第 42 卷第 6 期　　声　学　学　报　　Vol. 42, No. 6
2017 年 11 月　　ACTA　ACUSTICA　　Nov., 2017

# 短时频谱通用背景模型群联合韵律的年龄语音转换*

惠　琳　　俞一彪
(苏州大学 电子信息学院　苏州　215006)

2016 年 5 月 6 日收到
2016 年 10 月 17 日定稿

**摘要**　提出一种短时频谱通用背景模型群与韵律参数相结合进行年龄语音转换的方法。谱参数转换方面，同一年龄段各说话者提取语音短时谱系数并建立高斯混合模型，然后依据语音特征相似性对说话者进行聚类，每一类训练一个通用背景模型，最终得到通用背景模型群和一组短时频谱转换函数。谱参数转换之后再对共振峰进一步微调。韵律参数转换方面，基频和语速分别建立单高斯和平均时长率模型来推导转换函数。实验结果显示，提出的方法在 ABX 和 MOS 等评价指标上比传统的双线性法有明显的优势，相对单一通用背景模型法的对数似然度变化率提高了 4%。这一结果表明提出的方法能够使转换语音具有良好目标倾向性的同时有较好的语音质量，性能较传统方法有明显提升。
PACS 数：43.72

## Voice conversion of different ages using universal background model groups of short-time spectra and prosodic features

HUI Lin　YU Yibiao
(*School of Electronic and Information Engineering, Soochow University*　Suzhou　215006)

Received May. 6, 2016
Revised Oct. 17, 2016

**Abstract**　For the voice conversion of different ages, a method using Universal Background Model Groups (UBMG) of short-time spectra and prosodic features is proposed. In spectrum aspect, Gaussian Mixture Model (GMM) is trained for every speaker after extracting linear predictive cepstrum coefficients, then the speakers in the same age period are clustered based on their voice similarity, and each cluster is further trained to be a UBM of spectrum distribution. Finally, an UBM group and corresponding spectrum conversion functions are obtained in each age period. Formants adjustment is further used after spectrum conversion. Furthermore, fundamental frequency and speech rate are modeled by single Gaussian and average duration rate respectively to derive their conversion functions in the aspect of prosodic features. The results of objective and subjective evaluation experiments such as ABX and MOS show that the proposed method has a distinct advantage compared with conventional bilinear method and its change rate of log-likelihood ratio increases by 4% compared with single UBM method. The results show the proposed method can make the converted speech more close to the speech of target age period with good speech quality while the performance has been improved evidently compared with conventional methods.

## 引言

年龄语音转换作为具有表现力的语音合成技术之一，是指在保持文本信息不变的情况下，改变语音中说话者的年龄信息，将具有源年龄段特征的语音转换为具有目标年龄段特征的语音。年龄语音转换一般都要求是非特定人的，即转换模型适用于某个年龄段的所有说话者，而非某个特定说话者，这与一般的语音转换有很大区别。年龄语音转换在智能人

* 国家自然科学基金项目 (61271360) 资助

机交互、动画配音和医疗健康等领域具有重要应用价值。

20 世纪 80 年代以来, 国内外的语音工作者在语音合成和转换方面做了大量的研究工作并取得了一定的成果。随着语音合成技术的发展, 人们开始关注合成语音的表现力, 说话者性别、情感和年龄等特征信息的提取与合成得到了空前关注 [1−2]。在年龄语音的特征分析方面, Schötz 研究得出, 从声学特征角度, 说话者语音的年龄信息主要决定于频谱信息、基频和语速, 其中频谱信息较为重要 [3]。Stölten 和 Engstrand 的研究表明, 可以从基频和语速的变化来反映年龄特征 [4]。Minematsu 等人的研究认为, 增加语速、基频、能量这些韵律参数可以提高年龄语音的表现力 [5]。方尔庆等人对年龄语音的分析表明, 对于成年人来说, 随着年龄的增长, 基频、语速、共振峰、声压值等声学特征会发生明显的变化 [6]。在年龄语音转换和合成方面, Traunmüller 通过调整基频、时长等参数来合成年龄语音 [7]; Schötz 用带有权重系数的线性插值法得到语音特征参数并用共振峰来合成非特定人的年龄语音 [8]。这些方法主要考虑了韵律参数和声道长度与说话者年龄的关系, 缺少对不同年龄语音短时频谱的差异性分析与运用, 效果并不理想。2005 年, Pitz 等发现不同说话者在语音倒谱空间呈现一种双线性变换映射关系 [9], 之后, Erro 等许多学者基于双线性变换尝试年龄语音的转换研究 [10−11], 但由于无法很好确定不同说话者以及不同年龄段的转换因子, 客观上存在同一年龄段不同说话者语音特征散度大的问题, 使得这一方法难以得到理想的效果。年龄语音转换的主要困难在于: (1) 语料库的建立较为困难, 很难在同一条件下录制同一人不同年龄段的语音; (2) 即使是同一年龄段的说话者, 其语音频谱和韵律等声学特征的差异性很大。总体来说, 目前非特定人年龄语音的转换合成方法在鲁棒性、目标倾向性和自然度方面的性能还有待提高, 仅仅依靠韵律特征或者短时频谱都不能有效实现高性能的年龄语音转换, 两者的综合运用和有效处理需要进一步深入研究。

随着年龄的改变, 说话者语音的特征参数会随之改变且相同年龄段说话者语音的特征参数具有一定的相似性 [12]。因此, 本文提出一种非特定人的年龄语音转换方法: 针对儿童、青年、中年和老年这 4 个年龄段, 依据不同性别, 从短时频谱特征和韵律特征两个方面分别进行分析, 建立年龄语音的转换模型。短时频谱特征由线性预测倒谱系数 LPCC (Linear Prediction Cepstrum Coefficient) 表示。韵律特征参数包括基频和语速, 其中语速由平均时长表示。实验和评价结果表明, 本文提出的方法是有效的, 转换合成的年龄语音不仅有很好的目标倾向性, 而且有较高的语音质量。

## 1 年龄语音转换系统结构

本文提出的年龄语音转换系统框图如图 1 所示。系统使用 STRAIGHT 分析合成工具进行短时谱和基频的提取以及转换阶段的语音合成处理。

在训练阶段, 属于源年龄段的各说话者语音样本通过 STRAIGHT 提取 LPCC 倒谱和基频参数, 然后依据每个说话者的 LPCC 倒谱参数分别训练高斯混合模型 GMM (Gaussian Mixture Model), 并对各个 GMM 模型进行自适应聚类, 进一步根据聚类结果训练得到一个描述源年龄段语音短时谱特征分布的 UBM (Universal Background Model) 群, 同时, 结合聚类结果和目标年龄段语音样本训练得到与该 UBM 群对应的一组短时谱转换函数。另一方面, 对各年龄段语音的基频、时长和共振峰进行统计分析, 推导相应的转换函数。

在转换阶段, 源年龄段测试语音可以通过 STRAIGHT 来提取 LPCC 倒谱参数和基频, 依据 LPCC 倒谱参数计算每个 UBM 的似然度, 依据最大似然准则选择相应的短时谱转换函数进行短时谱转换, 然后依据共振峰转换函数进行共振峰微调, 并进一步根据基频和平均时长转换函数修正韵律参数, 最终由 STRAIGHT 合成并输出目标年龄语音。

图 1 所示的非特定人年龄语音转换系统具有宽泛的适应能力。UBM 群模型以及后续的共振峰微调减小了相同年龄段不同说话者发音的差异性。改变说话者, 系统无需进行重新训练, 系统的灵活性和鲁棒性都比较好。

## 2 谱参数的转换

不同年龄段说话者的语音声学特征是不同的, 从生理角度上来看主要是由声道长度决定。随着年龄的增长, 成长阶段人的声道长度逐渐变长 [13]。从声学角度来看, 声道长度的变化主要反映在语音的频谱包络以及共振峰上的区别。因此, 年龄语音转换的一个重要方面就是短时频谱的转换。

### 2.1 UBM 群与短时谱转换函数

与说话者语音转换不同, 年龄语音转换要考虑非特定人特性, 短时谱转换模型必须对同一年龄段的说话者具有普适性。UBM 模型可以用来描述与说

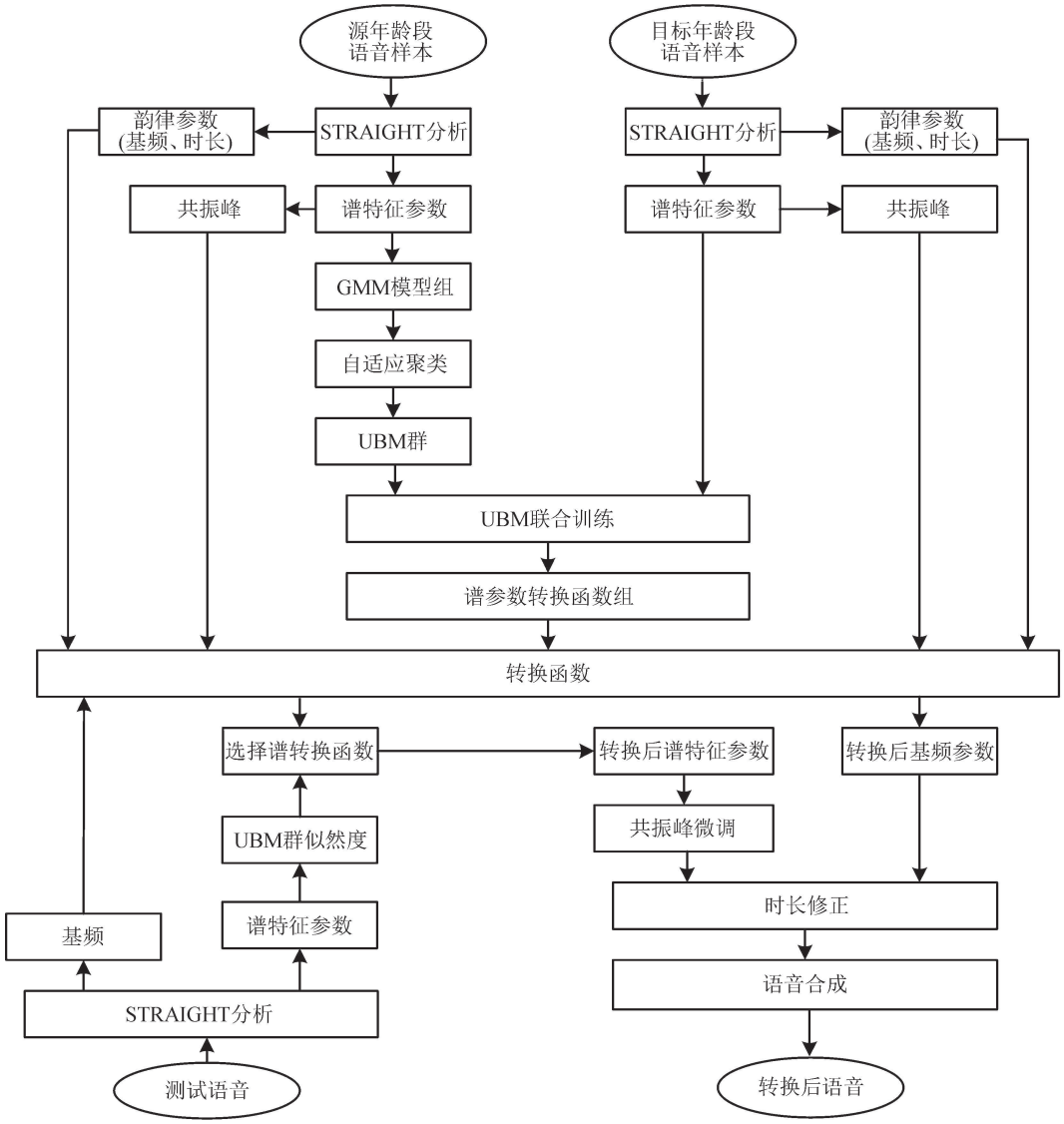


图 1 非特定人年龄语音转换系统框图

话者无关的短时谱平均分布统计特性，训练数据是这个年龄段所有说话者的语音数据。但由于同一年龄段说话者之间发音的差异性，会使得 UBM 模型空间的离散度较大，描述该年龄段的短时谱分布特征过于平滑，从而影响转换语音的质量。

针对上述情况，本文采用一种自适应聚类技术来训练各年龄段语音短时谱的 UBM 群模型，并由此推导年龄语音的短时谱转换函数。首先，提取训练集中该年龄段每个说话者语音的 LPCC 倒谱系数并训练各自的 GMM 模型，然后依据模型间距离度量准则，将该年龄段中具有相似语音特征的说话者聚为一类，进一步以类为单位训练每一类的 UBM，最终形成一个 UBM 群。短时谱转换函数基于 UBM 群中每一个 UBM 模型对应的源年龄段语音和目标年龄段语音进行联合训练推导得到，每一个 UBM 对应一个转换函数。年龄语音转换时，首先计算测试语音对应年龄段的 UBM 群中每一个 UBM 的似然度，并依据最大似然准则选择相应的短时谱转换函数进行短时谱转换。

### 2.1.1 GMM 模型间距离的度量

GMM 模型的参数概括了说话者语音特征的分布特性，因此把 GMM 模型参数看成该说话者语音特征参数分布的代表，考察这个代表在另一个 GMM 模型中的对数似然度，由此来衡量它们之间的距离。

设说话者 1 和 2 对应的模型分别是 GMM1 和 GMM2，则 GMM1 中 $n$ 个高斯分量的均值向量在 GMM2 中的对数似然度为：

$$L(\mathrm{GMM1},\mathrm{GMM2})=\sum_{i=1}^{n}a_{i1}\log p(\boldsymbol{\mu}_{1i}|\lambda_2), \qquad (1)$$

其中，$\boldsymbol{\mu}_{1i}$ 是 GMM1 中第 $i$ 个高斯分量的均值向量，$\lambda_2$ 是 GMM2 中的参数，$p(\boldsymbol{\mu}_{1i}|\lambda_2)$ 则是 GMM1 中第

$i$ 个高斯分量的均值向量在 GMM2 模型中的概率，$a_{i1}$ 是 GMM1 中第 $i$ 个高斯分量的权重值。同样可计算 GMM2 中 $n$ 个高斯分量的均值向量在 GMM1 中的对数似然度。

考虑到对称性，GMM1 和 GMM2 之间的距离用两者似然度均值的相反数来描述：

$$d_{1,2}=-\frac{L(\mathrm{GMM1},\mathrm{GMM2})+L(\mathrm{GMM2},\mathrm{GMM1})}{2}. \tag{2}$$

GMM 间对数似然度越大，则距离越短，表明两个说话者的语音特征越相似。

### 2.1.2 聚类与 UBM 群

聚类以说话者 GMM 为基础，目的是将同一年龄段各说话者根据其语音特征分布的相似性进行归类，在此基础上训练每一个类别的 UBM 形成 UBM 群，以减少单一 UBM 模型在说话者和短时频谱包络两方面的过平滑问题。

设定初始系数 $k\in(0,1)$，按最大最小距离聚类算法得到初始分类 ((1)—(5))，然后训练 UBM 群 ((6)—(8))。具体步骤如下：

(1) 随机选择该年龄段第一个说话者为第一聚类中心，并选择与第一聚类中心具有最大距离 $d$ 的说话者为第二聚类中心。

(2) 计算每个说话者与确定的聚类中心的距离 $d_{i1},d_{i2},\cdots,d_{im}$，$i=1,2,\cdots,N$，其中 $m$ 和 $N$ 分别是聚类中心数和说话者个数，并选择较小的距离 $d_i=\min(d_{i1},d_{i2},\cdots,d_{im})$。

(3) 在所有的最小距离中选择一个最大值，若该值超过阈值 $T=k\times d$，则产生新的聚类中心，即最大距离所对应的那个说话者，并继续，否则聚类结束。

(4) 重复以上步骤 (2) 和 (3)，直到没有新的聚类中心出现。

(5) 将所有说话者根据其与聚类中心的距离按最近邻准则进行归类，完成初始分类。

(6) 根据类内所有说话者语音特征参数训练 UBM 模型。

(7) 采用式 (1) 和式 (2) 计算各说话者到各 UBM 的距离，并按最近邻原则把每个说话者进行归类。

(8) 返回步骤 (6)，直到各类成员不再变化为止。

需要注意的是，阈值 T 的选取将会决定最终分类的个数，需要不断调节系数 $k$(通常来说，$0.5\leqslant k<1$)，比较多次分类结果，以满足类中的说话者个数分布都比较均衡为宜。

### 2.1.3 短时谱转换函数

在 UBM 群形成后，将 UBM 群中的每个 UBM 所对应的源年龄段语音 LPCC 参数 $\boldsymbol{x}_i$ 分别与目标年龄段语音 LPCC 参数 $y_i$ 进行联合训练，推导出一组短时谱转换函数。一个 $M$ 阶的 UBM 用 $M$ 个高斯分量来描述联合矢量 $\boldsymbol{z}_i$ 的概率分布如下式所示：

$$p(\boldsymbol{z}_i|\lambda)=\sum_{m=1}^{M}\alpha_m N(\boldsymbol{z}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m), \tag{3}$$

其中联合矢量 $\boldsymbol{z}_i=[\boldsymbol{x}_i^{\mathrm{T}},\boldsymbol{y}_i^{\mathrm{T}}]$；$\alpha_m$ 是混合权重值且 $\sum_{m=1}^{M}\alpha_m=1$；$N(\boldsymbol{z}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ 是第 $m$ 个高斯分布的概率密度函数，$\boldsymbol{\mu}_m$ 和 $\boldsymbol{\Sigma}_m$ 分别是联合矢量的均值和协方差矩阵。

$$\boldsymbol{\mu}_m=\begin{bmatrix}\boldsymbol{\mu}_m^x\\ \boldsymbol{\mu}_m^y\end{bmatrix},\quad \boldsymbol{\Sigma}_m=\begin{bmatrix}\boldsymbol{\Sigma}_m^{xx} & \boldsymbol{\Sigma}_m^{xy}\\ \boldsymbol{\Sigma}_m^{yx} & \boldsymbol{\Sigma}_m^{yy}\end{bmatrix}. \tag{4}$$

通过最大期望 (EM) 算法可以估计出该 UBM 的参数 $\lambda=(\alpha_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$。则短时谱转换函数为：

$$F(\boldsymbol{x}_i)=\sum_{m=1}^{M}p(m|\boldsymbol{x}_i,\lambda)[\boldsymbol{\mu}_m^y+\boldsymbol{\Sigma}_m^{yx}(\boldsymbol{\Sigma}_m^{xx})^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_m^x)], \tag{5}$$

其中 $p(m|\boldsymbol{x}_i,\lambda)$ 是特征矢量 $\boldsymbol{x}_i$ 属于第 $m$ 个高斯分布的后验概率：

$$p(m|\boldsymbol{x}_i,\lambda)=\frac{\alpha_m N(\boldsymbol{x}_i;\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^{xx})}{\sum_{k=1}^{M}\alpha_k N(\boldsymbol{x}_i;\boldsymbol{\mu}_k^x,\boldsymbol{\Sigma}_k^{xx})}. \tag{6}$$

在转换阶段，计算测试语音在 UBM 群中每个 UBM 的似然度，似然度最大的 UBM 所对应的谱转换函数即为最佳转换函数。

### 2.1.4 UBM 高斯分量数分析

图 2 描述了在不同的模型分量数下，Y-C, Y-M 和 Y-O 三种转换方向上转换语音平均对数似然度变化率 (见 4.1 小节) 的情况。可以看到，随着模型分量数从 32 增加到 96，平均对数似然度变化率提升很快，但在 96 以上提升开始变小，虽然因为模型拟合

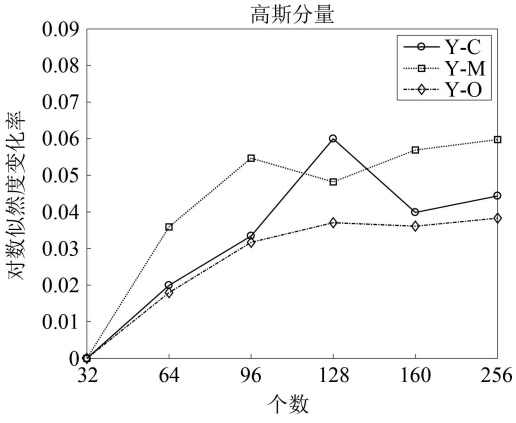


图 2 不同分量数下的对数似然度变化率

问题造成 Y-C 和 Y-M 两个方向有一定波动，但总体上开始趋于平稳，说明再进一步增加分量数对性能提高的贡献已经不大。考虑到运算量与转换性能间的平衡，本文选取的模型分量个数为 96。

### 2.2 共振峰微调

基于 UBM 群的似然度计算和短时频谱转换函数选择较好地解决了转换频谱的过平滑现象，但仍然有可能导致共振峰等一些重要的频谱细节模糊甚至信息的丢失[14]。因此有必要通过对转换频谱的共振峰进行微调来更细致地描述转换年龄语音的频谱包络。

共振峰反映了说话者的声道特性。共振峰频率越高，声道长度越短。而前两个共振峰又描述了元音最重要的频谱特征，其统计特性体现了某个年龄段语音在该元音上的共性。本文中共振峰值通过 LPC 求根法得到候选值，并通过一些条件来约束和消除虚假峰，规则如下：

(1) 删除共振峰频率小于 200 Hz 所对应的共轭极点。一般来说第一共振峰频率大于 200 Hz。

(2) 删除共振峰带宽大于 800 Hz 所对应的共轭极点。

(3) 当两个共轭极点所对应的共振峰频率之差小于 350 Hz, 且其中一个共轭极点的共振峰带宽大于 450 Hz, 则删除该虚假峰所对应的共轭极点。

共振峰公式为：

$$F_i = \frac{\theta_i}{2\pi} f_s, \tag{7}$$

其中 $f_s$ 是采样频率，$\theta_i$ 是共轭极点所对的相位角。共振峰偏移量为：

$$dF = F_i' - F_i = \frac{(\theta_i' - \theta_i)}{2\pi} f_s = \frac{d\theta f_s}{2\pi}, \tag{8}$$

其中 $F_i$ 和 $F_i'$ 分别是偏移前后的共振峰值；$\theta_i$ 和 $\theta_i'$ 分别是偏移前后共轭极点所对的相角；$d\theta$ 则是相角偏移量。所以相角偏移量：

$$d\theta = \frac{2\pi dF}{f_s}. \tag{9}$$

例如，把青年男性元音\a\的一帧语音进行共振峰微调，对其前两个共振峰分别下调 100 Hz, 结果如图 3 所示 (图中实线和虚线分别是是搬移前和搬移后的声道传递函数功率谱曲线)。

经过统计分析，表 1 为青年语音共振峰参数与其他各个年龄段共振峰参数相比的偏移量。其中，C, Y, M, O 分别代表了儿童、青年、中年以及老年段。由于在儿童段，说话者还未经历变声期，共振峰参数的偏移量并不稳定，但总体上来说，随着年龄的增长，共振峰逐渐降低且女性的共振峰偏移量大于男性。

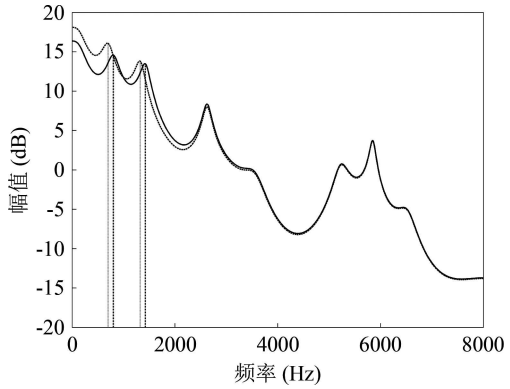


图 3 青年男性元音\a\的共振峰搬移

表 1 青年与其它年龄段共振峰偏移值 (单位：Hz)

| 男 (女) | $\Delta F1$ | $\Delta F2$ |
|---|---|---|
| Y-C | 86(9) | 35(45) |
| Y-M | -12(-18) | -27(-29) |
| Y-O | -20(-30) | -57(-50) |

## 3 韵律参数的转换

韵律特征是表现语音个性化的重要因素，同样也是年龄语音转换的重要内容。文献 12 详细介绍分析了各年龄段的时长和基频统计分布信息。本文主要考虑采用基频和平均时长两种韵律参数进行年龄语音转换[15]。

### 3.1 时长

不同年龄段的说话者其发音语速是不同的，本文用语句的平均时长率 $g$ 来描述目标年龄段说话者时长与青年说话者的时长比值。

$$T_t = g \times T_s, \tag{10}$$

$T_t$ 是目标年龄段的平均时长，$T_s$ 是源年龄段的平均时长。

从表 2 中可以看出，儿童段说话者的语速较慢，在青年段时语速逐渐变快；随着年龄继续增长，语速从中年开始又变慢，老年的语速甚至比儿童的语速还慢。

表 2 平均时长率

| 性别 | 儿童 | 青年 | 中年 | 老年 |
|---|---|---|---|---|
| 男 | 1.2333 | 1 | 1.1154 | 1.3175 |
| 女 | 1.3248 | 1 | 1.2018 | 1.3893 |

### 3.2 基频

基音频率 F0 (Fundamental frequency) 反映说话者发音时的激励特性, 不同年龄段的说话者由于其生理特性的变化而具有不同的基音频率特性。所以, 在年龄语音转换的过程中, 基频也是一种重要的韵律转换参数。

本文用单高斯模型来描述源和目标年龄段说话者语音的基频参数分布情况。$\mu_s$ 和 $\mu_t$ 是源和目标年龄段语音样本的平均基频值, $\sigma_s$ 和 $\sigma_t$ 是它们的方差。则基频转换函数如下所示：

$$F0_t = \frac{\sigma_t}{\sigma_s}(F0_s - \mu_s) + \mu_t. \tag{11}$$

## 4 实验与分析

实验采用的年龄语音库选用苏州大学语音技术实验室的 SUDA_NL 语音库 [12]。该库包含了儿童、青年、中年和老年 4 个年龄段 169 名说话者的语音。每个说话者的语音样本长度为 150 s, 文本选自于小学语文课本。语音采集采用 ZOOM H4 设备, 在安静的房间中录制, 采样频率为 16 kHz, 16 bit 量化。年龄段的划分为：儿童 (8～10 岁)、青年 (20～25 岁)、中年 (39～53 岁)、老年 (62 岁以上)。每个年龄段中分别选取 10 名男性和女性说话者的语音作为训练语料。

为了检验本文提出的年龄语音转换方法的有效性, 将双线性法 [9,11] 和 UBM 年龄语音转换法 [12] 与本文提出的年龄语音转换方法 (UBMG+ 韵律) 进行对比。每个方法中设计了 3 组年龄语音转换实验, 分别是青年转儿童 (Y-C)、青年转中年 (Y-M)、青年转老年 (Y-O), 每组都包含有男性和女性组。以上实验通过主观评价和客观评价对结果进行评价。实验中训练集语音和测试集语音没有重合, 训练集和测试集语音文本一致。谱特征参数由 24 维 LPCC 参数表示。

### 4.1 客观评价

由于年龄语音转换的非特定人性质, 客观评价中的谱失真测度并不适用。为此提出用平均对数似然度变化率来评价转换性能。主要是比较测试语音在目标年龄段样本中的平均对数似然度, 似然度越大, 测试语音转换的效果越好。其中平均对数似然度计算公式为：

$$D = \frac{1}{N}\log\prod_{i=1}^{N} p(\boldsymbol{x}_i), \tag{12}$$

$p(\boldsymbol{x}_i)$ 是第 $i$ 帧语音样本在目标年龄段 UBM 模型中的概率分布, $N$ 是帧数, 则平均对数似然度变化率 $d$ 为：

$$d = \frac{D_2 - D_1}{|D_1|}, \tag{13}$$

其中双线性法的平均对数似然度为 $D_1, D_2$ 是其余两个方法的平均对数似然度。通过比较 $d$ 值, 来考察上文提出的其余两个方法较双线性法的转换性能。$d$ 值越大, 该方法的转换效果越好。实验结果如表 3 所示, 其中括号内外的数字分别是女性和男性说话者的对数似然度变化率。可以看出, 较传统的双线性法, UBM 法的似然度变化率有了明显的提高。在 UBM 法的基础上, 对谱参数转换方式进行改进后, 本文所提出的 “UBMG+ 韵律” 方法的似然度变化率进一步提高了 4%。这个客观评价实验说明了本文提出的方法能够有效提高年龄语音的转换效果, 使转换语音更具有目标倾向性。

表 3　对数似然度变化率

| 男 (女) | Y-C | Y-M | Y-O |
|---|---|---|---|
| UBM+ 韵律 | 14%(21%) | 38%(31%) | 42%(35%) |
| UBMG+ 韵律 | 20%(25%) | 41%(34%) | 46%(39%) |

### 4.2 主观评价

(1) ABX 评价是用于判断转换后语音的目标倾向性。对转换后的语音 X 进行判断, 若接近源年龄段语音 A 则评分为 0 分, 接近目标年龄段语音 B 评分则为 1 分。若最终结果表明转换语音与目标年龄段语音接近, 则转换成功, 反之则为失败。在 ABX 测试中, 选择 8 个听者对每种转换方向的 8 句语音进行评分, 实验结果如表 4 所示。

表 4　ABX 测试

| 男 (女) | Y-C | Y-M | Y-O |
|---|---|---|---|
| 双线性 | 0.67(0.71) | 0.68(0.69) | 0.67(0.68) |
| UBM+ 韵律 | 0.81(0.85) | 0.78(0.82) | 0.79(0.81) |
| UBMG+ 韵律 | 0.85(0.88) | 0.81(0.86) | 0.83(0.86) |

从表 4 中可以看出, 相比双线性法和 UBM 法, 本文提出的 “UBMG+ 韵律” 方法的 ABX 评分更高, 转换效果更好。同时 Y-C, Y-O 的转换效果要好于 Y-M, 女性的转换效果要好于男性。该主观测试结果表明, “UBMG+ 韵律” 方法的转换性能较好, 转换后的语音更接近于目标年龄段的语音。

(2) MOS 评价是对转换后语音的质量进行测试。评分有 5 个等级, 1 分表明转换后语音质量很差, 5 分表明转换后语音质量很好。得分越高, 转换

后语音的自然度和可懂度越高。在 MOS 测试中选择 8 个听者对每种转换方向的 8 个语音进行评分，实验结果如表 5 所示。

表 5 MOS 测试

| 男 (女) | Y-C | Y-M | Y-O |
|---|---|---|---|
| 双线性 | 4.10(4.11) | 4.12(4.10) | 4.04(4.02) |
| UBM+韵律 | 4.22(4.27) | 4.23(4.26) | 4.21(4.25) |
| UBMG+韵律 | 4.23(4.29) | 4.23(4.27) | 4.20(4.26) |

从表 5 中可以看出本文提出的方法的 MOS 评分要高于双线性方法，与 UBM 法的评分相差不大。同时，Y-C 和 Y-M 的语音质量要高于 Y-O，且女性的转换效果要好于男性。通过 MOS 主观评价可以看出，“UBMG+韵律” 方法相对于双线性和 UBM 法，转换语音的质量具有一定的优势。

## 5 总结

本文采用短时频谱 UBM 群联合韵律特征的方法，分别从谱参数和韵律参数两个方面对语音的个性特征进行调整，实现了非特定人的年龄语音转换。由于年龄语音转换的非特定人特性，系统无需重复训练，提高了系统的灵活性和效率。在谱参数方面，通过自适应聚类技术把同一年龄段语音特性相似的说话者聚类组成 UBM 群，提高了同一年龄不同说话者语音特征描述的精准性，并在转换阶段进行共振峰微调，更好地表现不同年龄段的谱参数变化。在韵律参数方面，对年龄语音影响较大的平均时长和基频进行了调整，使得转换语音更具目标倾向性。主客观评价也显示了本文提出的方法要优于双线性和单一 UBM 方法，似然度变化率在单一 UBM 法的基础上提高了 4%，能够使转换语音更接近目标年龄段说话者语音的同时具有较好的语音质量。理论分析和实验结果都证明了该方法的有效性。

以后的研究还需深入考虑与说话者年龄相关的其它语音个性特征，并进一步运用到年龄语音转换中。

## 参 考 文 献


1 Kuwabara H, Sagisak Y. Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication*, 1995; **16**(2): 165—173

2 Türk O, Schröder M. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *IEEE Trans. On Audio, Speech and Language Processing*, 2010; **18**(5): 965—973

3 Schötz S. Towards synthesis of speaker age: A perceptual study with natural, synthesized and resynthesized stimuli. *Phonum*, 2003; **9**: 153—156

4 Stölten K, Engstrand O. Effects of perceived age on perceived dialect strength: A listening test using manipulations of speaking rate and F0. *Phonum*, 2003; **9**: 29—32

5 Minematsu N, Sekiguchi M, Hirose K. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In: Proc. ICASSP, 2002; **1**(1): I-137-I-140

6 方尔庆，耿新. 基于视听信息的自动年龄估计方法. 软件学报，2011; **22**(7): 1503—1523

7 Traunmüller H, Branderud P, Bigestans A. Paralinguistic speech signal transformations. PERILUS X, 1989: 47—64

8 Schötz S. F0 and segment duration in formant synthesis of speaker age. In: Proc. Speech Prosody. Dresden, 2006: 515—518

9 Pitz M, Ney H. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Trans. on Speech and Audio Processing*, 2005; **13**(5): 930—944

10 李金中，李贤，汪增福. 基于声道长度对齐的年龄语音转换. 中国科学技术大学学报， 2015; **45**(7): 575—581

11 Erro D, Navas E, Hernaez I. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Trans. on Audio, Speech, and Language Processing*, 2013; **21**(3): 556—566

12 Hui Lin, Yu YiBiao. Acoustic feature analysis and conversion of age speech. In: Proc. ICWMMN, 2015: 147—151

13 Vorperian H K, Wang S, Chung M K *et al.* Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study. *The Journal of the Acoustical Society of America*, 2009; **125**(3): 1666—1678

14 陈雪勤，赵鹤鸣. 有效高斯分量通用背景模型下耳语音声道系统转换研究. 声学学报， 2013; **38**(2): 195—200

15 李力，俞一彪. 采用超音段韵律特征联合短时频谱的语音转换. 信号处理， 2012; **28**(2): 289—294