

基于 ISODATA 聚类算法的语音转换研究

崔立梅, 李燕萍, 吕中良

(南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘要:提出了一种基于迭代自组织聚类算法(ISODATA)的双线性频率弯折语音转换模型。根据语音特征参数分类不充分产生残差成分的问题,在基于高斯混合模型的聚类过程中引入了迭代自组织聚类算法。该算法将聚类得到的类内均值作为训练模型初始均值,改善了EM算法初始值选取不当导致算法不能收敛的问题,从而对特征参数的拟合更加准确,结合后续的双线性频率弯折(BLFW)模型实现语音转换。实验测试结果表明:提出的算法具有较好的自适应聚类特性,能够使特征参数分类更合理,进而得到更准确的转换函数,使得转换的语音更接近目标语音。选择合适的初始值参数,对提出的算法与高斯混合模型及双线性频率弯折模型进行比较,平均MCD值相差很小,平均MOS值有所提高。这说明合理精确的聚类有利于提高语音转换系统的性能。

关键词:迭代自组织聚类算法;双线性频率弯折语音转换模型;残差成分;聚类特性

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2017)06-0106-04

doi:10.3969/j.issn.1673-629X.2017.06.022

Research on Voice Conversion Based on Self Organizing Clustering and Frequency Warping

CUI Li-mei, LI Yan-ping, LYU Zhong-liang

(College of Communication & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: A voice conversion model of bilinear frequency warping based on Iterative Self-Organizing clustering Data Analysis Techniques Algorithm (ISODATA) is put forward. According to the residual components generated by insufficient classification of speech feature parameters, in the clustering process based on Gaussian mixture model, the iterative self-organizing clustering algorithm is introduced. It takes average value within class obtained by clustering as the initial mean for training model, which improves the problem that the algorithm cannot converge due to inappropriated initial value selection of EM algorithm, thus making the characteristic parameters fitting more accurate, realization of voice conversion with subsequent bilinear frequency warping (BLFW) model. The experimental results show that the proposed algorithm has better adaptive clustering characteristics, which can make the characteristic parameters classification more reasonable, and get more accurate conversion function, making the speech more close to the target speech. Choosing appropriate initial value parameters, the algorithm proposed is compared with the Gauss mixture model and the bilinear frequency warping model. The average MCD value is very small, and the average MOS value is high. This shows that reasonable and accurate clustering is beneficial to improve the performance of speech conversion system.

Key words: iterative self-organizing clustering algorithm; bilinear frequency warping voice conversion model; residual components; clustering characteristics

0 引言

语音包含很多信息,其中最主要的就是语义信息,其次是个性化信息。语音转换(Voice Conversion)就是要改变一个说话人(源说话人, source speaker)的语

音个性特征信息,使之具有另外一个人(目标说话人, target speaker)的个性特征信息^[1]。语音转换是一种改变源说话人的声音,使其听起来具有目标说话人特性的技术。它在改变说话人个性特征的同时,保持语

收稿日期:2016-06-08

修回日期:2016-10-11

网络出版时间:2017-04-28

基金项目:国家自然科学基金资助项目(61401227);江苏省博士后基金(1402067B)

作者简介:崔立梅(1988-),女,硕士研究生,研究方向为语音转换;李燕萍,博士,副教授,研究生导师,通讯作者,研究方向为语音转换和说话人识别。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20170428.1702.024.html>

音的语义信息不变。

语音转换的本质是对语音特征参数的转换,因此首先是选取分析和合成语音的系统模型,提取好的语音特征参数。然后训练并得到合适的转换函数,最后进行转换以及语音合成处理。

现在常用的语音特征参数包括 LPC 系数和 MFCC 参数,以及由 LPC 系数推演得到的包括 LSP 参数在内的一系列推演参数等。文中采用 MFCC 参数。语音转换研究的核心问题是寻找能够精确反映源说话人特征参数和目标说话人特征参数之间的映射关系,即转换函数。目前较流行的语音转换函数是基于高斯混合模型 (Gaussian Mixture Model, GMM)。基于 GMM 模型的转换方法具有较好的转换效果,但存在转换频谱过平滑的问题,导致转换后的语音自然度下降,严重影响了该方法的实用性。为了提高语音质量,A. Pribilova 等提出了一种基于频率弯折的转换算法 (DFW)^[2],但转换效果不佳。D. Erro 综合了基于 GMM 模型的转换算法和频率弯折算法的优势,提出了一种在 GMM 模型的基础上进行加权的频率弯折算法 (Weighted Frequency Warping, WFW),较好地平衡了语音质量和转换性能之间的矛盾^[3]。但是由于在转换过程中未对幅度进行转换,转换的相似性一般。于是 D. Erro 提出了高斯混合模型+频率弯折+幅度压扩模型 (GMM+FW+AS)。

为了进一步提高语音转换的质量,D. Erro 提出了残差 (residual) 成分的概念,此处的残差是指没有被特征参数捕获的语音信号谱成分^[4]。其中一些语音信号谱成分未被捕获是由于分类不合理造成的,于是文中提出了迭代自组织聚类算法+高斯混合模型+双线性频率弯折加幅度压扩语音转换模型 (ISODATA + GMM + BLFW),采用 ISODATA 聚类方法^[5]替代 GMM 混合模型传统的 K 均值法进行自适应无监督分类,获得更为合理的聚类,能更好地捕获特征参数的信息;在频率弯折部分采用 BLFW (双线性频率卷绕) + AS,比 FW+AS 更容易实现^[6]。

文中研究在于,一方面利用 ISODATA 聚类方法实现语音特征参数的分类,结合后续的 EM 计算和 BLFW 训练及转换得到 ISODATA + GMM + BLFW 语音转换模型,在此基础上,调整 ISODATA 的初始参数得到最优分类数;另一方面在最优分类数的基础上,比较 ISODATA + GMM + BLFW 模型与 GMM 模型、FW 模型、GMM + BLFW 模型、GMM + BLFW 模型的转换效果。

1 传统的 GMM+BLFW 转换算法

在平行语料库,对目标语音和源语音分别提取特

征参数 MFCC,然后利用动态时间规整 (Dynamic Time Warping, DTW)^[7] 算法进行时间对齐。将对齐的 MFCC 特征参数进行 GMM 模型训练。

$$P(\boldsymbol{X}|\boldsymbol{\lambda})=\sum_{i=1}^MP(\omega_i)N(\boldsymbol{X};\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i)\tag{1}$$

其中, \boldsymbol{X} 为 p 维随机矢量; $P(\omega_i)$ 为混合加权重,且 $\sum_{i=1}^Mp(\omega_i)=1$; $N(\boldsymbol{X};\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i)$ 为每个子分布的 p 维联合高斯概率分布,表示如下:

$$N(\boldsymbol{X};\boldsymbol{\mu}_i;\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}\mathrm{e}^{\{-\frac{1}{2}\langle\boldsymbol{x}-\boldsymbol{\mu}_i\rangle^T\boldsymbol{\Sigma}_i^{-1}\langle\boldsymbol{x}-\boldsymbol{\mu}_i\rangle\}}\tag{2}$$

其中, $\boldsymbol{\mu}_i$ 为均值矢量; $\boldsymbol{\Sigma}_i$ 为协方差矩阵。

完整的混合高斯模型由协方差、参数均值向量和混合权重组合而成,表示为 $\boldsymbol{\lambda}=\{\boldsymbol{w}_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}$ ^[8]。对 GMM 模型参数 $\boldsymbol{\lambda}$ 的估计常常采用 EM (Expectation Maximization) 算法^[9-10]。在采用 EM 算法估计 GMM 模型参数时,必须要先确定 GMM 模型的高斯分量个数 M 和模型的初始参数 $\boldsymbol{\lambda}$ 。传统的 GMM 模型的高斯分量个数 M 一般为 8, 16, 32, 64 等;初始参数 $\boldsymbol{\lambda}$ 采用 K 均值算法将特征参数归为与高斯分量个数相等的各个类中,然后分别计算各个类的均值、方差,作为初始均值和方差;权值是各个类中所包含的特征矢量个数占总特征矢量个数的比率。确定初始参数 $\boldsymbol{\lambda}$ 后,采用 EM 算法估算出一个新的模型参数,使得新的模型参数下的似然度大于初始参数下的似然度。经过多次迭代得到最终 $\boldsymbol{\lambda}$ 。

经过 GMM 模型训练后进行双线性频率弯折训练,其中双线性函数的特征为只需要一个参数 α 确定。

$$z_{\alpha}^{-1}=\frac{z^{-1}-\alpha}{1-\alpha z^{-1}},|\alpha|<1\tag{3}$$

若给定一个因果且时间离散序列 $x[n]$ 及其 Z 变换 $X[z]$,可根据 $Y[z]=X[z_{\alpha}]$ 计算得到一个新序列 $y[n]$,即 $y[n]$ 为 $X[z_{\alpha}]$ 的逆 Z 变换^[11-13]。

$$y[n]=\sum_{m=0}^{\infty}\omega_{nm}x[m]\tag{4}$$

$$\omega_{nm}=\frac{1}{2\pi j}\oint_{z_{\alpha}^{-1}}z_{\alpha}^{-m}z^{n-1}\mathrm{d}z$$

实际上,上述理论也适用于倒谱序列,其中 Z 变换对应于 \log 幅度谱。给定一个 p 维倒谱矢量 \boldsymbol{X} ,倒谱矢量 \boldsymbol{y} 对应于频率弯折函数^[11-13]:

$$\boldsymbol{y}=\boldsymbol{W}_{\alpha}\boldsymbol{X}\tag{5}$$
$$\boldsymbol{W}_{\alpha}=\begin{bmatrix}1-\alpha^2&2\alpha-2\alpha^3&\cdots\\-\alpha+\alpha^3&1-4\alpha^2+3\alpha^4&\cdots\\ \vdots&\vdots&\ddots\end{bmatrix}$$

GMM+BLFW 模型中用到的双线性频率弯折函数为:

$$y = W\alpha(X, \theta)X + s(X, \theta) \quad (6)$$

其中, W 由式(5) 确定; $\alpha(X, \theta)$ 和 $s(X, \theta)$ 分别为频率弯折因子和幅度压扩因子, 由式(7) 确定:

$$\alpha(X, \theta) = \sum_{k=1}^M p_k^{(\theta)}(X) \alpha_k \quad (7)$$

$$s(X, \theta) = \sum_{k=1}^M p_k^{(\theta)}(X) s_k$$

其中, $p_k^{(\theta)}(X)$ 为 GMM 训练得到的语音特征矢量在第 k 个高斯分量中的后置概率, 即:

$$p_k^{(\theta)}(X) = \frac{P(\omega_i)N(X; \mu_i; \Sigma_i)}{\sum_{i=1}^M P(\omega_i)N(X; \mu_i; \Sigma_i)} \quad (8)$$

由式(7) 可知, 频率弯折因子和幅度压扩因子是根据 GMM 模型训练得到的, 此时 GMM 模型训练及 BLFWA 模型训练结束。频率弯折因子和幅度压扩因子确定后, 即可得到转换函数, 通过转换函数对新语音进行转换。

2 改进的 GMM+BLFWA 转换算法及系统框架

2.1 ISODATA+GMM+BLFWA 算法

从上节可以看出, 传统的 GMM+BLFWA 转换算法中每个说话人赋予的模型结构完全相同, 人为确定聚类数然后采用 K 均值法对特征参数进行分类。但是每个说话人语音信号短时频谱的概率分布并不完全相同, 这样就会导致语音特征参数分布聚类拟合不精确, 带来较大误差并影响频率弯折中参数的估计。因此, 文中提出了 ISODATA+GMM+BLFWA 模型, 根据每个说话人具体语音特征分布选择高斯混合数, 建立与之相应的模型结构, 使每个模型结构更好地拟合每个说话人的具体特征分布, 从而提高语音转换准确率。该模型利用 ISODATA 对特征参数矢量序列进行无监督分类, 在样本均值迭代中根据预先设定的阈值进行反复修改, 以达到合理分类数。

2.2 整个转换系统框图

系统转换框图见图 1。

从图中可以看出, 语音转换可以分为两个阶段, 即训练阶段和转换阶段。在训练阶段, 语音信号首先利用 AhoTransf^[14] 语音信号建立模型。该模型可作为语音信号分析/合成模型。提取出参数 MFCC 和 logf0, 其中 MFCC 用于训练频率弯折转换函数。得到 MFCC 参数后进行时间对齐 DTW, 形成两个一一对应的时序, 再利用 ISODATA 算法进行聚类, 得到合理分类的梅尔特征参数。进而将梅尔特征参数进行 GMM 训练, 获得概率函数 P 及均值、方差等一系列参数。利用 GMM 训练得到的概率函数 P 及对齐的 MFCC 特征源

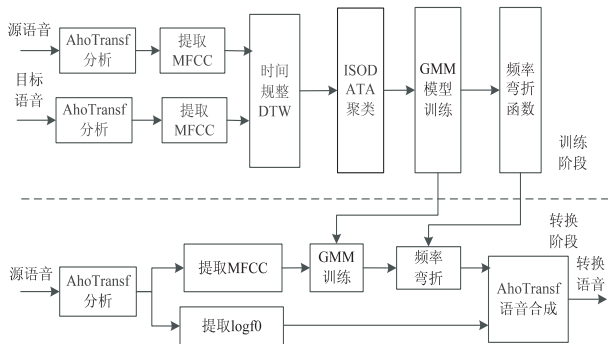


图 1 系统转换框图

序列和目标序列进行 BLFWA 训练, 得到频率弯折因子和幅度压扩因子, 根据式(6) 得到频率弯折曲线函数, 即转换函数。

在转换阶段, 源语音信号经过 AhoTransf 得到 MFCC 和 log 基音频率, 其中对 MFCC 进行 GMM 训练, 利用训练阶段获得的 GMM 模型均值、方差等参数训练得到概率 P 。获得概率 P 后加上训练阶段得到的频率弯折因子和幅度压扩因子(即转换函数)对输入的待转换源语音进行 BLFWA 转换, 转换后的频谱加上 log 基音频率通过 AhoTransf 模型合成出转换后的语音信号。

3 实验结果分析

3.1 语音库

实验采用的语音库 CMU ARCTIC 是由卡内基梅隆大学的语言技术研究所创建的美式英语单说话人平行语音库, 包括 5 男 2 女。该实验采用的特征参数为 MFCC 矢量, 信号的采样率为 16 kHz。抽取其中 4 个人的语音, 即 2 个男声和 2 个女声, 分别命名为 M_1 、 M_2 和 F_1 、 F_2 。每个人都取 60 个语句, 每个语句大概为 3~4 s 时长的短语, 其中 50 个用于训练, 10 用于测试。而且每个人的发音内容相同, 为对称的语音库。

经过大量实验发现, 在 ISODATA 聚类过程中, $\theta_c = 0.2$ (合并依据的聚类中心距离阈值), $\theta_s = 0.01$ (类内标准差阈值), $C > 35$ (预期的类数) 时, 获得最大分类数 31, 且实验所获得转换语音最佳。于是不同模型分类数均设置为 31 并进行比较。

3.2 客观评价

整个实验根据转换方向的不同分为 4 部分, 分别是女声转换为男声 ($F_1 - M_1$)、女声转换为女声 ($F_1 - F_2$)、女声转换为男声 ($F_2 - M_2$) 和男声转换为男声 ($M_2 - M_1$)。采用梅尔倒谱失真 (Mel-Cepstral Distortion, MCD)^[15] 作为反映语音转换性能的客观准则。

$$MCD(V^{Arg}, V^{ref}) =$$

$$\frac{10\sqrt{2}}{T \ln 10} \sum_{t=0}^{T-1} \sqrt{\sum_{d=1}^D (\nu_d^{Arg}(t) - \nu_d^{ref}(t))^2} \quad (9)$$

其中, MFCC 参数为 20-D 梅尔倒谱参数, 使用 $v_d(t)$ 表示, $0 \leq d \leq 19$, 在计算 MCD 时, 去掉第一维参数; T 为 MFCC 经过 DTW 对齐后的总帧数。

不同转换模型下的 MCD 值见图 2。

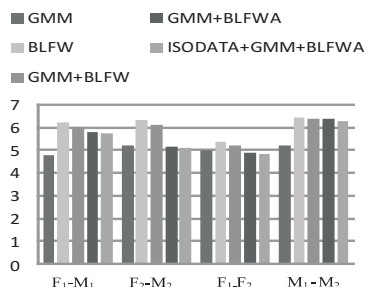


图2 不同转换模型下的 MCD 值

从图 2 可以看出, BLFW、GMM + BLFW、GMM + BLFWA 和 ISODATA+GMM+BLFWA 的 MCD 值是依次递减的, 这说明 ISODATA+GMM+BLFWA 的转换效果比上述几种模型要好。经过计算, ISODATA+GMM+BLFWA 的平均 MCD 值为 5.496, GMM 模型的平均 MCD 值为 5.0431, 说明 ISODATA+GMM+BLFWA 模型和 GMM 模型的转换相似性基本相当。

3.3 主观评价

实验的主观评价采用平均主观意见分 (MOS)。让听音人听完转换语音后, 给出意见分 (5: 优秀, 4: 良好, 3: 一般, 2: 较差, 1: 很差)^[16]。测试结束后, 统计出平均意见得分。MOS 越高, 说明转换语音的清晰度和可懂度越好。结果如图 3 所示。

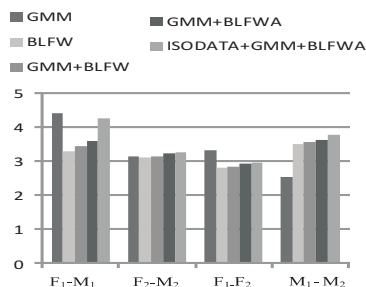


图3 不同转换模型下的 MOS 值

从图 3 可以看出, 采用文中方法训练的主观意见得分明显高于 BLFW、GMM+BLFW 和 GMM+BLFWA 的 MOS 分, 表明改进的 ISODATA+GMM+BLFWA 模型转换的语音目标倾向性和质量有明显改善, 降低了残差分量造成的影响; 经过计算, ISODATA+GMM+BLFWA 的平均 MOS 值为 3.568, GMM 模型的平均 MOS 值为 3.342, 说明 ISODATA+GMM+BLFWA 模型转换音质比 GMM 模型好。

4 结束语

ISODATA+GMM+BLFWA 模型通过 ISODATA 聚类算法对语音特征参数进行处理和分析, 得到更为精

确的分类。从 MOS 及 MCD 测试结果表明, 改进的 ISODATA+GMM+BLFWA 模型有效地降低了残差分量造成的影响, 在保证变换语音目标倾向性的同时, 提高了转换语音的音质。

参考文献:

- [1] 赵力. 语音信号处理[M]. 北京: 机械工业出版社, 2003.
- [2] 李波, 王成友, 蔡宣平, 等. 语音转换及相关技术综述[J]. 通信学报, 2004, 25(5): 109-118.
- [3] Erro D, Moreno A, Bonafonte A. Voice conversion based on weighted frequency warping[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(5): 922-931.
- [4] Erro D, Polyakova T, Moreno A. On combining statistical methods and frequency warping for high-quality voice conversion[C]//International conference on acoustics, speech and signal processing. [s. l.]: IEEE, 2008: 4665-4668.
- [5] 孙即详. 现代模式识别[M]. 长沙: 国防科技大学出版社, 2002.
- [6] Erro D, Navas E, Hernaez I. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(3): 556-566.
- [7] 徐小峰. 基于 GMM 的独立建模语音转换系统研究[D]. 苏州: 苏州大学, 2010.
- [8] 王韵琪, 俞一彪. 自适应高斯混合模型及说话人识别应用[J]. 通信技术, 2014, 47(7): 738-743.
- [9] Demrsra A P, Lamb N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B, 1977, 39(1): 1-38.
- [10] Xu L, Jordan M I. On convergence properties of the EM algorithm for Gaussian mixtures[J]. Neural Computation, 1996, 8(1): 129-151.
- [11] McDonough J, Byrne W. Speaker adaptation with all-pass transforms[C]//International conference on acoustics, speech, and signal processing. [s. l.]: IEEE, 1999: 757-760.
- [12] Pitz M, Ney H. Vocal tract normalization equals linear transformation in cepstral space[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(5): 930-944.
- [13] Emori T, Shinoda K. Rapid vocal tract length normalization using maximum likelihood estimation[C]//Proceedings of Eurospeech. [s. l.]: [s. n.], 2001: 1649-1652.
- [14] Saratxaga I, Hernández I, Navas E, et al. AhoTransf: a tool for multiband excitation based speech analysis and modification[C]//Proceedings of LREC. [s. l.]: [s. n.], 2010: 3733-3737.
- [15] Shuang Z, Meng F, Qin Y. Voice conversion by combining frequency warping with unit selection[C]//International conference on acoustics, speech and signal processing. [s. l.]: IEEE, 2008: 4661-4664.
- [16] 张雄伟, 陈亮, 杨吉斌. 现代语音处理技术及应用[M]. 北京: 机械工业出版社, 2003.