文章编号:1004-9037(2013)02-0141-08

基于卷积非负矩阵分解的语音转换方法

孙 健[1] 张雄伟[2] 曹铁勇[2] 杨吉斌[2] 孙新建[1]

(1. 解放军理工大学通信工程学院,南京,210007; 2. 解放军理工大学指挥自动化学院,南京,210007)

摘要:为了在语音转换过程中充分考虑语音的帧间相关性,提出了一种基于卷积非负矩阵分解的语音转换方法。卷积非负矩阵分解得到的时频基可较好地保存语音信号中的个人特征信息及帧间相关性。利用这一特性,在训练阶段,通过卷积非负矩阵分解从训练数据中提取源说话人和目标说话人相匹配的时频基。在转换阶段,通过时频基替换实现对源说话人语音的转换。相对于传统方法,本方法能够更好地保存和转换语音帧间相关性。实验仿真及主、客观评价结果表明,与基于高斯混合模型、状态空间模型的语音转换方法相比,该方法具有更好的转换语音质量和转换相似度。

关键词:语音转换;卷积非负矩阵分解;时频基

中图分类号:TP391 **文献标志码**:A

Voice Conversion Based on Convolutive Nonnegative Matrix Factorization

Sun Jian[1], *Zhang Xiongwei*[2], *Cao Tieyong*[2], *Yang Jibin*[2], *Sun Xinjian*[1]

(1. Institute of Communication Engineering, PLA University of Science & Technology, Nanjing, 210007, China;
2. Institute of Command Automation, PLA University of Science & Technology, Nanjing, 210007, China)

Abstract: In order to fully consider the inter-frame correlation in voice conversion, a voice conversion method based on convolutive nonnegative matrix factorization is proposed. The personal characteristics and inter-frame correlation in voice can be well preserved in the time-frequency bases obtained from convolutive nonnegative matrix factorization. With this feature, during the training phase of voice conversion, the matching time-frequency bases of source and target speakers can be extracted from training data through convolutive nonnegative matrix factorization. Then in the conversion phase, the voice of source speaker is converted through time-frequency bases substitution. Compared with traditional methods, the inter-frame correlation in voice can be better preserved and converted in the proposed method. Experimental results using objective and subjective evaluations show that the proposed method outperforms the methods based on Gaussian mixture model and the state space model in the view of both speech quality and conversion similarity to the target speech.

Key words: voice conversion; convolutive nonnegative matrix factorization; time-frequency bases

引 言

语音转换是一种通过改变源说话人语音信号中的个人特征信息,使之具有目标说话人语音个人特征信息的技术[1]。语音转换技术在个性化语音合成、信息安全及多媒体娱乐领域都有着广阔的应用前景。例如,将语音转换技术加入到语音合成系统中,可实现个性化语音合成;通过语音转换,可伪造敌方人员声音来突破声纹识别准入系统;通过语音转换可再现历史人物演讲等。

最早的声道谱转换方法由 Abe 等人在 1988 年提出[2]。现有的转换方法主要包括:基于高斯混合模型(Gaussian mixture model, GMM)的方

基金项目:国家自然科学基金(61072042)资助项目;解放军理工大学预先研究基金(20110205,20110211)资助项目。
收稿日期:2011-11-17;**修订日期**:2012-02-17

法[3-4]、基于隐马尔科夫模型(Hidden Markov model, HMM)的方法[5]、基于频谱弯折(Frequency warping, FW)的方法[6-7]和基于人工神经网络(Artificial neural network, ANN)的方法[8]。但现有方法中仍存在一些尚未妥善解决的问题,如基于 GMM 和 ANN 的方法,一个重要的前提假设是:各语音帧间是相互独立的。这就造成语音帧间的时序相关性被忽略,从而导致转换结果的不连续性[9]。文献[9]提出通过决策树在转换模型中加入部分动态信息和语音信息,如音素信息、清浊音信息等。文献[10]基于语音声道谱静态参数与动态参数间的线性相关假设,对 GMM 模型进行了改进,提出了基于 GMM 的频谱参数轨迹转换方法。文献[11]基于同样的假设,提出了轨迹 HMM 模型并应用于语音转换。文献[12~13]使用状态空间模型(State space model, SSM)对语音信号的频谱进行建模,利用 SSM 可对频谱轨迹整体建模的特点来提升转换后语音的连续性。上述方法在转换中考虑了帧间相关性,并在一定程度上改善了转换后语音的质量,但是由于其试图以某种固定的参数来表征语音帧间的相关性,从而导致转换效果仍不尽如人意。文献[9]中使用的音素信息和清浊音信息是无法表征帧与帧间这种较小时间尺度上的相关性的;在文献[10~11]中,帧间的相关性被限定为明确的线性关系,从而限制了相关性范围;文献[12~13]中所提出的方法虽然试图从声道谱中直接提取语音帧间的相关性,但在转换阶段并没有考虑不同说话人语音帧间相关性的差异。

针对以上问题,本文考虑从声道谱参数中直接提取和保存语音帧间的相关性,并将体现说话人特征的相关性信息应用到声道谱转换中。通过对语音信号的分析,本文认为语音帧间的相关性可分为两类:(1) 局部帧间相关性,即语音局部相邻帧间存在的相关性,这一相关性可认为主要体现在音素内部;(2) 全局帧间相关性,即更大时间尺度上语音帧间的相关性,体现在音素间的关系上,主要受语音内容所决定。对局部帧间相关性的转换应考虑两方面问题,一是保证转换后声道谱帧间的连续性,二是去除局部帧间相关性中源说话人特征,使之具有目标说话人特征。

卷积非负矩阵分解(Convolutive nonnegative matrix factorization, CNMF)是一种针对语音信号处理所提出的非负矩阵分解方法,该方法在保证分解结果非负性的前提下,使用了二维时频基代替原非负矩阵分解中的一维基向量,从而有效地承载了语音信号的局部帧间相关性。该方法在多说话人语音的分离上已有成功的应用[14]。通过该方法可以将语音声道谱分解为一组非负时频基和对应的编码矩阵。分解得到的时频基可认为是承载了说话人个人特征的一个子空间,因而集中体现了说话人的个人特征信息,而编码矩阵则是声道谱参数在此子空间上的投影。因此,可考虑利用目标说话人时频基替换源说话人时频基以实现声道谱的转换。但 CNMF 存在分解结果不唯一的问题,即在不同初始条件下对同一语音声道谱参数分解得到的时频基和编码矩阵并不唯一。虽然这种不唯一性可理解为特征子空间的不同表现形式,但却造成了分解得到的源说话人和目标说话人时频基难以匹配的问题,阻碍了其在语音转换中的应用。

1 卷积非负矩阵分解

1.1 非负矩阵分解方法

基分解是信号分析的一种常用方法,通过基分解可将待分析的数据矩阵 $\boldsymbol{V}$ 投影到低维基矩阵 $\boldsymbol{W}$ 所构成的子空间上,并得到投影系数或称为编码矩阵 $\boldsymbol{H}$

$$\boldsymbol{V} \approx \boldsymbol{W} \cdot \boldsymbol{H} \tag{1}$$

式中:$\boldsymbol{V}$ 为 $M \times N$ 的矩阵,基矩阵 $\boldsymbol{W}$ 由 R 个列向量构成,且 $R \ll N$。式(1)也可写为

$$v_{ij} \approx (\boldsymbol{W} \cdot \boldsymbol{H})_{ij} = \sum_{r=1}^{R} w_{ir} h_{rj} \tag{2}$$

式中 v_{ij} 为 $\boldsymbol{V}$ 中第 i 行第 j 列的元素。

非负矩阵分解(Nonnegative matrix factorization, NMF)不同于一般基分解之处在于,分解结果中加入了非负性限制[15],即式(2)中 $w_{ir} \geqslant 0$, $h_{rj} \geqslant 0$。此外,R 应满足 $(M+N)R < MN$ 的条件。

由于具有非负性限制,使得 $\boldsymbol{V}$ 中的列向量是由各个非负的基向量通过叠加方式进行重构的。因此,各个基向量可看作是对 $\boldsymbol{V}$ 的部分分解,也可认为每个基向量都承载了 $\boldsymbol{V}$ 的部分特征。这种由整体到局部的分解,使 NMF 在应用于图像、语音等信号的分析时,结果具有更明确的物理意义[15]。

对于非负矩阵分解的具体实现步骤,Lee 等人[15-16]提出了两种方法:基于欧氏距离的误差最小化方法(式(3))和基于 Kullback-Leibler 散度的误差最小化方法(式(4))

$$E(\boldsymbol{V} \| \boldsymbol{WH}) = \sum_{i,j} (v_{ij} - (\boldsymbol{W} \cdot \boldsymbol{H})_{ij})^2 \tag{3}$$

$$D(\boldsymbol{V} \| \boldsymbol{WH}) = \sum_{i,j} \left[v_{ij} \log \frac{v_{ij}}{(\boldsymbol{W} \cdot \boldsymbol{H})_{ij}} - v_{ij} + (\boldsymbol{W} \cdot \boldsymbol{H})_{ij} \right] \tag{4}$$

基于欧式距离的误差最小化方法是在加性高斯噪声条件下对 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 的极大似然估计,而基于 Kullback-Leibler(KL)散度的误差最小化方法则是在观测数据由均值为$(\boldsymbol{W}\cdot\boldsymbol{H})_{ij}$ 的 Poisson 过程生成时的极大似然估计。相对而言,基于 KL 散度的误差最小方法使分解结果对于观测数据中的低频信息具有更好的重构效果。由于人耳对语音低频部分的变化更加敏感,因此在对语音信号进行分析时,采用基于 KL 散度的 NMF 较为合理。

1.2　卷积非负矩阵分解方法

在非负矩阵分解基础上,文献[14]考虑到语音信号的时间相关性,将原有基矩阵 $\boldsymbol{W}$ 中的列向量扩展为时频二维矩阵,提出了卷积非负矩阵分解,并将该方法应用于多路语音信号的分离中。NMF 中基矩阵与编码矩阵的点乘关系也由此扩展为时频二维基与编码矩阵的卷积关系

$$\boldsymbol{V}\approx\sum_{r=0}^{R-1}\boldsymbol{W}(r)\cdot\overrightarrow{\boldsymbol{H}}^{r} \tag{5}$$

式中:$\boldsymbol{W}(r)$表示第 r 个时频基,大小为 $M\times t$;$\overrightarrow{\boldsymbol{H}}^{r}$表示对编码矩阵 $\boldsymbol{H}$ 以列向量形式右移 r 个单位。

具体来说,如果

$$\boldsymbol{H}=[\boldsymbol{h}_1,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_N] \tag{6}$$

式中 $\boldsymbol{h}_i$ 为编码矩阵 $\boldsymbol{H}$ 的第 i 个列向量,$i=1,\cdots,N$,则

$$\overrightarrow{\boldsymbol{H}}^{r}=[\boldsymbol{0},\cdots,\boldsymbol{0},\boldsymbol{h}_1,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_{N-r}] \tag{7}$$

式中"**0**"为零值列向量。

依照文献[14]中给出的方法可通过多次迭代实现对 $\boldsymbol{W}(r)$和 $\boldsymbol{H}$ 的求解。

取 ARCTIC 语音库中 BDL 子库的前 10 句语音,首先通过 STRAIGHT 模型进行分析[17],得到其 STRAIGHT 谱矩阵 $\boldsymbol{S}$。其中每一列为一帧语音的 STRAIGHT 谱数据,帧移取为 5 ms。当 $R=40$,$t=12$ 时,通过 CNMF 对 $\boldsymbol{S}$ 进行分解,可得到如图 1 所示的一组时频基。

CNMF 相对非负矩阵分解来说,更好地发掘了时序信号的时间相关性,并将这种相关性体现在时频基中。通过图 1 可以看出,得到的各个时频基类似于语音信号各个音素的时频谱。但不同于音素时频谱的是,这些时频基是通过对声道谱参数直接分析提取得到的。此外,这些时频基在整个语音声道谱中的变化还需要通过编码矩阵加以体现。因此可认为,时频基中主要包含了语音信号中相对不变的频域结构特征,而编码矩阵则包含了这些特征的动态变化过程。由此,可进一步认为说话人个人特征信息主要包含在时频基中,而编码矩阵则主要承载了语义信息。

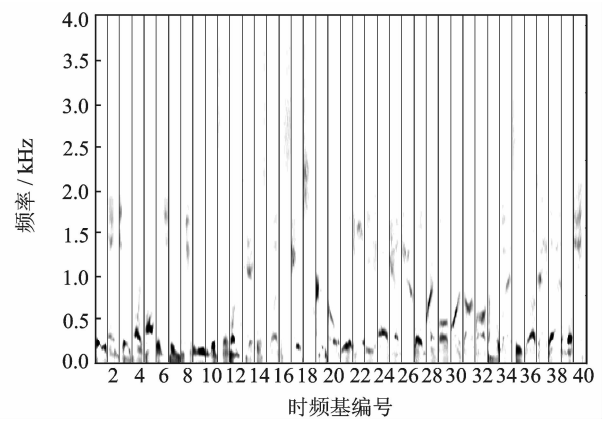


图 1　BDL 语音 STRAIGHT 谱的时频基示意图

2　基于卷积非负矩阵分解的语音转换

由第 1 节分析可知,通过 CNMF 对语音声道谱进行分析,可得到承载说话人个人特征信息的时频基及其相应的编码矩阵。因此,很自然可想到通过使用目标说话人的时频基替换源说话人的时频基,之后再与编码矩阵进行卷积,实现语音转换。但此方案在实施时却面临 CNMF 分解结果不唯一的问题。

同一声道谱矩阵 $\boldsymbol{V}$,在 $\boldsymbol{W}(r)$和 $\boldsymbol{H}$ 的初始值不同时,通过 CNMF 进行分析,最终得到的 $\boldsymbol{W}(r)$和 $\boldsymbol{H}$ 并不相同,即同一时频谱矩阵 $\boldsymbol{V}$ 具有多种时频基和编码矩阵的组合形式。因此,如果对源说话人和目标说话人的平行声道谱矩阵进行独立地卷积非负矩阵分解分析,则无法保证能够得到相同的、可表征内容信息的编码矩阵。为解决这一问题,本文借鉴文献[9]的方法,通过加入编码矩阵一致性限制,从而使源说话人和目标说话人声道谱矩阵经非负矩阵分解后,具有相同的编码矩阵和各自的时频基。之后再利用时频基替换在转换阶段实现转换。该方案流程如图 2 所示。

整个声道谱转换过程可分为训练和转换两个阶段,其中训练阶段由数据预处理(图 2 中①)和基于 CNMF 的时频基提取(图 2 中②)两部分组成。在图中③部分是基于时频基替换的声道谱转换过程。

2.1　数据预处理

用于训练的源语音和目标语音具有相同的内

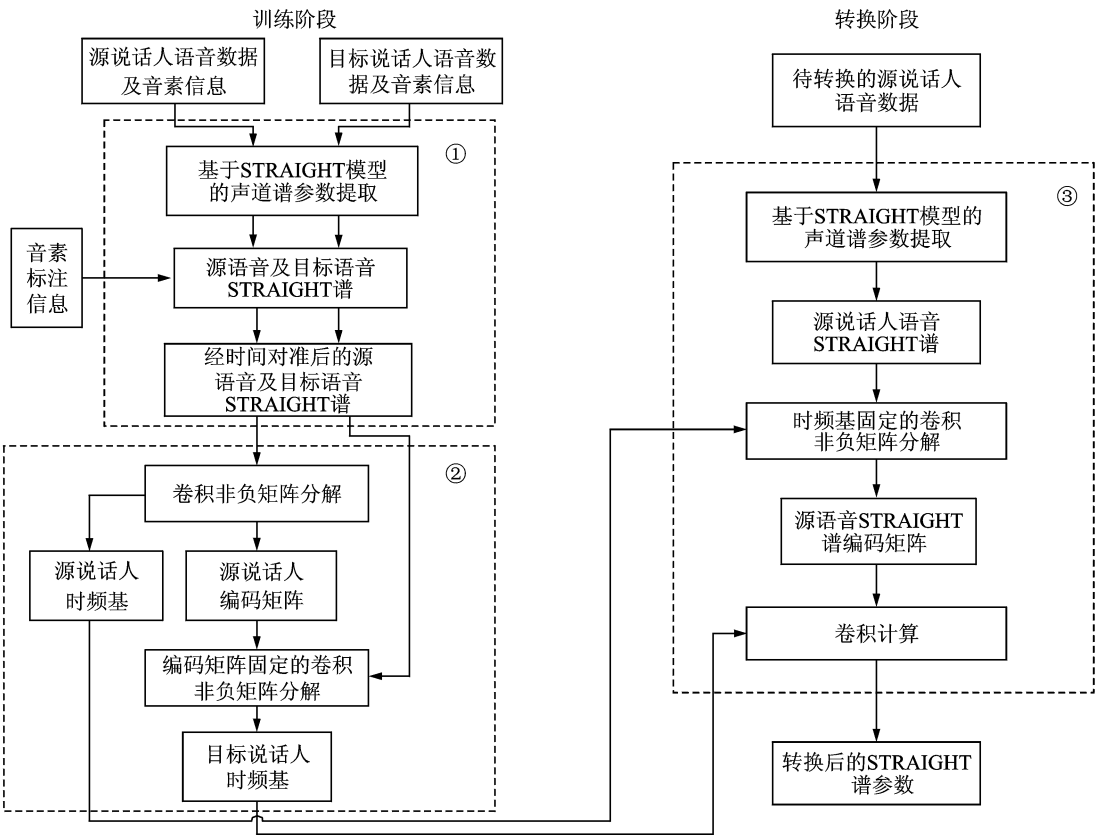


图 2 基于 CNMF 的语音声道谱转换流程示意图

容,分别用 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 表示。首先需要提取语音的声道谱参数。由于 STRAIGHT 模型可以较好地分离声源激励信息和声道信息[17],因此本文中采用 STRAIGHT 模型分解得到的 STRAIGHT 谱来表征语音的声道谱。对训练语音数据通过 STRAIGHT 模型进行分析,得到源语音和目标语音的 STRAIGHT 谱,分别用 $\boldsymbol{S}_x$ 和 $\boldsymbol{S}_y$ 表示。

为保证后续分解得到的编码矩阵的一致性,在进行非负矩阵分解前,还需要进一步对 $\boldsymbol{S}_x$ 和 $\boldsymbol{S}_y$ 进行时间对准处理。在现有的语音转换方法中,最常用的时间对准方式是动态时间规整(Dynamic time warping, DTW)[2-4,7-10,12-13]。但由于 DTW 中的帧插入和帧删除未考虑语音帧间的连续性,因此 DTW 处理过程中单帧的多次插入或连续多帧的删除,有时会造成对准后语音质量的大幅下降,不利于后续的 CNMF 处理。本文采用了基于音素标注信息的时间对准方法。

首先根据音素标注信息对源语音和目标语音的 STRAIGHT 谱以帧为单位,即列向量,进行分段处理。无声段也被当作独立音素进行分段。由于源语音和目标语音为含有相同内容的语音数据,因而分段后可得到源语音和目标语音相匹配的各段 STRAIGHT 谱数据,分别表示为

$$\boldsymbol{S}_x=[\boldsymbol{S}_1^x,\boldsymbol{S}_2^x,\cdots,\boldsymbol{S}_i^x,\cdots,\boldsymbol{S}_I^x] \tag{8}$$

$$\boldsymbol{S}_y=[\boldsymbol{S}_1^y,\boldsymbol{S}_2^y,\cdots,\boldsymbol{S}_i^y,\cdots,\boldsymbol{S}_I^y] \tag{9}$$

式中 $\boldsymbol{S}_i^x$ 和 $\boldsymbol{S}_i^y$ 分别为 $\boldsymbol{S}_x$ 和 $\boldsymbol{S}_y$ 经划分后,相对应的第 i 段 STRAIGHT 谱。具体来说

$$\boldsymbol{S}_i^x=[s_{i1}^x,s_{i2}^x,\cdots,s_{ij}^x,\cdots,s_{iJ_i}^x] \tag{10}$$

$$\boldsymbol{S}_i^y=[s_{i1}^y,s_{i2}^y,\cdots,s_{ik}^y,\cdots,s_{iK_i}^y] \tag{11}$$

式中:s_{ij}^x 和 s_{ik}^y 分别为 $\boldsymbol{S}_i^x$ 和 $\boldsymbol{S}_i^y$ 中的第 j 和第 k 列数据,J_i 和 K_i 分别是 $\boldsymbol{S}_i^x$ 和 $\boldsymbol{S}_i^y$ 中所含的帧数(列数)。

以源说话人语音各段 STRAIGHT 谱中的帧数为基准,对目标说话人语音的各段 STRAIGHT 谱中帧数进行调整,使两者帧数相等。使用源说话人 STRAIGHT 谱帧数为基准是因为在转换阶段从语音信号中提取的声道谱也为源说话人 STRAIGHT 谱。因此,在训练阶段保持源说话人语音声道谱参数不变,有利于保持这种一致性。

鉴于 STRAIGHT 较为平滑,因此在对目标说话人各段 STRAIGHT 谱进行帧数调整时,可通过对各行数据采用二阶 B 样条插值的方法进行处理。通过比较语音 STRAIGHT 谱经插值后再还原的结果与原始 STRAIGHT 谱间的差异可知,该处理方式对 STRAIGHT 谱的影响非常小。表 1 给出使用 ARCTIC 库中 BDL 子库和 SLT 子库的

第 1 句语音数据的 STRAIGHT 谱进行实验仿真,计算插值后再还原处理造成的误差。其中 α_{in} 为初始插值比例,当其大于 1 时为增加点数,小于 1 时为减小点数。

表 1　不同插值、还原比例对语音声道谱造成的误差

误差值	语音库子库	
	BDL	SLT
$\alpha_{\text{in}}=0.4$	−21.1	−21.4
$\alpha_{\text{in}}=0.5$	−24.3	−25.4
$\alpha_{\text{in}}=0.6$	−27.0	−28.1
$\alpha_{\text{in}}=0.8$	−28.9	−30.6
$\alpha_{\text{in}}=1.25$	−42.9	−43.7
$\alpha_{\text{in}}=1.7$	−54.2	−55.3
$\alpha_{\text{in}}=2.0$	−58.5	−61.3
$\alpha_{\text{in}}=2.5$	−63.1	−64.9

误差值由式(12)计算[9]得到

$$e=10\log_{10}\left(\frac{\frac{1}{N_S}\sum_{i=1}^{N_S}\|\boldsymbol{s}_i-\hat{\boldsymbol{s}}_i\|^2}{\frac{1}{N_S}\sum_{t=1}^{N_S}\|\boldsymbol{s}_i\|^2}\right) \tag{12}$$

式中:N_S 为 STRAIGHT 谱中包含的列数;$\boldsymbol{s}_i$ 和 $\hat{\boldsymbol{s}}_i$ 分别为原始 STRAIGHT 谱和插值并还原后 STRAIGHT 谱的第 i 列数据。

通过实验仿真结果可看出,即使初始插值比例达到 0.4 和 2.5 时,STRAIGHT 谱的插值误差仍较小,这也说明本文所采用的插值处理方法的可行性。经过时间对准的源语音和目标语音 STRAIGHT 谱表示为 $\boldsymbol{S}_x'$ 和 $\boldsymbol{S}_y'$。

2.2　时频基提取

具体提取流程如图 2 中②所示。首先对源语音 $\boldsymbol{S}_x'$ 进行 CNMF 处理,进而得到源语音 STRAIGHT 谱的时频基$\{\boldsymbol{W}_x(r)\}$和编码矩阵 $\boldsymbol{H}_x$。由于 CNMF 的结果存在不唯一性问题,因此如果采用同样方法对目标说话人 STRAIGHT 谱进行分析,则很难保证得到的目标说话人编码矩阵与 $\boldsymbol{H}_x$ 相同。而根据前面的假设,编码矩阵 $\boldsymbol{H}_x$ 主要是由语音内容信息决定的,而 $\boldsymbol{S}_x'$ 和 $\boldsymbol{S}_y'$ 为经过时间对准的平行语音声道谱,因此对 $\boldsymbol{S}_y'$ 分解得到的编码矩阵应与 $\boldsymbol{H}_x$ 相同。

因此,为了保证分解结果的一致性,在对 $\boldsymbol{S}_y'$ 进行 CNMF 分析时,采用固定其编码矩阵为 $\boldsymbol{H}_x$ 的方法,即令

$$\boldsymbol{H}_y=\boldsymbol{H}_x \tag{13}$$

由此得到相应的目标说话人语音时频基$\{\boldsymbol{W}_y(r)\}$。

2.3　基于时频基替换的声道谱转换

基于训练阶段得到的时频基实现声道谱转换的流程如图 2 中③部分所示。首先提取源说话人语音信号中的 STRAIGHT 谱,之后对该谱在固定编码矩阵为$\{\boldsymbol{W}_x(r)\}$的前提下进行 CNMF 分析,从而得到相应的编码矩阵 $\boldsymbol{H}_x^{\text{convert}}$。之后基于 $\boldsymbol{H}_x^{\text{convert}}$ 和训练阶段得到的目标说话人时频基$\{\boldsymbol{W}_y(r)\}$,通过式(5)得到转换后的 STRAIGHT 谱,即

$$\boldsymbol{S}_y^{\text{convert}}=\sum_{r=0}^{R-1}\boldsymbol{W}_y(r)\cdot\overrightarrow{\boldsymbol{H}_x^{\text{convert}}} \tag{14}$$

3　仿真实验

3.1　卷积非负矩阵分解结果的实验分析

在 1.2 节中已经给出了对 CNMF 分解结果的假设,即时频基中承载了说话人特征信息并有效保存了语音帧间相关性,编码矩阵则主要承载了与语义相关的信息。基于以上假设,对于声道谱参数使用 CNMF 进行分析,之后在编码矩阵固定的情况下,使用目标说话人的时频基矩阵替换源说话人的时频基矩阵,则可实现声道谱中个人特征信息的转换。

为了验证时频基矩阵中是否包含说话人的个人特征信息,可通过如下方法进行初步检验:对说话人 A 的声道谱参数矩阵通过 CNMF 方法进行分析,得到相应的时频基$\{\boldsymbol{W}_A(r)\}$和编码矩阵 $\boldsymbol{H}_A$。之后在固定时频基为$\{\boldsymbol{W}_A(r)\}$的情况下,分别对说话人 A 其他语音数据的声道谱参数以及说话人 B 语音的声道谱参数进行 CNMF 分析,观测重构误差。如果 $\boldsymbol{W}_A$ 中包含个人特征信息,则对说话人 A 语音声道谱分析的误差应该远小于对说话人 B 语音声道谱分析的误差。

在实验中,设定时频基的个数 $R=40$,各时频基中列向量个数 $t=12$。首先取 ARCTIC 语音库中 BDL 子库和 SLT 子库中的前 10 句语音,分别计算其 STRAIGHT 谱,STRAIGHT 谱中各列向量的维度为 256。通过 CNMF 得到时频基$\{\boldsymbol{W}_{\text{BDL}}(r)\}$和$\{\boldsymbol{W}_{\text{SLT}}(r)\}$。然后取 BDL 子库和 SLT 子库中第 11～15 句语音数据,计算 STRAIGHT 谱后,分别计算固定时频基为$\{\boldsymbol{W}_{\text{BDL}}(r)\}$和$\{\boldsymbol{W}_{\text{SLT}}(r)\}$时 CNMF 的重构误差,重构误差采用式(12)进行计算。具体误差值如表 2 所示。

表 2 重构误差分析

语音库子库	时频基为 $\boldsymbol{W}_{\mathrm{BDL}}$ 时的重构误差	时频基为 $\boldsymbol{W}_{\mathrm{SLT}}$ 时的重构误差
BDL	−18.22	−15.65
SLT	−13.03	−19.01

通过重构误差结果对比可知，对于 BDL 和 SLT 中的语音声道谱，采用与说话人语音相匹配的时频基进行 CNMF 分析时的重构误差要明显小于采用其他时频基分析时的重构误差。这说明 CNMF 得到的时频基中的确包含了与说话人相关的个人特征信息。

3.2 卷积非负矩阵分解参数

对于 CNMF 来说，有两个参数会影响最终的转换效果，即时频基的个数 R 和时频基中列向量的个数 t。显然 R 和 t 的增加，会使非负矩阵分解中可变参数的个数增加，有助于减小非负矩阵分解的重构误差。但对于最终的转换效果，这两个参数并不一定越大越好。例如，当 t 的值过大时，各个时频基将可能会覆盖更多的音素，从而导致时频基中承载了部分语义信息。将其应用于其他语音信号的分析时反而会造成语音质量的下降。对这两个参数的选择，本文通过尝试穷举两参数不同组合情况下的语音转换，并通过主观评价选出最优转换语音效果时的参数值。进行主观评价时，采用具有一定语音评价专业背景知识的 5 人作为评价者，评价主要分语音质量评价和转换相似度评价两方面。对语音质量采用 MOS 分方式评价，而对于转换语音与目标语音相似度，则采用类似 MOS 分的评价机制，给出 0～5 分的相似度分数，0 分表示完全不相似，5 分表示完全相似。

在实验中，采用 BDL 和 SLT 语音库中的语音作为训练和测试数据。其中训练数据从第 1 句开始选取，尝试了不同训练数据量下，总量不超过 50 句的情况，而测试数据则选取第 51～55 句。转换前对训练数据和测试数据都进行时间对准处理，尝试了 BDL 到 SLT 和 SLT 到 BDL 两种转换情况。声道谱的转换流程如图 2 所示。生成转换语音时，为去除基音周期等参数转换不准确对转换效果评价的影响，在转换后语音的合成中，直接使用了目标说话人的基音周期。

具体来说，参数设定如下：(1) 训练数据量分别取[10 20 30 40 50]，单位：句；(2) 时频基的个数 R 分别取[20 30 40 50 60]；(3) 时频基中列向量个数 t 分别取[4 8 12 16 20]。

通过实验及主观评价结果发现，在不同训练数据条件下，$R=40$，$t=12$ 时，转换语音在语音质量和相似度上基本都是最优。因此，在后续实验中设定 $R=40$，$t=12$。

在 ARCTIC 语音库中，音素的种类共有 40 种，与选定的时频基的个数正好吻合。每种音素代表了一种不同的声音基本单元，而不同的时频基也是从语音声道谱中提取的不同特征单元。虽然在 R 的选择上，仅尝试了 5 个不同值，并不能说明 40 即为最优结果，但其在一定程度上反映了时频基与音素间的内在相似性。此外，在 $t=12$ 时，每个时频基持续的时长为 60 ms，其长度也与音素长度大体相当。这也在一定程度上反映了两者的相似性。

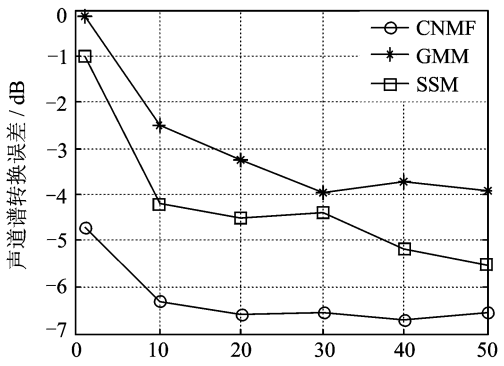
3.3 声道谱转换性能的实验分析

由于 STRAIGHT 模型分解得到的 STRAIGHT 谱中各个列向量的维数为 256，并不适合 GMM 和 SSM 的建模。因此，在进行 GMM 和 SSM 模型训练前，首先需要将 STRAIGHT 谱的列向量变换为维数较小的线谱频率(Line spectrum frequencies，LSF)参数。实验发现当转换后 LSF 的阶数达到 16 阶以上时，转换过程对语音质量的影响即可忽略。因此，在后续实验中 LSF 的阶数选为 16。

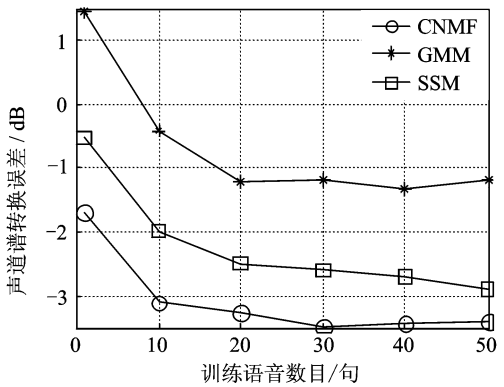
在 GMM 的训练中，将源说话人与目标说话人对应帧的 LSF 参数联合构成 28 维向量后，再进行 GMM 训练，而 GMM 中高斯分量的个数选为 16[9]。模型的训练采用最大期望(Expectation maximization，EM)算法实现[18]。而对基于 SSM 的声道谱转换，参照文献[9]结论，其阶数选为 9，训练算法采用 EM 算法实现。实验中为 ARCTIC 库中 BDL 和 SLT 子库的语音，其中前 50 句语音用于训练，51～55 句用于测试。实验前首先对语音数据进行时间对准处理。

对于客观评价，采用了式(12)对比转换后语音声道谱与目标语音声道谱的差异。为使误差计算标准统一，计算误差前首先将 GMM 和 SSM 的 LSF 转换结果变换回 STRAIGHT 谱参数。图 3 给出了 3 种方法转换结果的误差情况。

通过客观评价结果可知，在 BDL 到 SLT 和 SLT 到 BDL 声道谱转换中，3 种方法的转换误差随着训练数据量的增加都进一步减小，在 30 句左右时基本达到稳定。ARCTIC 语音库中语音数据的长度基本上为 3 s 左右，因此可认为训练数据长度在 100 s 左右时训练效果基本达到稳定。



(a) BDL(男声)到SLT(女声)的转换



(b) SLT(女声)到BDL(男声)的转换

图 3　3 种不同声道谱转化方法的客观评价结果

此外,可以看出这 3 种方法中基于 GMM 的转换方法效果最差。虽然随着训练数据量的增加,误差进一步减小,但与其他方法相比,始终存在 1～2 dB的差距。基于 SSM 转换方法的性能处于 GMM 方法和 CNMF 方法之间,而基于 CNMF 的方法转换效果最好。此外,对比 BDL 到 SLT 和 SLT 到 BDL 的转换效果可知,SLT 到 BDL 的转换误差要略小于 BDL 到 SLT 的转换误差。上述客观评价结果证明了本文方法的有效性。

为进一步对比转换效果,按照 2.2 节中的方法基于声道谱转换结果合成转换后的语音,并对转换语音的语音质量以及与目标语音相似度进行主观评价。主观评价时,评价方法依照 2.2 节内容进行。表 3 给出了 3 种方法转换结果的主观评价结果。

表 3　3 种转换方法的主观评价结果

声道谱转换	转换方法	MOS 分均值（质量）	类 MOS 分均值（相似度）
BDL→SLT	CNMF	3.41	3.58
	GMM	2.84	2.52
	SSM	3.12	2.98
SLT→BDL	CNMF	3.34	3.49
	GMM	2.72	2.46
	SSM	2.92	3.12

从主观评价结果来看,基于 CNMF 的转换方法在转换语音质量和与目标语音的相似度上都优于其他两种方法。通过主观评测结果可发现,BDL 到 SLT 的转换效果基本与 SLT 到 BDL 的转换效果一致,这与客观评价结果存在出入。这说明客观评价结果虽然与主观评价结果大体一致,但两者还是存在一定的误差,此时应以主观评价为准。CNMF 的语音转换方法转换效果的提升,主要应归结于基于 CNMF 的语音转换方法实现对语音帧间相关性更好的保存和转换。

4　结束语

本文提出了一种基于 CNMF 的语音转换方法。由于 CNMF 分解得到的时频基可承载语音个人特征信息的同时,较好地保存语音帧间相关性。利用 CNMF 的这一特点,本文通过时频基替换实现从源说话人到目标说话人的语音转换,且在具体的转换方法设计中,克服了 CNMF 分解结果不唯一的问题。实验仿真结果表明,本文提出的方法在转换语音的质量及与目标语音的相似度上都优于基于 GMM 和 SSM 的声道频谱转换方法。

虽然本文所述方法给出了一种语音转换的新思路,且得到了较好的实验仿真结果,但基于 CNMF 的转换方法仍存在以下不足:(1) 其转换效果与模型中部分参数值的选取,如时频基个数、时频基中列向量个数,甚至与训练数据量都有较大关系。本文中这些参数的确定还依赖于实验结果,具有较大的随意性,缺乏更为有效的选取方法。(2) 本质上,基于 CNMF 的方法仍旧是线性分解方法,因而可能对语音信号中一些非线性特性表征不足,从而在一定程度上制约了转换效果的提升。此外,CNMF 仅仅将语音信号分为两部分,因此还无法对包含情感信息的语音信号进行有效分析。

参考文献:

[1] Stylianou Y. Voice transformation: a survey [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. China: IEEE, 2009: 3585-3588.

[2] Abe M, Nakamura S, Shikano K, et al. Voice conversion through vector quantization [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Seattle, Washington: IEEE, 1988: 655-658.

[3] Stylianou Y, Cappe O, Moulines E. Continuous probabilistic transform for voice conversion [J].

IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131-142.

[4] 岳振军，邹翔，王浩. 基于隐马尔可夫模型和高斯混合模型结合的声音转换方法[J]. 数据采集与处理，2009,24(3): 285-289.
Yue Zhenjun, Zou Xiang, Wang Hao. Voice conversion with the combination of HMM and GMM[J]. Journal of Data Acquisition and Processing, 2009, 24(3): 285-289.

[5] Yamagishi J, Kobayashi T, Nakano Y, et al. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm [J]. IEEE Transactions on Audio, Speech and Language Processing, 2009, 17(1): 66-83.

[6] 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.

[7] 秦勇，双志伟，张世磊. 语音转换分析及相似度改进[J]. 清华大学学报:自然科学版, 2009, 49(S1): 1408-1412.
Qin Yong, Shuang Zhiwei, Zhang Shilei. Warped source spectrum for voice conversion and similarity [J]. Journal of Tsinghua University: Science and Technology Edition, 2009, 49(S1): 1408-1412.

[8] Desai S, Black A W, Yegnanarayana B, et al. Spectral mapping using artificial neural networks for voice conversion [J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 18(5): 954-964.

[9] Duxans H, Bonafonte A, Kain A, et al. Including dynamic and phonetic information in voice conversion systems [C]//8th International Conference on Spoken Language Processing. Jeju Island, Korea: [s. n.], 2004: 5-8.

[10] Toda T, Black A W, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(8): 2222-2235.

[11] Zen H, Nankaku Y, Tokuda K. Continuous stochastic feature mapping based on trajectory HMMs [J]. IEEE Transactions on Audio, Speech and Language Processing, 2011, 19(2): 417-430.

[12] Xu Ning, Yang Zhen, Zhang Linghua, et al. Voice conversion based on state-space model for modelling spectral trajectory [J]. Electronics Letters, 2009, 45(14): 763-764.

[13] 徐宁，杨震，张玲华. 基于状态空间模型的子频带语音转换算法 [J]. 电子学报, 2010, 38(3): 646-653.
Xu Ning, Yang Zhen, Zhang Linghua. Sub-band voice morphing algorithm based on state-space model [J]. Acta Electronica Sinica, 2010, 38(3): 646-653.

[14] Paris S. Convolutive speech bases and their application to supervised speech separation [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(1): 1-12.

[15] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401(6755): 788-791.

[16] Lee D D, Seung H S. Algorithms for nonnegative matrix factorization [C]//Advances in Neural Information Processing Systems 13. Cambridge, Mass, USA: MIT Press, 2001: 556-562.

[17] Kawahara H, Masuda-Katsuse I, Cheveign A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds [J]. Speech Communication, 1999, 27(3/4): 187-207.

[18] Lee K S. Statistical approach for voice personality transformation [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(2): 641-651.

作者简介：孙健(1984-)，男，博士研究生，研究方向：多媒体信号处理，E-mail: sunjian001 @ gmail. com；张雄伟(1965-)，男，教授，研究方向：多媒体信号处理、智能计算；曹铁勇(1971-)，男，教授，研究方向：多媒体信号处理；杨吉斌(1978-)，男，讲师，研究方向：数字通信、多媒体信号处理；孙新建(1980-)，男，博士研究生，研究方向：多媒体信号处理。