TTS语音单元边界的自动切分

王丽娟 曹志刚

(清华大学电子工程系 微波与数字通信技术国家重点实验室, 北京 100084)

摘 要: 语音单元边界的准确切分对基于波形拼接的语音合成系统至关重要。文章采用了两步切分方法,第一步中先由基于HMM模型的强制对齐方法得到初始的边界,在第二步中提出用基于前后音素的边界模型来修正初始边界。为解决训练数据不足的问题,提出用分类与衰退树将前后因素发音相近的边界模型进行聚类。这样可以根据训练数据的多少,动态调节边界模型的数目,以保证模型训练的可靠性。在对中文语音库的实验中,自动切分的准确度由78.7%提高到91.5%。

关键词: 前后音素相关,边界模型,分类与衰退树,自动切分,TTS

中图法分类号: TN912.3 文献标识码: A 文章编号: 1000-7180(2005)12-008-04

Automatic Segmentation for TTS Units

WANG Li-juan, CAO Zhi-gang

(1 State Key Laboratory of Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084)

Abstract: Correct unit segmentation are, though laborsome, very crucial to the performance of a concatenation based TTS system. This paper suggests a two-step procedure for automatic unit segmentation, which coarsely segments speech data in the first step and refines segment boundaries in the secord step. A new Context-Dependent Boundary Model (CDBM) to describe the evolution across the segment boundary is proposed. To reduce manual segmentation, Classification and Regression Tree (CART) is used to structure the available data into a more efficient usage. Acoustically similar boundaries are clustered together and corresponding tied CDBM models are thus trained and used for boundary refinement during the secord step. After a series of experiments, the optimal CDBM parameters and the training conditions are found. The segmentation accuracy is raised from 78.7% to 91.5% in Mandarin syllable segmentation with about 1,000 manually segmented sentences as CDBM training data.

Key words: Context-dependent boundary model, CART, Automatic segmentation, TTS

1 引言

基于波形拼接的语音合成技术,日益成为文语转换系统中的主流技术。它采用自然语言波形直接拼接的方法,进行拼接的语音单元都是从一个预先录制好的自然语言数据库中挑选出来的。只要语音库足够的大, 包括了各种可能语境下的语音单元,有可能拼接出任何语句,并且有可能最大限度地保留原有语音的自然度。语音单元的切分和标注的准确性直接影响到合成语音的质量。在目前大部分的TTS系统中, 语音单元的切分和标注都由人工来完成,并且需要进行反复校对来保证一致性。当语音数据库较大时,进行切分,标注,校对需要大量的时间和成本。因此,语音单元的自动切分成为目前的研究热点。

收稿日期: 2005-03-21

2 自动切分两步法

常用的自动切分方法有两种:(1) 基于模板;(2) 基于模型, 以隐马尔可夫模型HMM(Hidden Markov Model)为代表。它们都是运用动态规划方法将一串语音单元的模板(方法一)或是模型(方法二)与给定的一句语音进行对齐,从而得到每个语音单元的起始时间。或用自动语音识别ASR(Automatic Speech Recognition)中的术语,称为强制对齐(Forced Alignment)。文献[1]研究表明,基于HMM模型的强制对齐比基于模板的方法能得到更好的切分准确度。因而,我们采用了基于HMM模型的强制对齐方法来产生初始的语音单元切分边界。然而,强制对齐产生的边界点往往不是用于拼接合成的最佳拼接点。为了减小自动切分与人工标注之间的差距,人们用两步法来进行自动切分。如图1所

示,在第一步中,基于 HMM 模型的强制对齐产生初始边界;在第二步中,通过学习一部分人工标注的边界,训练得到的边界模型对初始边界进行修正。

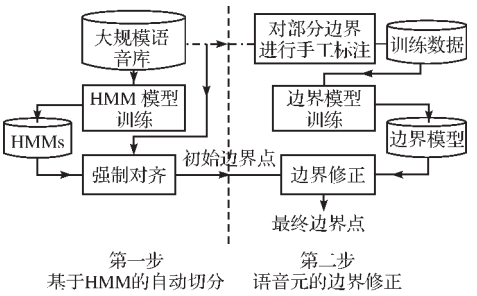

图1 自动切分两步法的整体框图

在以往的研究中，人们提出了多种边界模型，如神经网络模型 NN(Neural Network)[1]、混合高斯模型 GMM(Gaussian Mixture Model)[2]、隐马尔可夫模型、多层感知模型 MLP(Multi-Level Perception)[3,4]等。这些模型在不同程度上提高了语音单元的切分准确度。在本文中,我们提出了基于前后音素的边界模型 CDBM(Context-Dependent Boundary Model)来对初始边界进行修正。由于人工标注的边界数量有限,为解决训练数据不足的问题,我们提出了用分类与衰退树 CART (Classification and Regression Tree)将前后因素发音相近的边界模型进行聚类。这样,可以根据训练数据的多少,动态调节边界模型的数目,以保证模型训练的可靠性。

3　基于前后音素的边界模型 CDBM

3.1　边界建模与模型聚类

在音素边界处的语音变化是由这个边界前后两个音素所共同决定的,因此,我们提出了基于前后音素的边界模型 CDBM。每一个 CDBM 都可以表示为 $X-B+Y$，其中 B 代表边界,X 代表边界前面的音素,Y 代表边界后面的音素。如果 X 有 N_X 类,Y 有 N_Y 类,(N_X 和 N_Y 不一定相同)，那么就可能 N_X*N_Y 个这样的 CDBM。为表征边界的特征,我们提出将边界周围的语音帧特征向量联合成一个特征向量来描述边界信息，并用一个 GMM 来对一种边界进行训练。如图 2 所示,首先在边界的左右各取 N 帧语音信号,包括边界本身语音帧$(t_{-N},\dots,t_0,\dots,t_N)$,共 $2N+1$ 帧 m 维的语音信号特征向量。然后将这 $2N+1$ 个语音特征向量联合在一起就构成了一个 $(2N+1)*m$ 维的边界特征向量。元素为:

$$b_j(y_t)=p(y_t|j) \tag{1}$$

表示在状态 j,对于观测语音特征 y_t 的似然值。

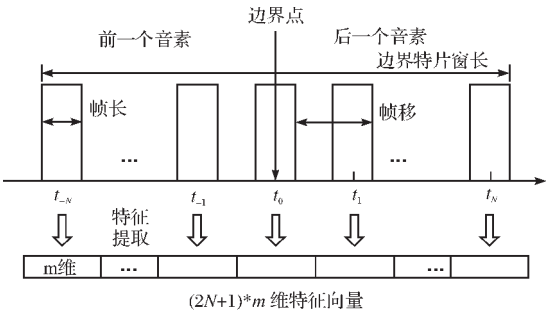

图2 边界特征向量的提取

对经过矢量量化(VQ)的语音特征来说,B 为产生各码本的离散概率。复杂度比较高的语音识别系统通常采用连续概率密度隐含马尔可夫模型 CDHMM (Continuous Density HMM),即语音特征取值是连续的，其统计规律用概率密度函数 PDF(Probability Density Function)来表示,通常用高斯分布的线性组合来拟合 PDF:

$$b_j(y_t)=\sum_{m=1}^{M} c_{jm}N(y_t;\mu_{jm},\sum_{jm}) \tag{2}$$

其中 $N(y_t;\mu_{jm},\sum_{jm})$是多维高斯分布，$\mu_{jm}$ 和 $\sum_{jm}$ 分别为其均值向量和协方差矩阵,c_{jm} 表示第 m 个高斯分量在状态 j 中的权重，$\sum_{m=1}^{M} c_{jm}=1$,M 为状态中高斯分量的总数。

为了能够准确地描述每一种边界的特性,我们希望能为每一个 CDBM 建立一个 GMM 模型。然而通常用于训练的手工切分边界数目有限,因此无法为每个 CDBM 训练出一个可靠的 GMM 模型。为了解决训练数据不足的问题,我们采用分类与衰退树(CART)[6] 来将前后音素发音相近的 CDBM 进行聚类,如图 3 所示。这样即使在训练中没有出现过的 CDBM 模型也能够通过 CART 找到与它最接近的模型来代替。其中模型的训练与聚类的实现方法与自动语音识别中对三音子 HMM 模型训练聚类相似[5]。

3.2　边界修正

一旦边界模型 CDBM 训练完毕,就开始对第一步得到的语音单元初始边界进行修正。这里我们所采用的方法与文献[2]类似。如图 4 所示,对每一个待修正的初始边界,假设准确的边界在此初始边界周围的搜索区间内,我们将在此区间内搜索出最合适的边界点作为修正后的结果。首先根据这个给定边界的前后音素,即,它所对应的 $X-B+Y$,在 CART 上找到相应的边界模型。然后,对这个搜索区域内每一个候选边界点,提取其边界特征信息,根据模型算出它的似然值。将其中似然值最大的候选边界

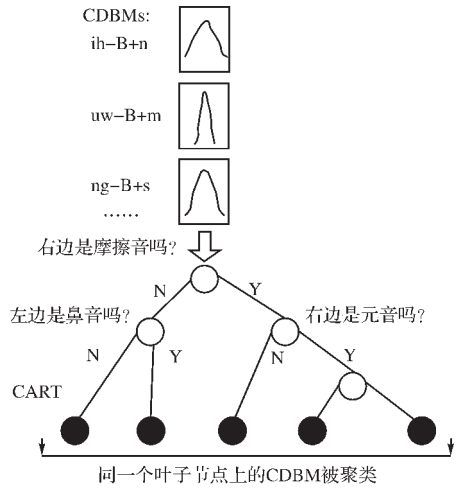

图 3 用CART对边界模型进行聚类

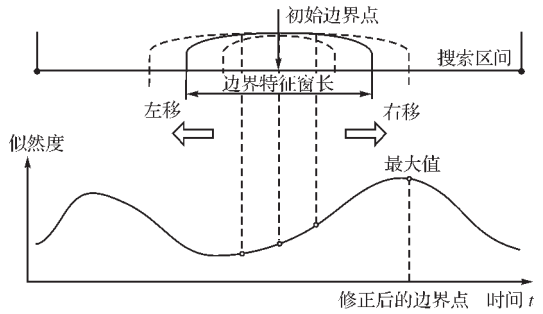

图 4 边界修正示意图

点,作为修正后的边界结果。

4 试验及结果

4.1 自动切分性能评价方法

由于在汉语的拼接语音合成中,通常选用汉字的音节作为拼接的语音单元,因而在我们衡量切分准确度的时候只考虑所有音节的边界,忽略音节内部的音素之间的边界。以人工标注的边界为正确标准,自动切分的边界点与正确边界间的偏差在±T 之内时,我们认为是正确的,否则,是错误的,T 称为容错门限。正确自动切分边界的数目占所有切分边界的百分比称为在此容错门限下的自动切分准确度。在语音合成中, 通常考察的是容错门限为 20ms 的自动切分准确度。

实验采用微软中文 TTS 语音库,其中包含经人工校对过的所有音节标注和音节的边界切分信息。数据库共包含一万两千个句子,将近 18 万字。所有句子都是由一个受过良好播音训练的女性发音人用放松的朗读语调读出的。该数据库在设计时,充分考虑了连续语流中音节在各种韵律层次、声调环境和音联环境中的可能变化,使各基本音节都有足够多的韵律变体。在第一步中,我们用用 HTK 工具对整个语音库[5]训练得到不带声调的三音子 HMM 模型。然后 HMM 模型对整个语音库进行强制对齐的切分,得到初始的边界点。

第二步中,部分手工切分的边界(1000~50,000 个)将用于训练边界模型 CDBM,另外的 10,000 个边界将用于测试自动切分的性能。在提取边界特征时, 我们为每个边界取 5 帧帧长为 20ms 的语音信号,帧移为 30ms,如图 2 所示。对于每一帧语音信号,提取 39 维语音特征向量,它包括 12 MFCC,语音帧能量的自然对数,以及它们的一阶、二阶差分。这样,产生了 5*39 维的边界特征向量。

为了测试第一步中的 HMM 模型的性能, 我们用训练集外的 500 句话做了自动语音识别的试验,测试结果为音节的误识率为 7.3%,这表明 HMM 模型已经训练得足够好了。然而,在容错门限为 20ms 时,强制对齐的切分准确度仅为 73.6%。我们把强制对齐的切分结果作为基线系统,在下面的所有实验中,经过边界模型修正后的切分结果将与这个基线结果进行比较。

4.2 实验一 不同混合高斯模型对切分准确度的影响

这个实验的设计目的是测试不同混合高斯模型对边界切分准确度的影响, 即确定边界模型 CDBM 的混和高斯成分数目。在下面的实验中,用于训练边界模型的是手工切分的 20,000 个边界。边界模型通过 CART 分类成 154 个叶子节点, 其中 CART 停止分裂的判定准则是保证每个叶子节点上至少有 10 个训练数据。在为每个叶子节点训练的边界模型中,我们分别选用了 1 到 8 个高斯成分的 GMM 模型。这些模型对应的边界切分的实验结果分别在表 1 被列出。实验表明,随着边界模型 GMM 中高斯成分数目的增加,边界切分的准确度在逐渐下降。

表 1 混合高斯成分的数目与切分准确度

容错门限 (ms)	基线系统	混合高斯成分的数目			
		1	2	4	8
10	43.1	69.9	69.4	68.0	65.9
20	78.7	91.5	91.1	90.3	89.2
30	92.6	96.9	96.7	96.5	96.1

这说明单一高斯成份的 GMM 模型已能很好地来描述一个 CDBM。或者同一类中的边界特征在统计分布上非常一致,只需用单高斯描述就足够。另一方面,当增加模型中的混合高斯成分时,虽然对

模型的描述能力增强了,但同时增加了需要训练的模型参数。这样,当训练数据较少时,不能使得这些模型参数得到可靠估计, 从而使得多高斯的 GMM 模型的切分性能反倒下降。

4.3 实验二 边界模型数目对于切分准确度的影响

这实验的设计目的是考察经 CART 聚类后边界模型最合适的数目。基于实验 一的结果,边界模型中高斯成份的个数被设定成 1。通过调整 CART 的每个叶子节点上最少的训练数据量 MTT(Minimum Training Tokens),可以控制 CART 的叶子节点数。当 MTT 减少时,CART 分裂的会较深,这样叶子节点的数目会增加, 相应的边界模型数目也会增加。反过来,当 MTT 增大时,边界模型的数目就减少。从表 2 中可以看到, 当边界模型的训练数目是 20,000 时,只是在 MTT 大于 40 时,边界切分的准确度开始下降,而对于 MTT 小于 40 的所有的值,切分的准确度基本不变。在训练集的数目被减少到 5,000 时, 切分准确度一直随着 MTT 的减小而增大直到它到达 10 为止。 这个结果说明在保证每个叶节点上的训练数据足够的前提下,模型数目越多越好。因此,边界模型的细分类比以往人们采用的简单分类能得到更好地切分性能。

表 2 边界模型数目与边界切分准确度(%)

训练数据量	基线系统	MTT					
		2	5	10	20	40	80
5,000	78.7	81.7	89.8	89.9	89.8	88.8	86.3
20,000	78.7	91.4	91.4	91.5	91.2	91.1	90.3

图 5 显示的是当训练数据量为 20,000 时,采用两步法修正后的边界在不同容错门限上的切分准确度的提升。

5 结束语

本文采用了两步切分方法,先由强制对齐方法得到初始的边界,然后用基于前后音素的边界模型修正初始边界。并通过实验讨论了模型中高斯成分的数目和模型聚类的程度对切分准确率的影响。实验结果表明这种方法能大幅提高边界切分的准确性。在对中文语音库的实验中,自动切分的准确度由 78.7%提高到 91.5%。将该方法应用于中文 TTS 平台,得到了高质量的合成语音。

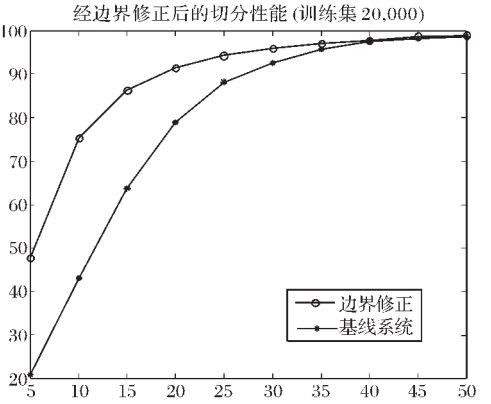

图 5 切分准确度在各容错门限上的性能提高

参考文献

[1] Doroteo Torre Toledano, Luis A Hernández Gómez. Automatic Phonetic Segmentation [J]. IEEE Transactions on speech and audio processing, November 2003,11(6): 617~625.

[2] Abhinav Sethy, Shrikanth Narayanam. Refined Speech Segmentation for Concatenative Speech Synthesis[C]. Proceeding of ICSLP, Denver, Colorado, USA, September 2002: 145~148.

[3] KI－Seung Lee, Jeong Su Kim. Context－adaptive Phone Boundary Refining for a TTS Database [C]. Proceeding of ICASSP, Hongkong, China, April 2003: 252~255.

[4] Eun-Young Park, Sang-Hun Kim, Jae-Ho Chung. Automatic Speech Synthesis Unit Generation with MLP based Postprocessor Against Auto－segmented Phoneme Errors [C]. Proceeding of ICASSP, Phoenix, Arizona, March 1999: 2985~2990.

[5] Odell J, Ollason D, Woodland P, et al. The HTK Book for HTK V3.0 [M]. Cambridge University Press, Cambridge, UK, 2001.

[6] Xuedong Huang. Alex Acero, Hsiao-Wuen. Spoken Language Processing [M]. Prentice Hall PTR, Upper Saddle River, New Jersey, 2001.

[7] Lijuan Wang, Yong Zhao, Min Chu, et al. Refining Segmental Boundaries for TTS Database Using Fine Contextual-dependent Boundary Models[C]. Proceeding of ICASSP, 2004: 641~644.

王丽娟　女,(1979–),**博士**。研究方向为语音信号处理、语音合成、语音识别。

曹志刚　男,(1940–),**教授,博士生导师**。研究方向为通信理论、卫星通信、宽带移动通信、远程教育、语音信号处理。