Voice Conversion Based on STRAIGHT and UBM-GMM

GAO Yingying, ZHU Weibin

Institute of Information Science , Beijing Jiaotong University, Beijing, China

e-mail: wbzhu@bjtu.edu.cn

Abstract: A novel approach for voice conversion is proposed which uses STRAIGHT analysis/synthesis arithmetic to analyze and synthesize speech. UBM-GMM is adopted and the speaker characters are described by the parameters of GMM. Consequently the amount of the target speaker training dataset is reduced. The transformation function based on GMM is also modified and the parallel corpus is successfully avoided then, along with an improvement in the flexibility of the function to the character parameters.

Keywords: voice conversion; STRAIGHT; UBM-GMM

基于 STRAIGHT 和 UBM-GMM 的语音转换实现

高莹莹，朱维彬

北京交通大学，信息科学研究所，北京，中国，100044

e-mail: 05282034@bjtu.edu.cn, wbzhu@bjtu.edu.cn

【摘　要】本文提出一种语音转换实现方案，采用 STRAIGHT 分析-合成算法进行语音信号的分析与重构，在此基础上引入 UBM-GMM 模型，利用高斯混合模型参数来表征发音人特征，使目标说话人训练集规模得到缩减。还对基于 GMM 的转换方程进行了改进，避免了平行语料的使用，同时改善了方程对转换参数的适应性。

【关键词】语音转换; STRAIGHT; UBM-GMM

1 引言

语音是人类主要的交流方式之一，人们通过发音风格的调整来表达不同的语义内容和情感；不同人发出相同语义内容和情感的语音时也常有明显的差别，即具有明显的个性特征。语音转换就是保持语音的语义内容不变，改变说话人声音的个性特征，使之听起来像由另一个人说出来的，或者像是该说话人在另一种发音风格下的声音。

现有的发音人语音转换系统实现通常包括三个过程：提取代表说话人个性信息的声学特征；建立两说话人声学特征的映射规则；以及将转换后的声学特征合成为语音信号。其中建立两说话人声学特征的映射规则是系统的核心部分。目前比较流行的映射方法主要有码本映射法[1]、动态频率规整[2]、人工神经网络[3]和混合高斯模型[4]等，这些方法大多需要使用平行语料才能得到源和目标说话人声学特征的映射规则，这就很大程度上限制了语音转换在实际生活中的应用。UBM-GMM 模型在发音人识别中得到了成功应用[5]，它通过对训练好的背景模型(UBM，Universal Background Model)进行自适应估计来获得单个说话人的模型(GMM)，该模型很好地表征了说话人的个性特征。我们将这种方法引入语音转换，从而避免平行语料的使用，并缩减语音训练集的规模。

文章第二部分介绍了实现方案；第三部分是该方案的测听实验；最后一部分对全文工作做了总结和展望。

2 方案介绍

STRAIGHT[6]算法因为其能够较好的分离声源信息和声道信息，并且合成语音的品质很有保证，因此在我们的语音转换方案中使用该算法对语音信号进行分析与合成。

在声学特征的建模部分，我们使用目前较为成功的高斯混合模型(GMM)进行声道参数的建模，由于说话人的 GMM 将高度依赖于训练语音的数据量，但是实际应用中需要目标说话人的训练数据集要尽可能的小，这种情况下如果独立训练源和目标说话人的 GMM 就不能保证目标说话人 GMM 的精准性。因此我们引入 UBM-GMM 模型，即只需要源说话人的训练集满足一定的规模，将源说话人声道参数的 GMM

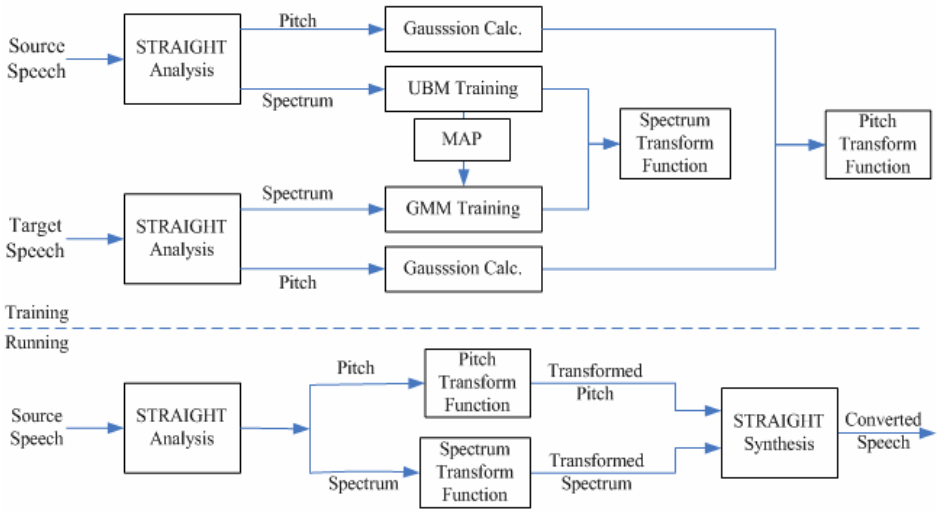


Figure 1. Flowchart of voice conversion approach based on STRAIGTH and UBM-GMM

图 1. 基于 STRAIGHT 和 UBM-GMM 的语音转换方案框图

作为全局背景模型（UBM），将目标说话人的声道参数对 UBM 进行最大后验估计（MAP）来得到目标说话人的 GMM。这样做既保证了目标说话人声学特征模型的准确性，又大大减少了目标说话人的训练语料的数量。

声学特征的映射规则即转换函数我们选用基于 GMM 的频谱转换函数和均值-方差基频转换函数，并在此基础上做了改进，避免了平行语料的使用，增强了转换方程对特征参数的适应性。图一是该方案的框图。

2.1 STRAIGHT分析-合成算法

STRAIGHT 分析-合成算法一方面利用一系列的补偿窗来消除时域信号的周期性干扰，另一方面在样条空间中进行逆滤波来削除频域干扰，从而大幅度提高了合成语音的质量;同时,它模仿声码器 VOCODER 的思路，将语音信号分解成声源部分和声道部分，使得能够单独对声道频谱包络进行修改，并且使人们可以对语音的一些基本参数如基频、时长、增益等进行调整而不会导致语音质量有太大的下降。

由于 STRAIGHT 分析算法求得的频谱参数维数非常大，当进行大规模参数训练时，会导致计算复杂度成倍增加，所以我们选用维数较小的线谱对参数作为声道参数进行后面模型的训练。

2.2 UBM-GMM模型训练

首先训练出源说话人的 GMM 作为全局背景模型（UBM），然后将目标说话人的参数对全局背景模型的参数进行自适应，即贝叶斯适应或者最大后验估计（MAP）来得到目标说话人的 GMM[5]。

与 EM 算法相同，自适应估计过程包括两步：第一步与 EM 算法中的求期望步骤相同，对目标说话人训练数据统计值进行估计，用来对 UBM 的每个混合分量进行相应的计算；与 EM 算法的第二步不同，在自适应中，这些新的统计值估计与从背景模型中得到的旧的统计值估计通过一个与数据相关的混合适应系数相结合得到最终的统计值估计。

2.3 转换函数

2.3.1 频谱转换函数

通常情况下基于GMM的频谱转换方法[4]要求源和目标说话人使用平行语料，即要求源和目标说话人的训练数据具有相同的语义内容、时长和录音背景，这样实际应用起来就很不方便，所以本文采用下面的方程进行转换从而避免平行语料的使用。

首先训练出源说话人s和目标说话人t谱参数的 GMM，得到 $\lambda_s=(Q,\alpha^X,\mu^X,\Sigma^X)$ 和 $\lambda_t=(Q,\alpha^Y,\mu^Y,\Sigma^Y)$。

然后将模型参数带入下面的方程求解转换后的谱参数：

$$F(X)=\sum_{i=1}^{Q}p_i(X)[\mu_i^Y+\frac{\Sigma_i^Y}{\Sigma_i^X}(X-\mu_i^X)] \qquad (1)$$

其中，$p_i(X)=\dfrac{\alpha_i N(X;\mu_i;\Sigma_i^X)}{\sum_{i=1}^{Q}\alpha_i N(X;\mu_i;\Sigma_i^X)}$

978-1-935068-06-8 © 2009 SciRes.

然而，这样做仍然没有避免特征参数过平滑和合成语音模糊化的缺点。文献[8]中指出这种现象是由于源说话人与目标说话人的相关性是非线性的，因此协方差之比会很小，导致偏移量部分的值相对均值部分会很小，这样就使最终转换结果接近于目标说话人的均值加权平均，而细节部分则丢失，因此造成合成语音的模糊化。该文献中提到一种改进做法，即假设源与目标说话人的方差相同，则二者之比衡为1，此时式(1)可改为：

$$F(X) = X + \sum_{i=1}^{Q} p_i(X)(\mu_i^Y - \mu_i^X) \qquad (2)$$

这样做可以在一定程度上增强转换函数的适应性，同时也可避免平行语料的使用，所以本文中暂定采用式(2)作为谱参数的转换函数。

2.3.2 基频转换函数

本文采用均值-方差线性转换算法，首先假设源、目标说话人的基频值都服从正态分布，根据源、目标说话人的语音训练数据可以计算出各自的均值和方差：(μ_s, σ_s)，(μ_t, σ_t)。由于人耳对音高的感知是和频率成对数关系的，所以我们首先需要将基频信号变换到对数域再做转换。基频的转换函数为：

$$F(\log(f_{0s})) = \mu_t + \frac{\sigma_t}{\sigma_s}(\log(f_{0s}) - \mu_s) \qquad (3)$$

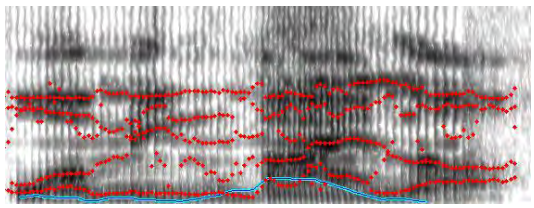
3 转换实验及测听

通过实验，我们有以下结论：高斯模型的混合度越高，分类越精细，转换语音更加连续、清晰，如图2。但同时计算量也越大，所以混合度并非越高越好。综合转换效果和计算量的考虑，本文最终设定混合度为512。同样，训练集的大小对转换结果也有影响，但这也是以增加计算量为代价的，所以训练集不是越大越好，实际应用中也要求目标说话人的数据不能过多。

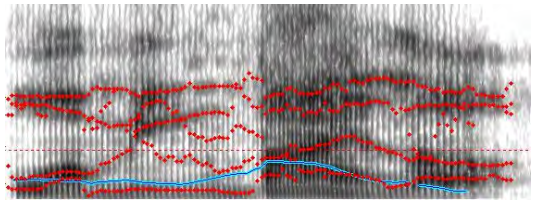
本方案采用1500句源说话人语料（女声,含1000句词和短语，500句长句，共计时长约2.5小时），10句目标说话人语料(男声，10句长句，共计时长约1分钟)，GMM混合度为512。

语音数据由精选的标准普通话发音人录制而成。声音文件采用16KHz采样率、16位数据、单声道WAV格式存储。

采用TC-STAR中的评估方法[7]，选择15名非专业人员进行测听，测听语句共10句，含3个词，2个短语，3个短句和2个长句；从相似度、语音品质两方面进行评测。表1、表2分别为转换语音品质和相似度打分标准、及测听的平均得分。



(a)



(b)

Figure 2. Spectrograms of conversed speeches whose mixtures are 64 and 512

图2. 混合度分别为64, 512(a, b)的转换语音语图

Table 1. Standards for five levels

表1. 五个等级标准

得分	5	4	3	2	1
相似度	一定相同	可能相同	不确定	可能相同	绝对不同
品质	好	较好	中等	较差	差

通过统计，测听结果为“品质”得分为3.93，“相似度”得分为2.13。可见该方案得到的转换语音有很好的品质，但相似度方面却不是很理想，非正式测听还表明转换后的语音在韵律、语调等方面更接近于源说话人。

4 总结与展望

本文实现了一个在保证一定相似度情况下使用非平行语料和小规模目标说话人训练集的语音转换方案，该方案基于STRAIGHT分析-合成算法，能得到品质很好的转换语音，但在相似度方面仍有较大改进余地。

从发音模型的角度，现方案主要是对谱包络进行调整，在一定程度上实现了发音调音功能的转换；对于激励源的转换，仅仅实现了基频的音域变换，而细致的嗓音品质(voice quality)的差异则基本未及；而韵律层面的差异：节奏、语调、语气方面的特点，则完

全没有体现。后续的研究将着重解决提高转换语音相似度，可能的切入点在于引入激励源模型或参数。

致谢

北京交通大学生物医学工程专业05级的15位同学参加了测听工作，在此表示感谢。

References (参考文献)

[1] M. Abe, *et al*. “Voice Conversion through Vector Quantization”. in Proc. ICASSP 1988: 655-658.

[2] H Valbret, *et al*. “Voice transformation using PSOLA technique”. Speech Communication. 1992, 11(2-3): 175-187.

[3] M Narendranath, *et al*. “Transformation of formants for voice conversion using artificial neural networks”. Speech Communication. 1995, 16(2): 207-216.

[4] Y. Stylianou, *et al*. “Continuous Probabilistic Transform for Voice Conversion”. IEEE Transactions on Speech and Audio Processing. 1998, 6(2): 131-142.

[5] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”. Speech Communication. 1995, 17(1-2): 91-108.

[6] H. Kawahara, I. Masuda-Katsuse and A. Cheveign´e, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and aninstantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds”. Speech Communication. 1999, 27(3-4): 187–207.

[7] Zhiwei Shuang, Fanping Meng and Yong Qin, “Voice Conversion by Combining Frequency Warping with Unit Selection”. in Proc. ICASSP 2008: 4661-4664.

[8] Y. Chen and M. Chu. “Voice Conversion with Smoothed GMM and MAP Adaptation”. in Proc. Eurospeech. 2003: 2413-2416.