Corpora; Label; Segment

目前语音识别系统的识别率和语音合成的自然度还不能令人满意，其根本原因是对自然语音的研究不够深入，不能准确归纳、描述和模拟自然语音的规律。因此，语音库建设已成为语音处理研究的基础。

1 语音库的建立过程

为了进行语音处理的研究，需要建立一个语音库，用来训练语音的声学模型，测试并比较不同算法的性能。一般来说，语音库的建立过程包括语音的录制、信号采集、分段和手工标注等，工作量十分繁重，需要开发辅助工具减轻任务强度。

连接数字语音库的话音录自19个男性说话人，其中每人朗读40个7位的连续数字串和20个孤立的数字。在安静的环境下，通过麦克风录入磁带中。录制的话音通过音频处理软件采集，存入计算机文件内。其中文件的编码格式为 Wave 波形文件，量化精度为16位，采样率为16kHz。随后，通过开发的语音辅助分段工具Segmenter，将该语音波形文件划分为连续数字串或孤立数字构成的语音段。针对每个独立语音段的波形文件，开发了语音辅助切分和标注工具Tag，用来对连续语音进行切分和标注。辅助切分工具Tag首先用简单的方法进行粗切分，然后手工对切分点进行调整。经过以上步骤，就可以建立已标注完整的语音库，用来进行声学模型的训练和测试。接着利用辅助混频工具Mixer，与标准噪音库合成，产生不同信噪比的语音，进行系统和算法的鲁棒性测试。

2 语音库建设的辅助工具

为了利用计算机辅助建设语音库，首先开发了一系列的辅助工具，主要包括语音分段工具Segmenter、语音切分和标注工具Tag和语音合成工具Mixer。所有工具利用Visual Basic的相关控件实现。其中Segmenter利用了录音时不同的语音段间明显的间隔手工分段，实现简单，在此不再详述。

辅助切分工具Tag包括粗切分、手工调整和标注3部分，具体实现如图1。其中粗切分方法是在可能的切分点中寻找一个能够将语音段平均划分的切分组合，方法如下：

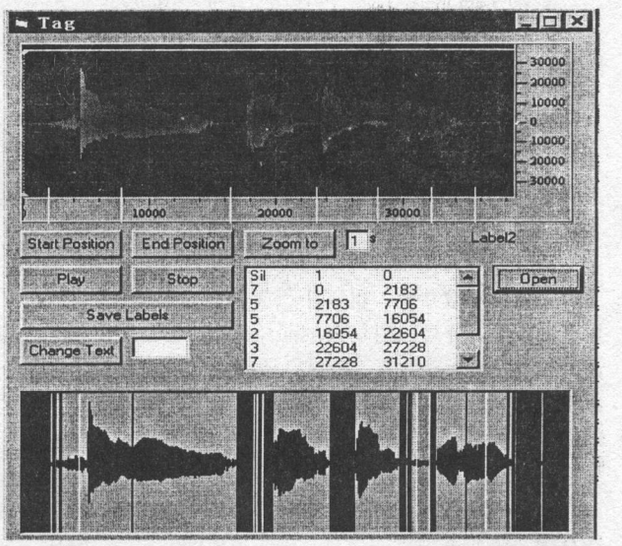

图1 辅助语音切分工具Tag

1) 首先将语音波形信号划分成一系列互不重叠的帧，计算每一帧的能量E_t，并计算相邻两帧之间能量的差

$$\Delta E_t = E_t - E_{t-1}$$

2) 将相邻两帧间的能量差进行平均值滤波，即

$$\Delta E'_t = \frac{\sum_{j=-2}^{2} \Delta E_{t+j}}{5}$$

3) 如果$\Delta E'_t < 0$且$\Delta E'_{t+1} > 0$，这说明在第t帧处语音信号变弱，而下一帧中信号增强，于是第t帧的结束位置是一个信号强度的谷点，记作候选切分点。根据这个标准找到所有的候选切分点构成集合C。在集合C中寻找一组N - 1个切分点，使得每一个音节的长度尽可能地接近。

作者简介：张文军(1967～)，男，博士生，主要从事智能控制，数字信号处理和语音信号处理等方面的研究；谢剑英，教授、博导；李 聪，本科毕业生

收稿日期：2001-06-27

4)在连续语音中可能存在着静音，故音节的长度中一般要去除静音的部分。定义任意两个候选切分点t_i和t_j之间的语音长度为

$$d(t_i,t_j)=\sum_{\substack{t_i\le t\le t_j\\ \text{s.t.}\Delta E_t>\theta}}1$$

其中，q为能量阈值，能量小于q的帧被认作是静音帧。切分的依据是不同音节语音长度的方差最小，即

$$S^*=\arg\min_S\sum_{i=1}^{N}\left(d(t_{i-1},t_i)-\frac{1}{N}d(t_0,t_N)\right)^2$$

粗切分算法

1) 初始化

$$f_1(t)=(d(t_0,t)-\frac{1}{N}d(t_0,t_N))^2 \qquad 1\le t\le T$$

$$g_1(t)=0 \qquad 1\le t\le T$$

2) 迭代

$$f_k(t)=\min_{t_{k-1}}[f_{k-1}(t_{k-1})+(d(t_{k-1},t)-\frac{1}{N}d(t_0,t_N))^2]$$

$$1\le t\le T \quad 1\le k\le N$$

$$g_k(t)=\arg\min_{t_{k-1}}[f_{k-1}(t_{k-1})+(d(t_{k-1},t)-\frac{1}{N}d(t_0,t_N))^2]$$

$$1\le t\le T \quad 1\le k\le N$$

3) 终止

$$f^*=f_N(T) \qquad t_N^*=T$$

4) 回溯

$$t_k^*=g_{k+1}(t_{k+1}^*), \qquad k=N-1,N-2,\cdots,0$$

语音辅助切分工具可以将当前的切分结果显示于波形图上，并具有回放功能，可以任意选择某一音节的当前切分结果进行回放，并可以对切分的起始点和终止点进行微调，微调的分辨率在5ms以内。经过自动粗切分和人工调整之后，最终生成的标注文件包括音节标号、音节的起始位置、音节的结束位置。

3 语音与噪声的合成

在采集过程所获得的语音信号是不含噪声的，鲁棒语音切分实验需要含噪语音信号，因此须按照一定的信噪比，利用辅助混频工具Mixer将语音信号与噪声信号进行合成，生成可以用于鲁棒语音实验的语音数据。

普通的信噪比定义为信号与噪声的方差(对于平稳信号，即为能量)之比。设语音信号为x(n)，噪声信号为r(n)，则普通的信噪比为

$$SNR=20\lg\frac{\sigma_x^2}{\sigma_r^2}(\text{dB})=20\lg\frac{\sum_n x^2(n)}{\sum_n r^2(n)}(\text{dB})$$

在不断的探索以期望得到主观上有意义的噪声度量过程中，人们提出了几种改进型的信噪比度量，以应用于语音、图像等信号。一个理想的度量是某个单一的数值，它应该通用、有意义、可靠、容易求得以便于判断和分析。由于语音信号是非平稳信号，对于相同量的噪声来说，随着环境信号电平的不同，对感觉的影响也会有所不同。同时，对于输入信号来说，不同频带中的噪声对信号的破坏也不同。

当信噪比在一定范围时，信号中的噪声对人类感觉的影响才会有意义，根据这一情况，定义一个感觉上限SNR_{max}和下界SNR_{min}是有意义的。一般来说，较为合理的值应取得不致于过分地影响最后结果。

在语音库的含噪语音合成过程中，综合考虑前面提及的几种因素，采用的信噪比度量是可称为分段频域加权信噪比，其值为

$$SNRSeg_f=E[\max\{\min\{SNR_f(m),SNR_{min}\},SNR_{max}\}]$$

其中，SNR(m)表示第m段的频域加权信噪比，而取期望即是在信号序列中对所有感兴趣的分段在时间上取平均，SNR_{max}一般取为35dB，SNR_{min}取为0dB。

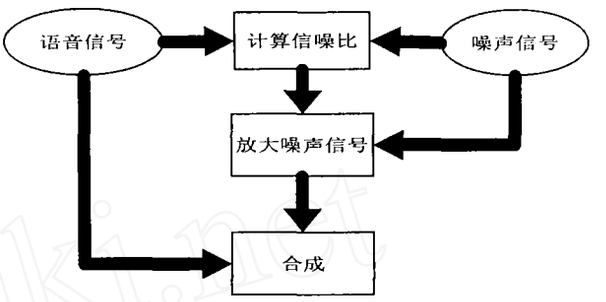

图2 语音与噪声的合成过程

在合成过程中，噪声信号取自Noisex-92噪声库。语音与噪声的合成过程如图2所示。首先，计算当前的语音信号和噪声信号的分段频域加权信噪比，通过与期望获得的带噪声语音信噪比进行比较，获得噪声信号所需的放大倍数。在放大噪声信号后，同原有的语音信号合成，得到了最终的含噪声的语音。

4 未来的工作

应改善用户界面的设计，使其更加方便易用；在Mixer工具中，考虑通道的影响和回声问题，使含噪语音满足不同的实验要求。未来可能的发展，诸如自动的一致性检验、语音信号与文本的自动对齐、视频播放、变速回放等，可能使用户得到更多的便利。而且，在新的应用领域中，要求提高语音文件管理方式和标注格式的柔性等。

参考文献

1 赵世霞，蔡莲红，常晓磊. 汉语语音合成语料库管理系统的建立. 小型微型计算机系统，2000,21(3)

2 陈肖霞. 连续话语语料库的语音切分和标记. 语言文字应用，2000,34(2)

3 Barras C.Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. Speech Communication,2001,33:(1-2):5-22