**不同性别语音转换的调研和实践**

院系：国家网络安全学院

作者：2019300003041 李潇楠

指导教师：蔡波

摘 要

本次技术报告的目的是运行不同的开源代码，将男女主播的声音进行风格转换，改变性别，并分析其中的技术原理。

关键词：语音 神经网络

**目录**

[1 研究背景 3](#_Toc20120)

[2 多音素的多对一非并行数据语音风格转换 4](#_Toc11707)

[2.1 研究亮点 4](#_Toc19404)

[2.2 关键技术解读 5](#_Toc1922)

[2.3 实验效果 11](#_Toc17444)

[2.3.1客观方法评估 12](#_Toc18788)

[2.3.2 主观方法评估 12](#_Toc13322)

[2.4 代码的模型构建 13](#_Toc9369)

[3 参考资料 14](#_Toc8787)

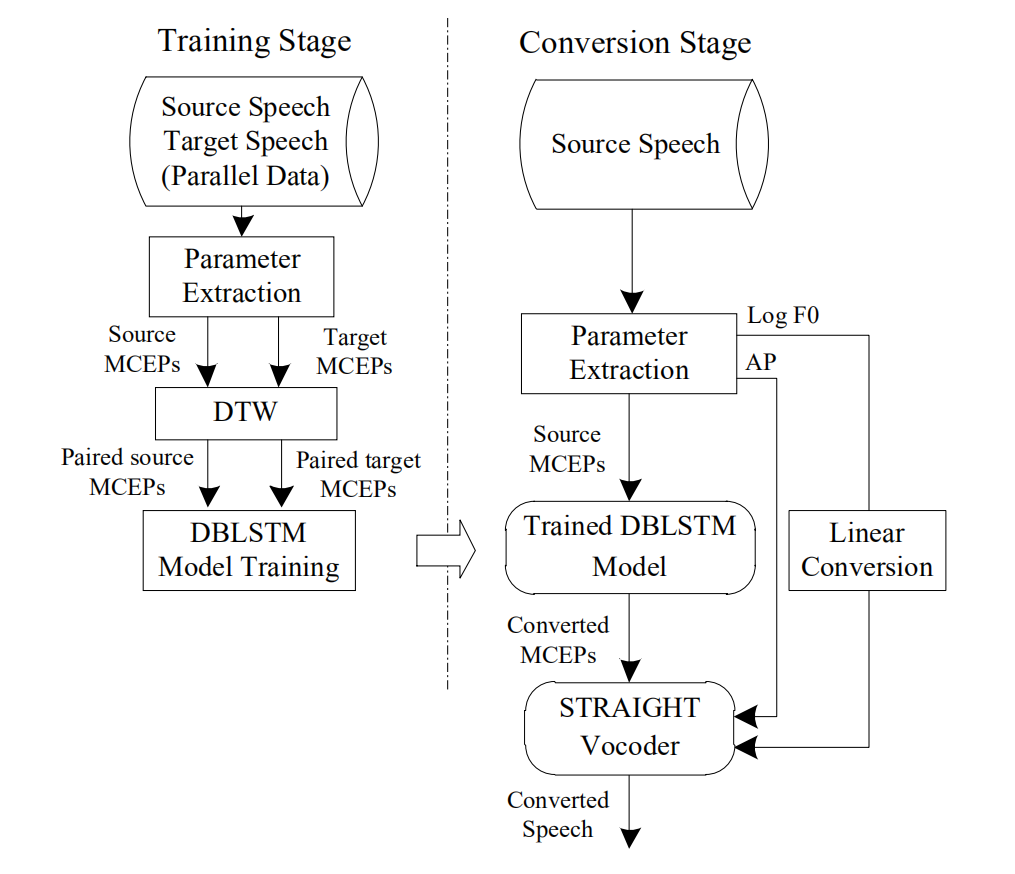
**1 研究背景**

语音特色转换，简称语音转换，是将源说话人的声音变为目标说话人的特色声音。随着智能语音技术借助大数据和人工智能的契机在医疗领域的应用愈加广泛，语音技术已经从一开始的语音识别技术逐步发展到智能语音辅助阶段，并且在医学领域取得了许多有意义的成果，例如医用陪护机器人和辅助治疗机器人等。在对于自闭症患儿的治疗中，个性化的语音对于自闭症的早期干预行之有效，因此，语音风格转移技术可以用于陪护机器人和辅助治疗机器人的个性化语音生成场景。

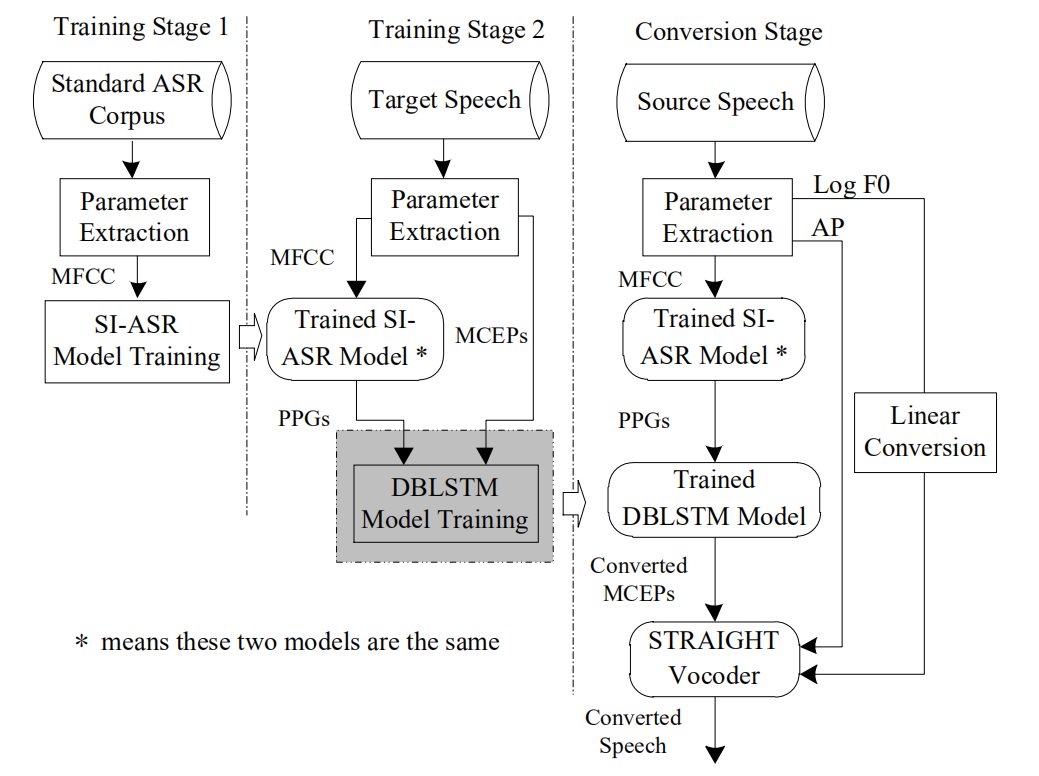
**2 多音素的多对一非并行数据语音风格转换**

**2.1 研究亮点**

对于下面的实验流程，忽略中间步骤，只关注输入和输出，进行初步分析。



并行数据（Baseline）：两阶段——第一阶段将原语音和目标语音一起作为DBLSTM模型的输入进行训练（并行）；第二阶段输入源语音进行风格的转换。



非并行：三阶段——第一阶段输入波形-音素的数据对，训练SI-ASR模型；第二阶段输入目标语音，训练DBLSTM模型（在代码里，这里是两个CBHG，将PPGs转换为梅尔频谱再转换为原始波形）；第三阶段将源语音转换为目标语音

因此我们可以看到，这个研究首先摆脱了数据集的难题，收集源语音和目标语音的大量并行数据是一项困难的工作，而非并行的方法只需要目标语音的数据，和任意人的波形-音素的数据作为模型的输入来训练，最后再输入源语音进行转换。这是研究的第一个亮点。

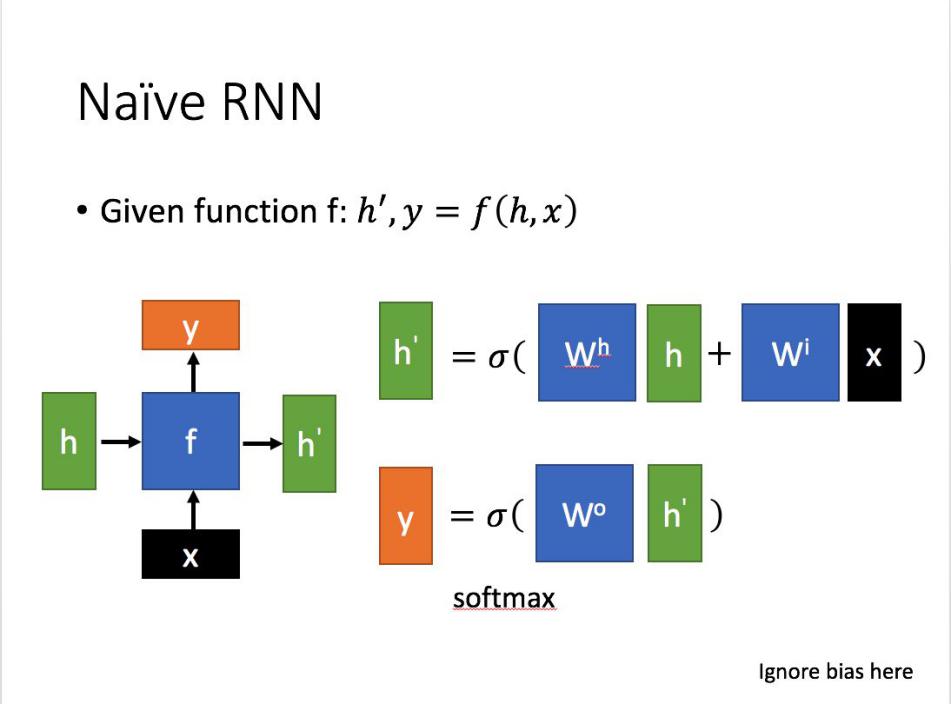
其二，并行数据的转换方法，每次训练模型都是针对特定的源语音和目标语音，是一对一的；而论文的新方法在训练之后，可以让任意语音转换成目标语音的风格，是多对一的。

其三，这种方法相较并行数据的转换方法，提高了语音的自然度，也提高了源语音和目标语音的相似性。

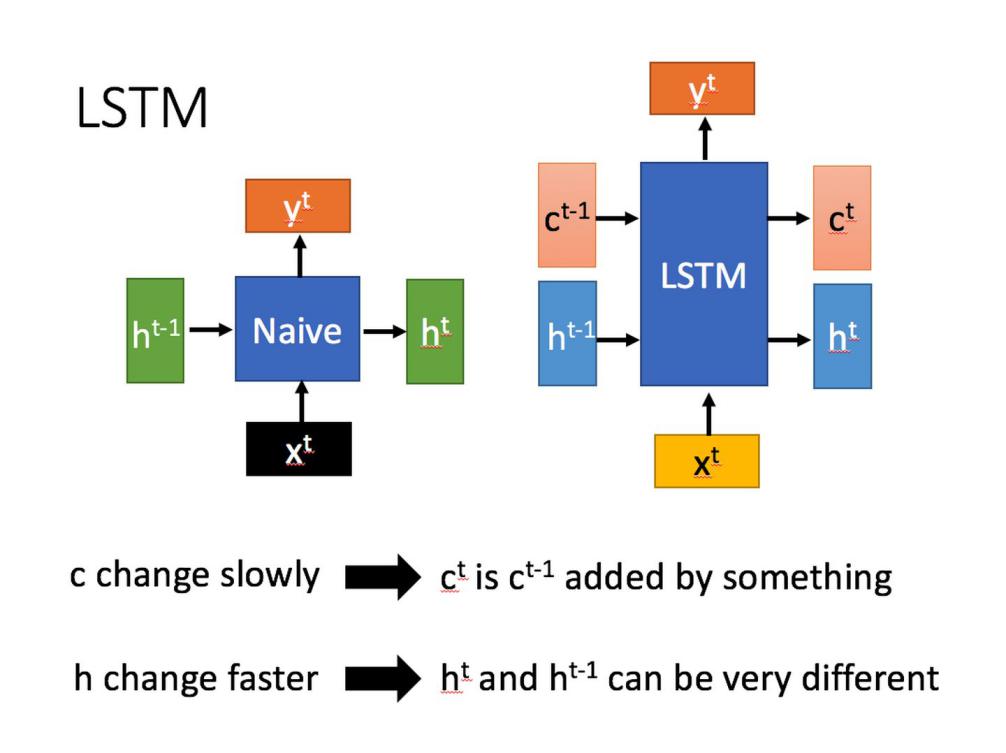
**2.2 关键技术解读**

**2.2.1 DBLSTM**

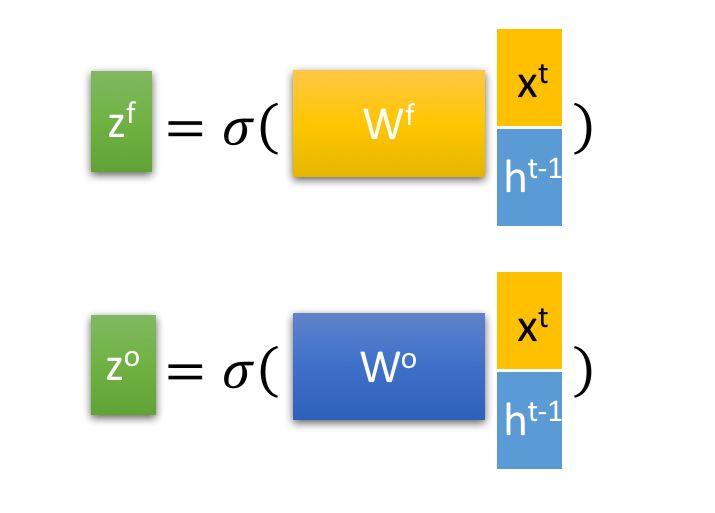
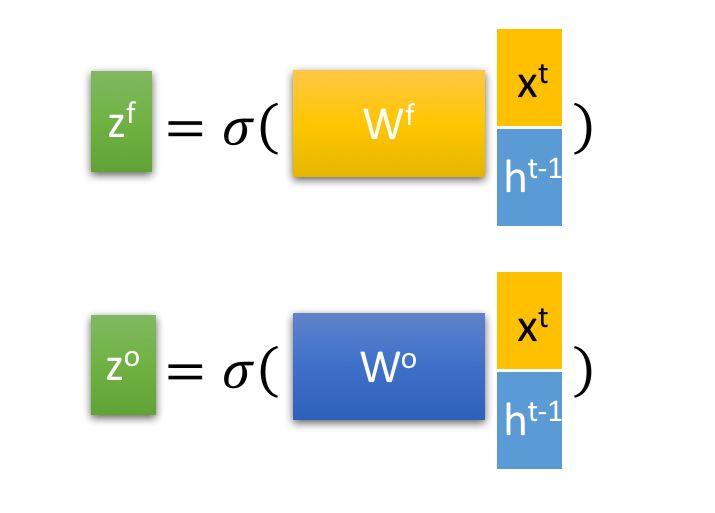
DBLSTM是贯穿并行和非并行方法的一个关键模型，首先需要了解模型的前身RNN和LSTM。

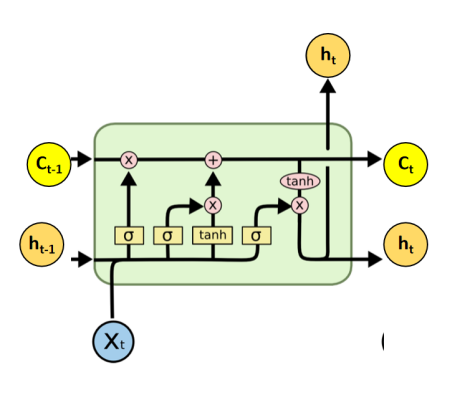
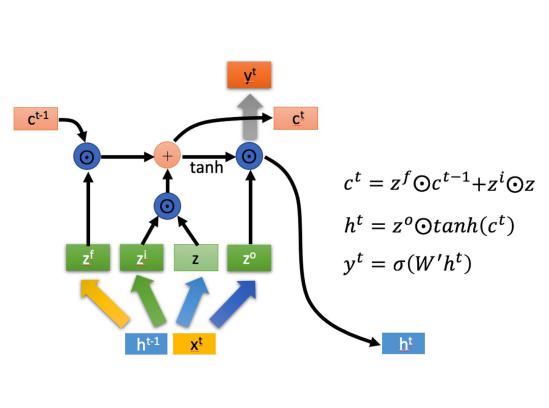


左边的是一个循环神经网络RNN（Recurrent Neural Network），IMG_256xt为当前状态下数据的输入， IMG_257 ht-1表示接收到的上一个节点的输入。yt为当前节点状态下的输出，而ht（hidden state）为传递到下一个节点的输出（yt与ht有关，可以用ht进行线性映射和分类得到）。这个模型很好的利用到了上文信息，但是容易产生梯度消失/爆炸现象。

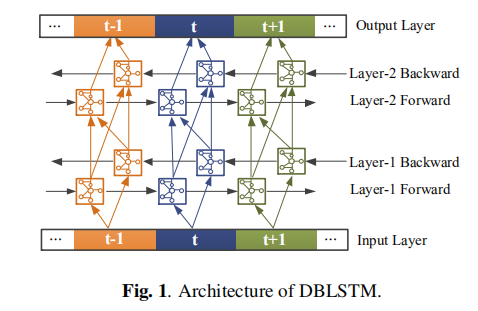


LSTM引入了一个新的状态（cell state）ct，相比于ht，由于它的值改变较慢，所以能让模型实现长期记忆的效果，而不仅仅只是专注到范围较近的上文。





上标f,i,o分别代表forget,information,output，也就是忘记、选择记忆和输出，通过精细化的控制状态转移来训练，比起RNN更加复杂，效果也更好。

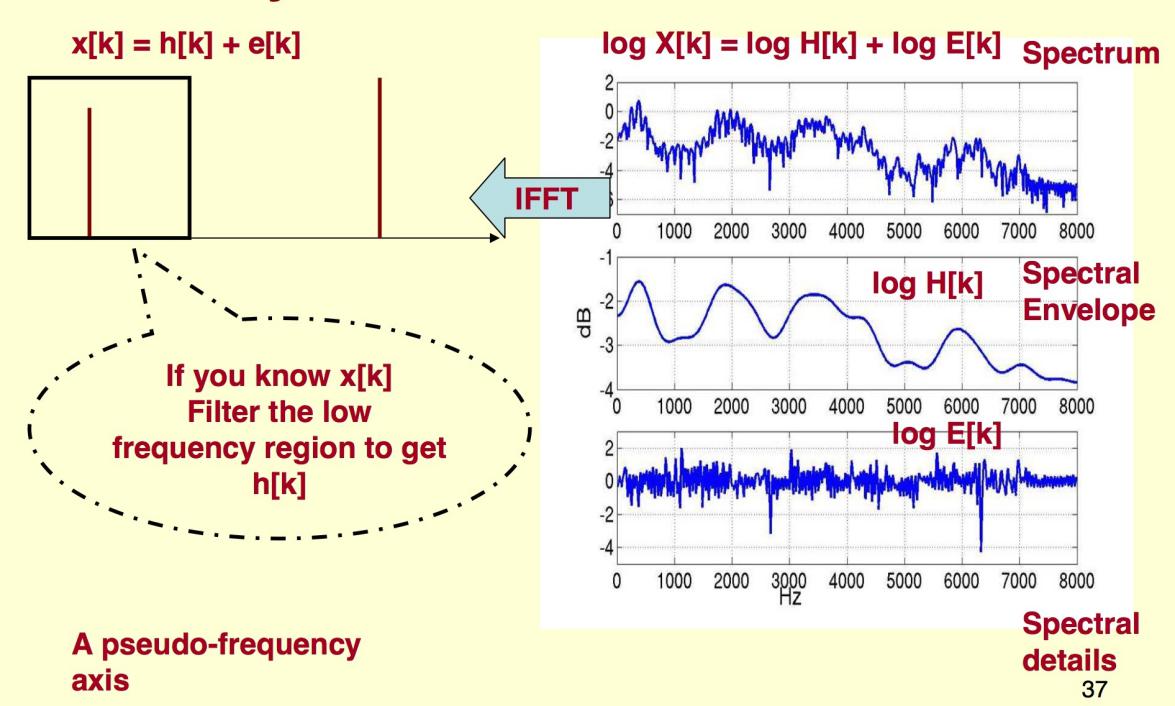


如图所示，DBLSTM是序列到序列映射模型，包括内存块和循环连接，可以在更长的时间内存储信息，并学习最佳的上下文信息量。 中间部分，左部分和右部分（用“t”，“t-1”和“t + 1”标记）分别代表当前帧，前一帧和后一帧。 图中的每个方块表示一个memory block，就是LSTM中的一个cell。 此外，每层的双向连接可以在前后两个方向上充分利用上下文信息，比LSTM能够利用到的信息范围更大了。

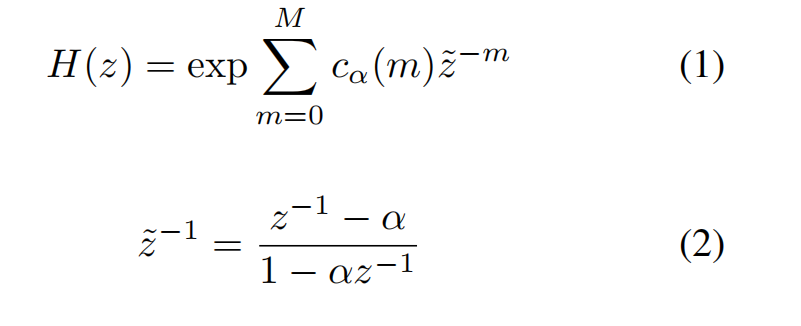
**2.2.2 梅尔倒谱系数MCEP、梅尔频率倒谱系数MFCC**

MCEP（Mel-cepstral coefficients）和MFCC（Mel Frequency Cepstrum Coefficient）是两个很相似的、表示语音特征的量，MFCC更常用。

将时域的音频加窗处理、做DFT，得到频域信号，取对数，做梅尔变换之后再做IDFT可以得到MFCC；也可以在梅尔变换之前先做三角滤波，再把IDFT换成离散余弦逆变换也可以得到MFCC。

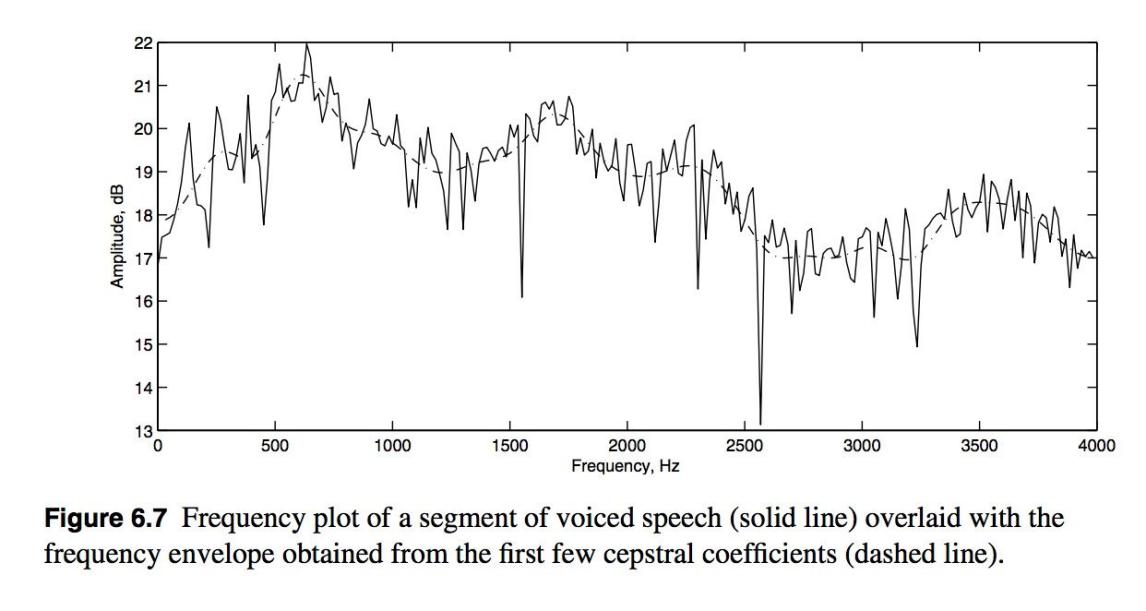


而MCEP与频域信号的包络线有关。频域信号可以看成低频的包络和高频的细节叠加，他们本来是乘积的关系，但因为取了对数，变成相加的关系，因此如上图所示，频谱取对数之后，得到的系数取低位部分（一般是前12个）做IDFT之后就是信号的包络。

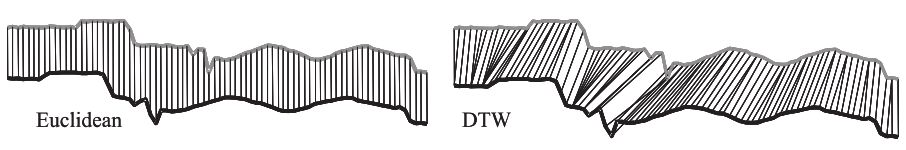


对频谱（包络）H(z)使用上述公式可以计算出cα(m)，第m位的MCEP

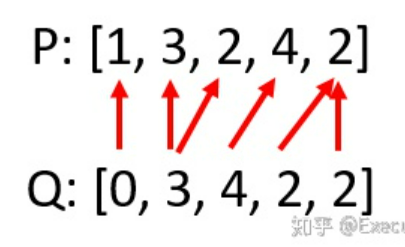
（式（2）是一个全通滤波函数，用来改变频谱的相位）



# **2.2.3 DTW 动态时间规整**

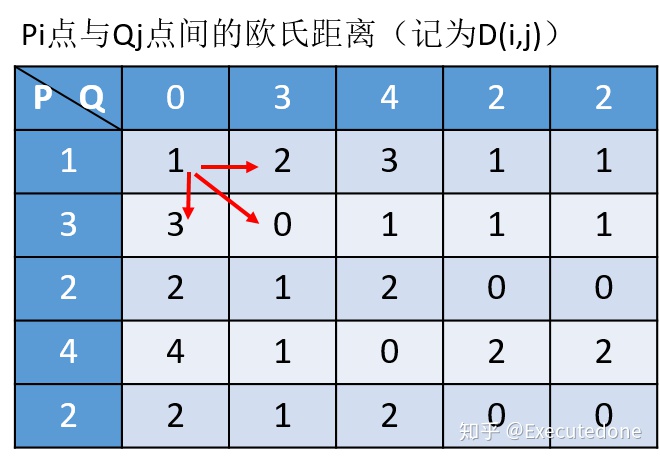


# 对于序列间的相似性，可以用欧几里得距离来衡量，但是对于上图的两个波形，形状相似而相位不同，计算出的距离会偏大，需要DTW（Dynamic Time Warping）算法在时间尺度上将序列配准。

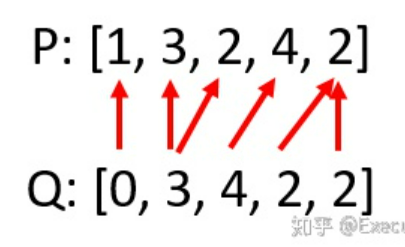
 

这里举一个例子，上图有两个序列，P和Q，长度分别为n和m，需要进行配准（用数值的差来代替欧氏距离方便计算），每个Pi和Qj必须与至少一个Qj和Pi匹配，因此它们可以是一对一、一对多或者多对一的关系，最终所有Pi和Qj都有箭头连接且不能交叉，并且距离之和最小。

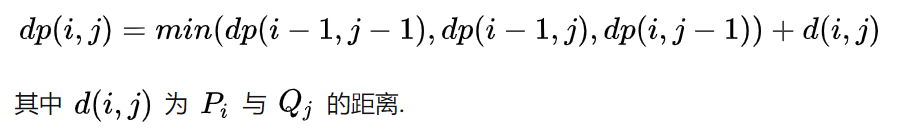
图上分别展示了两种方案，第一种ans=(1-0)+(3-3)+(4-2)+(4-2)+(2-2)=5，第二种ans=(1-0)+(3-3)+(3-2)+(4-4)+(2-2)+(2-2)=2，第二种是更优的（它是最优解）。这个问题用动态规划来求解。



首先计算所有Pi和Qj间的距离。为了求出距离之和的最小值，需要把这个问题转化为矩阵上的路径，对于D(i,j)，表示Pi和Qj有箭头连接。从左上角(1,1)走到右下角(n,m)，可以往右、往下或者往右下走，由于所有的行和列都有至少一个格子在路径上，也就保证每个i和j都有连接的箭头。



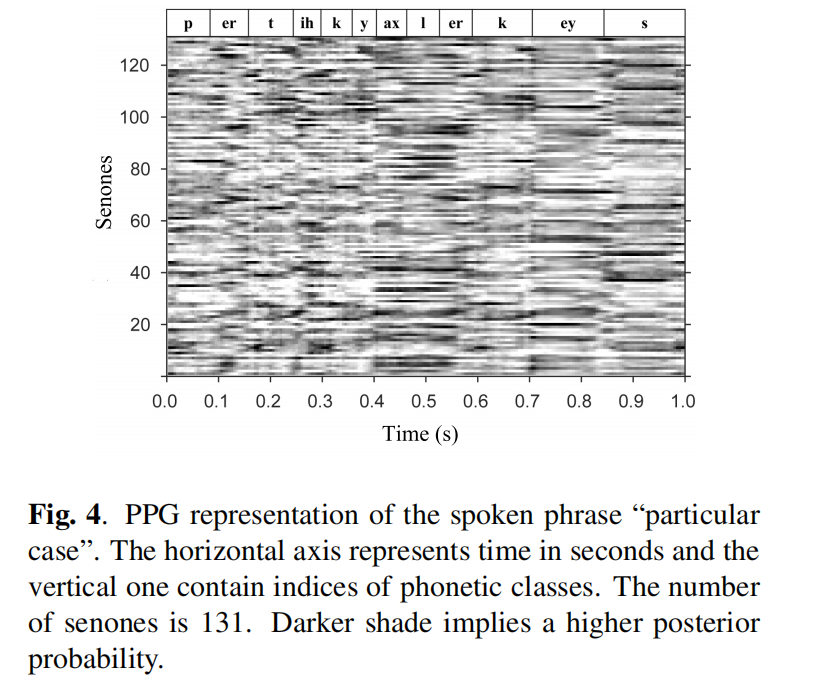
列出状态转移方程。



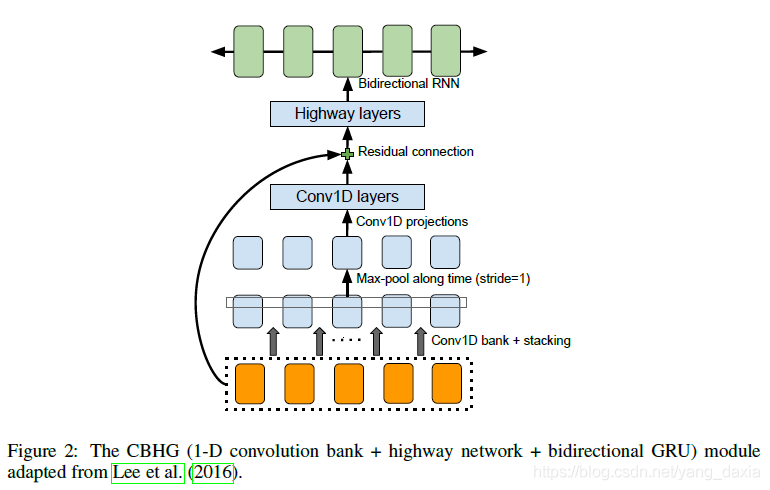
dp(i,j) 表示走到(i,j)的路径的最小值。

根据这个状态转移方程进行递推，dp(n,m)就是答案。黑色的格子是最优解的路径。

**2.2.4 SI-ASR非特定人语音识别和PPGs语音后验图**



SI-ASR（Speak-Independent Automatic Speech Recognition）模型实际上做的是分类工作，它的输出PPGs（Phonetic PosteriorGrams）指的是语音的每一帧对应音素的后验概率。上图颜色越深的地方表示概率越大。



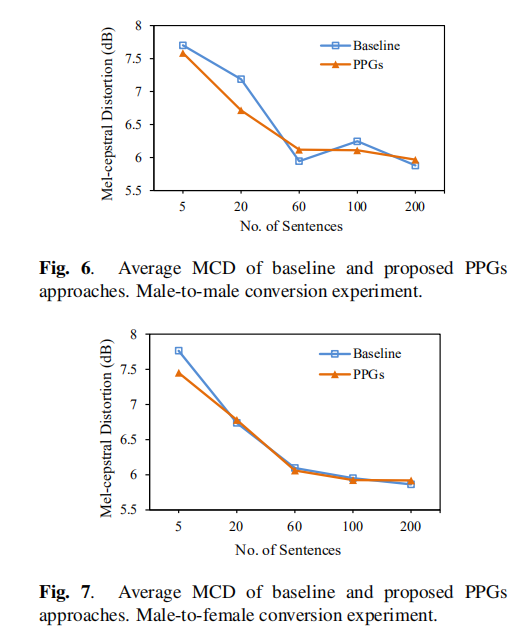
在官方提供的代码中，使用了CBHG（1-D convolution bank+highway network+bidirectional GRU）来实现SI-ASR，它用来提取序列特征、输出分类结果。它先实现的是卷积和池化，之后加入残差连接、highway 网络，解决梯度消失/爆炸的问题，最后用双向GRU做分类（类似BiLSTM）。

**2.3 实验效果**

使用CMU ARCTIC 语料库进行性别内和性别间转换。信号在16kHZ用单声道采样，25ms窗口，每5ms移动一次。提取了声学特征，包括光谱包络线、F0（1维）和AP（513维）通过直接分析。提取39阶MCEPs加上对数能量来表示光谱包络。

**2.3.1客观方法评估**

语音转换的客观评测标准是 MCD（Mel cepstral distortion），它表示的是转换后语音的 MFCC 特征与标准输出语音的 MFCC 特征的差距。

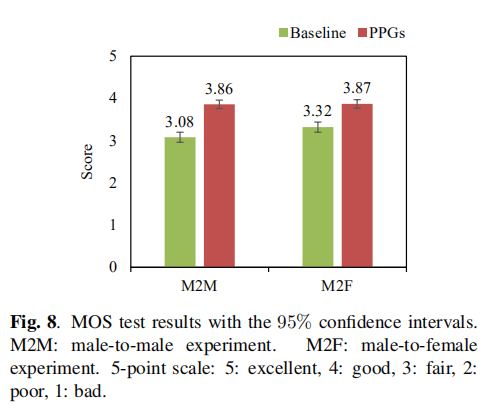


从实验结果来看，客观评估两种方法差距不大。

**2.3.2 主观方法评估**

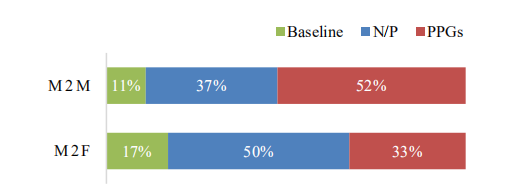
使用了MOS和ABX两种方法， 分别评估转换后语音的自然度和转换前后的相似性。

  MOS（mean opinion score），对于语音的音质，一般采用 5 分制，1 分为最差，5 分为最好。



论文的方法明显表现更好。

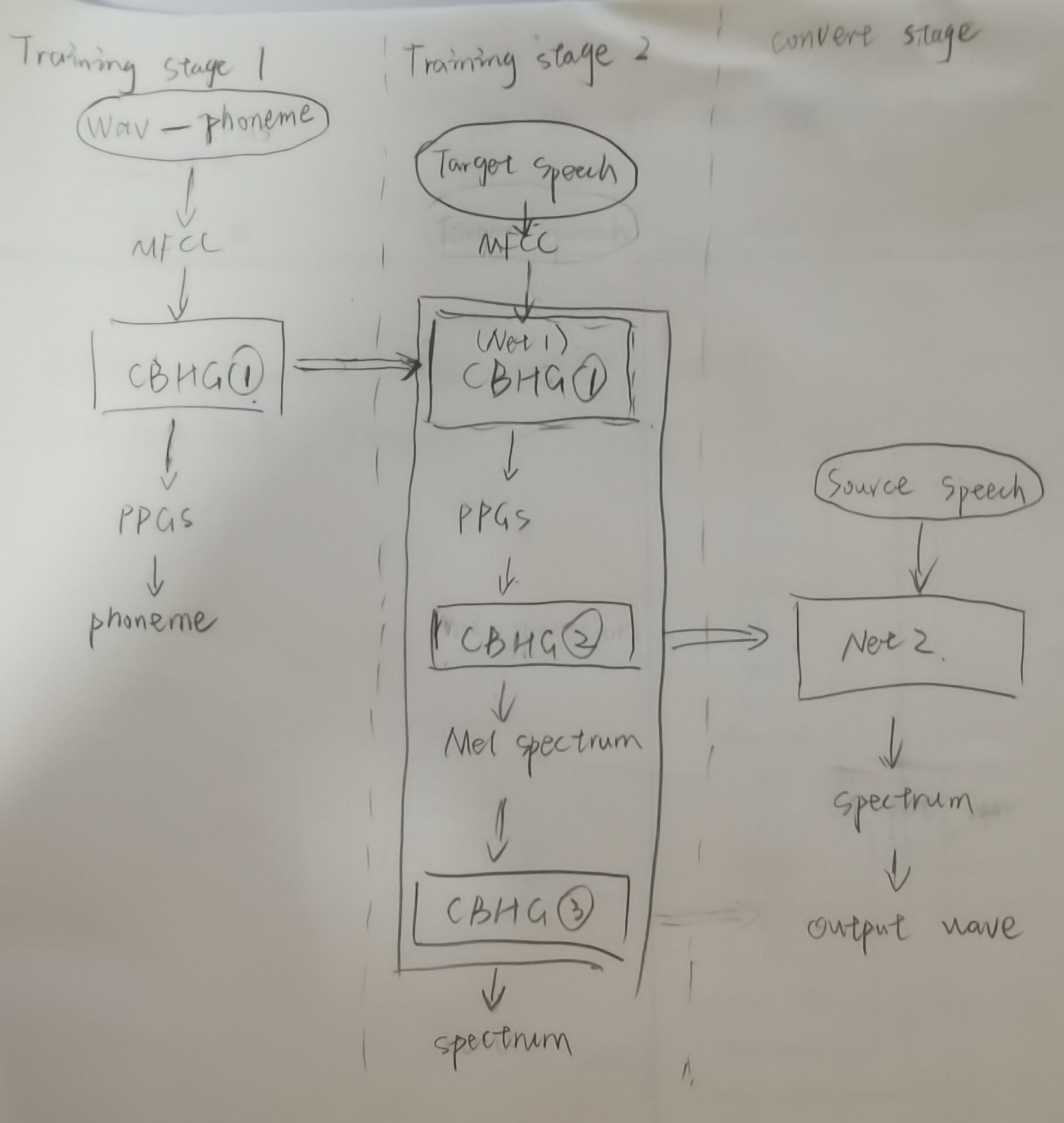
对于相似度，一般是让被试听取两个系统的输出（顺序不定）以及目标说话人的语音，选择哪一个系统的输出更像目标说话人。后面这种测试又常称为 ABX test 或 XAB test，其中 A、B 指两个系统的输出，X 指目标说话人的语音。



依然是论文的方法更好。对于语音转换效果的提升，作者分析了两点原因，并行的方法需要在输入训练数据时需要用DTW算法对齐数据，这个过程会产生误差，而非并行不存在这个问题；非并行方法的训练阶段不需要输入源语音的数据，这最大限度地减少了来自源语音特有信号的干扰。

**2.4 代码的模型构建**

这个实现和论文的出入比较大。



**3 参考资料**

https://github.com/andabi/deep-voice-conversion

<https://blog.csdn.net/zongza/article/details/85766746>

<https://blog.csdn.net/cicisensy/article/details/82696187>

<https://blog.csdn.net/yang_daxia/article/details/83897119>

[1] Sun L, Li K, Wang H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]//2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2016: 1-6.

[2]Song J, Kim H, Lee J, et al. Customization of IBM Intu’s Voice by Connecting Text-to-Speech Services and a Voice Conversion Network[C]//Proceedings of the 51st Hawaii International Conference on System Sciences. 2018.