引用格式：曾歆, 张雄伟, 孙蒙, 等. 基于 GMM 模型和 LPC-MFCC 联合特征的声道谱转换研究[J]. 声学技术, 2020, **39**(4): 451-455. [ZENG Xin, ZHANG Xiongwei, SUN Meng, et al. Research on vocal tract spectrum conversion based on GMM model and LPC-MFCC[J]. Technical Acoustics, **39**(4): 451-455.] **DOI**: 10.16300/j.cnki.1000-3630.2020.04.012

基于 GMM 模型和 LPC-MFCC 联合特征的声道谱转换研究

曾 歆，张雄伟，孙 蒙，苗晓孔，姚 琨
(陆军工程大学，江苏南京 210007)

摘要：声道谱转换是语音转换中的关键技术。目前，大多数语音转换方法对声道谱的转换都是先提取语音中的某一种声道特征参数，然后对其进行训练转换，进而合成转换语音。由于不同的声道特征参数表征着不同的物理和声学意义，因此这些方法通常忽略了不同声道特征参数之间可能存在的互补性。针对这一问题，研究了不同声道特征参数之间进行联合建模的方法，引入了一种由线性预测系数(Linear Prediction Coefficient, LPC)和梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)联合构成的 LPC-MFCC 特征参数，提出了一种基于高斯混合模型(Gaussian Mixture Model, GMM)和 LPC-MFCC 联合特征参数的语音转换方法。为验证文中方法的有效性，仿真实验选取了基于 GMM 和 LPC 的语音转换方法进行对比，对多组实验数据进行主观和客观测试，结果表明，文中提出的语音转换方法可以获得相似度更高的转换语音。

关键词：语音转换；声道谱转换；高斯混合模型；联合建模；线性预测系数-梅尔频率倒谱系数
中图分类号：TN912.3 **文献标识码：**A **文章编号：**1000-3630(2020)-04-0451-05

Research on vocal tract spectrum conversion based on GMM model and LPC-MFCC

ZENG Xin, ZHANG Xiongwei, SUN Meng, MIAO Xiaokong, YAO Kun
(*Army Engineering University*, *Nanjing* 210007, *Jiangsu*, *China*)

Abstract: Spectrum conversion is a key technique in voice conversion. At present, most of vocal tract spectrum conversion methods are first to extract one of characteristic parameters of the vocal tract then to train and convert it, and finally to synthesize the converted voice. Since different characteristic parameters of the vocal tract characterize different physical and acoustic meanings, these methods usually ignore the possible complementary effects between different characteristic parameters. To solve this problem, this paper studies the joint modeling method between different characteristic parameters of vocal tract, and introduces a new characteristic parameter called LPC-MFCC which is composed of Linear Prediction Coefficient (LPC) and Mel-Frequency Cepstral Coefficient (MFCC). And then, a voice conversion method based on Gaussian Mixture Model (GMM) with LPC-MFCC is proposed. In order to verify the effectiveness of the proposed method, the voice conversion method based on GMM with LPC parameter is selected for comparison in simulation experiments. Subjective and objective tests are conducted with multiple sets of experimental data, and the results show that the proposed voice conversion method can achieve a higher similarity of voice conversion.
Key words: voice conversion; vocal tract spectrum conversion; Gaussian Mixture Model (GMM); joint modeling; Linear Prediction Coefficient-Mel-Frequency Cepstral Coefficient (LPC-MFCC)

0 引 言

语音转换是一种在保留语义信息不变的前提下，修改源说话人的个性特征信息，使之具有目标说话人个性特征的语音处理技术[1]。语音转换要实现这一目的，就要提取表征个性特征信息的声学特征，建立不同说话人对应声学特征的对应关系，即转换规则，然后进行转换合成，得到转换语音。

语音转换是目前信号处理领域比较新的一个分支，该技术的研究兼具理论意义和实际应用价值。在多媒体娱乐方面，可通过语音转换实现特定人物配音；对于语音登入系统，可以利用转换语音

收稿日期：2019-01-04; **修回日期：**2019-02-28
基金项目：国家自然科学基金(61471394)、江苏省优秀青年基金(BK20180080)资助项目。
作者简介：曾歆(1995－)，男，四川泸州人，硕士研究生，研究方向为语音转换技术。
通讯作者：张雄伟，E-mail: 245104441@qq.com

攻击说话人认证系统。此外，还可以利用语音转换来消除个人特征差异对语音识别的影响等。由此可见，语音转换技术值得深入研究。

语音的特征信息大致划分为 3 类：音段信息、超音段信息和语言学信息[2]。相关研究表明，超音段信息中的平均基频和音段信息中的声道谱包络对说话人语音个人特征信息的贡献最为重要[3]。相对于平均基频而言，声道谱包络的建模、转换更为复杂，且是制约语音转换效果提升的瓶颈。因此，本文重点围绕声道谱转换展开研究。

语音转换技术研究可追溯到 20 世纪 80 年代。王志卫等[4]采用了基于码书映射的语音转换方法，该方法基于统计得到的直方图信息，通过加权求和的方法实现语音转换。这种“硬聚类”的转换方法虽然效果一般，但开辟了一条从统计学角度解决语音转换的思路。Toda 等[5]采用了基于高斯混合模型(Gaussian Mixture Model, GMM)的声道谱转换方法，对说话人的声道谱空间参数进行建模映射。相比基于码本映射的语音转换方法，该方法极大地提升了频谱平滑度，但基于概率的“软聚类”也导致结果中存在参数过平滑问题。Sundermann 等[6]采用动态频率规整(Dynamic Frequency Warping, DFW)的方法进行语音转换，即对源说话人声道谱频率进行 DFW 处理，使其共振峰位置匹配目标说话人频谱共振峰位置。此外，基于隐变量模型的转换方法[7]、基于深度神经网络模型的转换方法[8-9]等也相继被广泛研究和应用。

在语音转换系统中，声道谱参数是反映说话人个性的重要特征参数。在众多关于声道谱包络的建模转换中，GMM 方法的使用较为普遍。相较于近年流行的神经网络方法，GMM 方法依然具有模型体积小、转换耗时少、可本地化计算等优点。因此，本文考虑选用 GMM 方法进行相应的语音转换研究。

语音信号中包含着丰富的特征参数，不同的特征参数表征着不同的物理和声学意义。特征参数的选择对语音转换系统的转换效果至关重要。目前关于语音转换的研究中，大多数转换方法只选择对单一声道特征参数进行转换，而忽略了不同声道特征参数之间可能存在的互补性。本文在现有研究成果的基础上，对不同的声道特征参数进行联合建模和转换。具体来说，从语音信号中提取线性预测系数(Linear Prediction Coefficient, LPC)和梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)，联合二者得到 LPC-MFCC 特征参数，并借助转换性能较好的 GMM，实现对 LPC-MFCC 特征参数的转换。LPC 是表征声道信息的特征参数，主要反映声道响应；而 MFCC 是基于人听觉的临界带效应，在梅尔标度频率域提取出来的倒谱特征参数，更贴近人耳的听觉特性。因此，LPC 参数和 MFCC 参数存在一定的互补性。

1 相关知识

1.1 GMM 建模

在 GMM 建模阶段，采用了对源和目标联合建模的方法。联合建模一般选用并行语料，即源与目标训练的语料一致，以此来保证动态时间规整(Dynamic Time Warping, DTW)后的联合矢量源与目标的对齐，为 GMM 训练做好准备。转换规则的确立一般选用最小二乘法来估计转换函数的相关参数。与矢量量化语音转换方法相比，GMM 是对频谱包络特征参数进行软分类，使得特征参数能够以一定的概率属于多个不同的类，在一定程度上克服了矢量量化的不连续性，改善了转换后语音的音质。使用该方法进行语音转换能够得到较为满意的合成语音。

1.2 基于 LPC 特征参数的 GMM 声道谱转换方法

1.2.1 线性预测系数

语音线性预测的基本原理是：由于语音信号样点之间存在相关性，因此一个语音的采样值可以用过去若干语音采样值的线性组合来逼近。通过使实际语音信号抽样值和线性预测抽样值之间的误差在均方准则下达到最小值来求解预测系数，而预测系数就反映了语音信号的特征，故可以用这组语音特征参数进行语音转换或语音合成等。

设 n 时刻的语音采样值 $s(n)$ 可由其前面 p 个语音采样值的线性加权表示，则 $s(n)$ 可以表示为

$$s(n)\approx\sum_{i=1}^{p}a_i s(n-i) \tag{1}$$

其中，a_i 表示权值，p 个 LPC 参数可通过全极点模型进行求解。

线性预测最主要的优势在于可以较为精确地估计语音声道参数，能够较好地反映语音信号的声道特性。

1.2.2 基于 LPC 参数和 GMM 模型的语音转换

在训练阶段，首先分别提取源说话人和目标说话人的 LPC 参数；然后使用 DTW 算法对源和目标说话人的 LPC 参数进行时间对齐；最后运用 GMM 训练网络，建立映射转换规则。

在转换阶段，首先提取源说话人的 LPC 参数；然后根据训练阶段建立的映射转换规则，对源说话人的 LPC 参数进行转换；最后利用转换所得到的 LPC 参数合成转换语音。

2　本文方法

2.1　基于 LPC-MFCC 联合特征参数的语音转换

本文在基于 LPC 参数的 GMM 声道谱转换方法的基础上，引入了更贴近于人耳听觉特性的 MFCC 参数，构建了联合特征参数 LPC-MFCC 并用于语音转换。其语音转换框图如图 1 所示，转换步骤如下：

在训练阶段：(1) 分别提取源说话人和目标说话人的 LPC 参数和 MFCC 参数；(2) 联合 LPC 参数和 MFCC 参数，得到新的特征参数 LPC-MFCC；(3) 使用 DTW 算法对源和目标说话人的 LPC-MFCC 特征参数进行时间对齐；(4) 使用 GMM 模型训练网络，建立映射转换规则。

在转换阶段：(1) 提取源说话人的 LPC 参数和 MFCC 参数；(2) 联合 LPC 参数和 MFCC 参数，得到 LPC-MFCC 联合特征参数；(3) 根据训练阶段建立的映射转换规则，对源说话人的 LPC-MFCC 特征参数进行转换，转换所得 LPC-MFCC 特征参数中包含 LPC 参数对应转换的生成部分和 MFCC 参数对应转换的生成部分；(4) 考虑到基于 LPC 参数的语音转换方法的效果优于基于 MFCC 参数的语音转换方法，因此选取 LPC 参数对应转换生成部分进行语音合成，得到转换语音。

2.2　方法步骤

2.2.1　语音信号预处理

为了得到适合转换处理的语音帧，首先对语音进行加窗分帧、端点检测、预加重等前端预处理。其中，预加重的目的是为了对语音的高频部分进行加重，去除口唇辐射的影响，增加语音的高频分辨率。本文设置的预加重系数为 0.9。

2.2.2　MFCC 参数与 LPC 参数的提取

本步骤的目的是基于预处理后的语音帧，提取出反映信号特征的关键特征参数以便于后续处理。考虑到 GMM 模型更适用于低维度特征的建模，本文选取低维度的 MFCC 参数与 LPC 参数进行联合。MFCC 参数的提取过程如图 2 所示[10]。

基于 LPC 特征参数的语音转换在 1.2 节已经详细介绍，此处不再赘述。

2.2.3　LPC 参数与 MFCC 参数的联合

为了便于 LPC 参数和 MFCC 参数进行联合，在 LPC 参数和 MFCC 参数提取之前，对语音信号做同样的加窗分帧等预处理操作。本文实验设定滤波器阶数为 12。

为了便于阐述参数的联合过程，假设矩阵 $\boldsymbol{A}_{\mathrm{lpc}}$ 表示根据某一句语音提取得到的 LPC 参数，阶数为 $M\times N$，其中 M 表示帧数，N 表示特征维度。矩阵 A_{mfcc} 表示根据同一语音提取得到的 MFCC 参数，阶数为 $M\times N$。对两个矩阵按列拼接得到联合矩阵 $\boldsymbol{A}_{\mathrm{lpc\text{-}mfcc}}=\left[\boldsymbol{A}_{\mathrm{lpc}}\quad \boldsymbol{A}_{\mathrm{mfcc}}\right]$，即 LPC-MFCC 特征参数对应的矩阵，阶数大小为 $M\times 2N$。这一步对 LPC 参数的维度进行了扩充，使原本 N 维度的转换问题变成 $2N$ 维度的转换问题，同时也将 LPC 参数和 MFCC 参数之间可能存在的互补性纳入考虑范围。

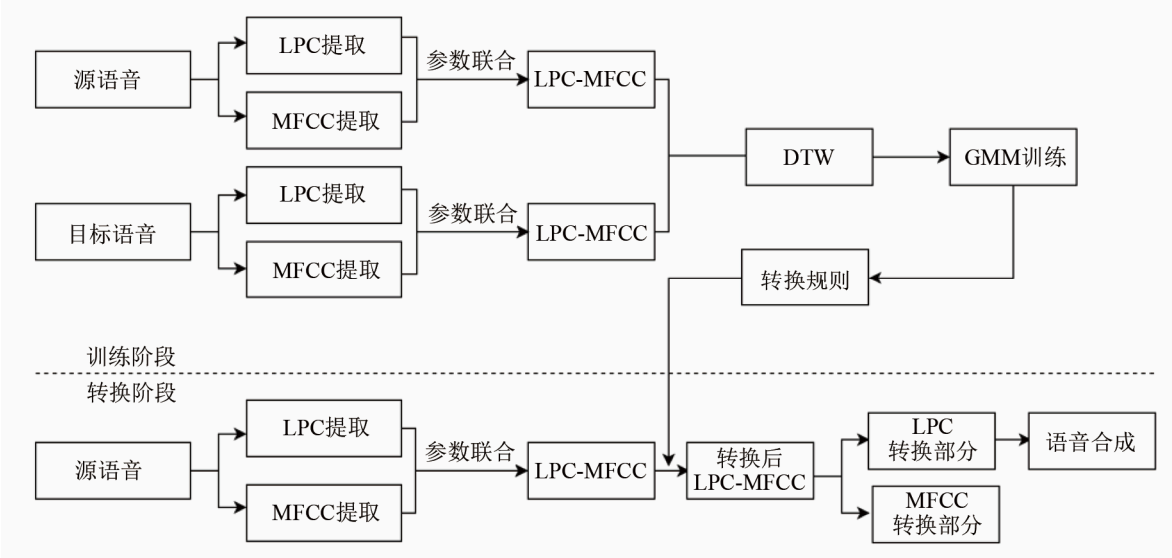

图 1　基于 GMM 模型和 LPC-MFCC 联合特征的转换框图

Fig.1　Block diagram of voice conversion based on GMM model with LPC-MFCC

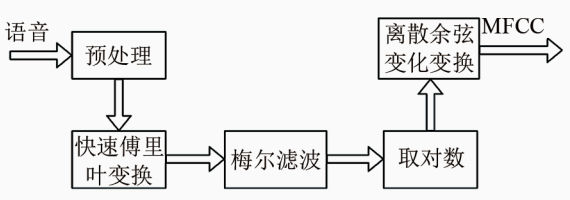

图 2 MFCC 特征提取流程

Fig.2 The procedure of extracting MFCC features

2.2.4 时间对齐

在建立源特征参数和目标特征参数映射关系之前，需要先将源和目标语音的特征参数进行时间对齐，确保转换的是描述同一音节的特征参数。使用 DTW 算法对源说话人和目标说话人的 LPC-MFCC 特征参数进行对齐，产生一对相等长度的源和目标的特征序列。

2.2.5 模型训练及参数转换

将源语音参数矢量 $\boldsymbol{X}$ 与目标语音参数矢量 Y 构成一个联合矢量 $\boldsymbol{Z}$，$\boldsymbol{Z}=[\boldsymbol{X}\ \boldsymbol{Y}]^{\mathrm{T}}$，利用联合概率 $P(\boldsymbol{X},\boldsymbol{Y})$ 来训练高斯混合模型。假设用 p 个单高斯分布的加权求和来表征 $\boldsymbol{Z}$ 的概率分布，则 GMM 的概率分布函数表示为[11]

$$P(\boldsymbol{z})=\sum_{i=1}^{p}\alpha_i N(\boldsymbol{z};\mu_i;\boldsymbol{\Sigma}_i) \tag{2}$$

其中，$N(\boldsymbol{z};\mu_i;\boldsymbol{\Sigma}_i)$ 表示均值为 μ_i、协方差矩阵为 $\boldsymbol{\Sigma}_i$ 的 p 维高斯分布，表达式为

$$N(\boldsymbol{z};\mu_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{\frac{p}{2}}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\exp[-\frac{1}{2}(\boldsymbol{z}-\mu_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{z}-\mu_i)] \tag{3}$$

约束条件为

$$\sum_{i=1}^{p}\alpha_i=1,\ \alpha_i\geqslant 0 \tag{4}$$

GMM 的 3 个模型参数 $(\alpha_i,\mu_i,\boldsymbol{\Sigma}_i)$，可以通过期望最大(Expectation-Maximization, EM)算法进行迭代求取[11]。

首先找到输入语音特征参数相对于源说话人 GMM 模型对应的分量，然后找到输入语音特征参数相对于目标说话人 GMM 模型对应的分量，然后在这两个分量之间建立转换规则，这样就可以将源语音的参数映射成目标语音的参数，从而实现对输入语音特征的转换。

运用上述的 GMM 训练 LPC-MFCC 特征参数，建立映射转换规则。在转换阶段，同样对源目标语音提取 LPC-MFCC 特征参数，根据训练好的网络模型进行转换。在合成阶段，只需取出 LPC 参数对应的转换部分，进行语音合成，从而得到转换语音。

3 实验结果及分析

为了更好地对比语音转换方法的性能，需要进行仿真实验测试。本文采用主观和客观相结合的测试方法来对两种方法的转换性能进行综合评价。

3.1 测试方法

3.1.1 客观测试

语音信号之间的差异一般采用语音信号频谱上的距离测度来描述。理论上可以使用各种类型频谱差测量来计算转换语音和目标语音之间的差异。转换后的频谱和目标频谱之间的距离越小，说明二者越接近，也即转换效果越好。语音转换相关文献中使用最多的客观测试衡量指标是梅尔倒谱失真(Mel Cepstral Distance, MCD)，单位 dB，其计算方法为

$$d_{\mathrm{MCD}}(y,\hat{y})=\frac{10}{\ln 10}\sqrt{2(y,\hat{y})^{\mathrm{T}}(y,\hat{y})} \tag{5}$$

其中，$\boldsymbol{y}$ 和 $\hat{\boldsymbol{y}}$ 分别是目标语音和转换语音的梅尔倒谱特征向量。

3.1.2 主观测试

主观测试也是对转换语音进行评价的一个很重要的方式。它根据一定的评价标准、靠人的主观听觉来对转换后的语音进行判断或打分，进而对语音转换方法的性能进行评估。语音转换相关文献中使用最多的主观测试衡量指标是平均意见得分(Mean Opinion Score, MOS)测试。MOS 测试的主要原理是让测评人根据 5 个等级划分对测试语音的主观感受进行打分。它既可以用于对语音自然度进行主观评价，也可以用于对说话人特征相似度的评价。测试要求测评人具有正常的听觉感知能力，并多年从事语音技术研究。

3.2 测试结果

本文使用由中国科学院自动化所(Institute of Automation, Chinese Academy of Sciences, CASIA)发布的 CASIA 汉语情感语料库进行了多组转换实验，包括：男声到男声(M-M)、男声到女声(M-F)、女声到男声(F-M)、女声到女声(F-F)的转换。客观测试结果如表 1 所示。其中优化比率表示联合特征参数方法相对于 LPC 参数方法的 MCD 的下降率。

结合表 1 分析可知，相比于基于 GMM 和 LPC 参数的语音转换方法，基于 GMM 和 LPC-MFCC 联合特征参数的语音转换方法，在男声到男声、男声到女声转换时，客观指标 MCD 值有较明显的下

降；但是当源目标语音是女声，目标语音是女声或者男声时，两种语音转换方法的 MCD 测试结果相差不大。可能的原因是女声音调高，将其作为待转换语音会影响转换效果。今后将会对其具体原因进行更深入的研究。

总体来说，基于联合特征参数的转换方法相比于基于 LPC 特征参数的转换方法，MCD 值明显降低，降低比率为 11%，客观测试结果更佳。

表 1　客观测试的 MCD 结果比较
Table 1　Comparison of MCD results in objective test

特征	MCD/dB					优化比率/%
	F-M	F-F	M-M	M-F	平均	
LPC 特征	8.4	7.5	8.3	11.5	8.9	11
联合特征	8.2	8.0	6.5	8.8	7.9	

在主观测试方面，依据转换语音和目标语音相似度的主观测试结果如表 2 所示。其中优化比率表示联合特征参数方法相对于 LPC 特征参数方法的 MOS 分提升率。

表 2　主观测试的 MOS 结果比较
Table 2　Comparison of MOS results in subjective test

特征	MOS					优化比率/%
	F-M	F-F	M-M	M-F	平均	
LPC 特征	2.41	2.52	2.57	2.46	2.49	25
联合特征	2.76	3.28	3.56	2.88	3.12	

结合表 2 分析可知，相比于基于 LPC 参数的转换方法，基于 LPC-MFCC 联合特征参数的转换方法，在男声到男声、女声到女声两组实验中的相似度显著提高；在男声到女声、女声到男声两组实验中略有提高。

总体来说，基于联合特征参数的转换方法，相比于基于 LPC 特征参数的转换方法，转换语音和目标语音更相似，相似度提升达到 25%，转换性能更佳。

4　结　论

本文在基于 GMM 模型和 LPC 参数语音转换方法的基础上，引入了更贴近人耳听觉特性的 MFCC 参数，将 LPC 和 MFCC 参数之间可能存在的互补性纳入考虑范围，在此基础上提出了一种基于 GMM 模型和 LPC-MFCC 联合特征参数的语音转换方法。主观和客观实验表明，相比于基于 GMM 模型和 LPC 参数的语音转换方法，基于 GMM 模型和 LPC-MFCC 联合特征参数的语音转换方法相似度更高，转换效果更佳。但 MFCC 参数的引入同时也会对 LPC 的合成阶段产生干扰，导致合成语音存在些许噪声。如何解决这一问题将是下一步工作的重点。此外，本文语音转换系统的输入和输出都是 LPC-MFCC，且合成阶段只选用 LPC 对应的转换部分进行语音合成。下一步拟继续研究以 LPC-MFCC 为输入，LPC 或 MFCC 为输出的语音转换方法，并且在语音合成阶段拟将 MFCC 纳入考虑范围，继续探究 LPC 和 MFCC 参数之间的互补性，以进一步提高转换语音的自然度和相似度。

参　考　文　献

[1] HELANDER E, VIRTANEN T, NURMINEN J, et al. Voice conversion using partial least squares regression[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, **18**(5): 912-921.

[2] 解伟超. 语音转换中声道谱参数和基频变换算法的研究[D]. 南京: 南京邮电大学, 2013.

[3] 周莹. 高质量语音转换系统中关键技术的研究[D]. 南京: 南京邮电大学, 2012.

[4] 王志卫，徐宁，刘小峰. 一种基于码书映射的高效语音转换方法[J]. 微处理机, 2014, **35**(1): 65-69.
WANG Zhiwei, XU Ning, LIU Xiaofeng. A highly efficient voice conversion method based on codebook mapping[J]. Microprocessors, 2014, **35**(1): 65-69.

[5] TODA T, BLACK A W, TOKUDA K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory[J]. IEEE Transactions on Audio, Speech & Language Processing, 2007, **15**(8): 2222-2235.

[6] SUNDERMANN D, NEY H. VTLN-based voice conversion[C]// ISSPIT 2003, Proceedings of the 3rd IEEE International Symposium on IEEE, 2003: 556-559.

[7] QIAO Y, SAITO D, MINEMATSU N. HMM-based sequence-to-frame mapping for voice conversion[C]//Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on IEEE, 2010: 4830-4833.

[8] MOHAMMADI S H, KAIN A. Voice conversion using deep neural networks with speaker-independent pre-training[C]//Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014: 19-23.

[9] SUN L, KANG S, LI K, et al. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015: 4869-4873.

[10] 单燕燕. 基于 LPC 和 MFCC 得分融合的说话人辨认[J]. 计算机技术与发展, 2016, **26**(1): 39-42.
SHAN Yanyan. Speaker identification based on score combination of LPC and MFCC[J]. Computer Technology and Development, 2016, **26**(1): 39-42.

[11] SAITO D, DOI H, MINEMATSU N, et al. Voice conversion based on matrix variate Gaussian mixture model[C]//Signal Processing (ICSP), 2014 12th International Conference on. IEEE, 2014: 567-571.