实验三：基于识别合成框架的语音转换

实验目标

语音转换 VC（Voice Conversion）是智能语音交互领域的一个重要研究方向，其目的在于给定源说话人（Source Speaker）的语音，在保持语音内容不变的情况下，将源说话人的音色转换为新的目标说话人（Target Speaker）的音色。

根据语音转换模型对源说话人、目标说话人是否有限制，可以分为：

（1）一对一的语音转换（one-to-one VC），也即只能实现由特定源说话人到特定目标说话人的语音转换；

（2）多对一的语音转换（any-to-one VC），也即输入可以是任意源说话人的语音，输出是特定的目标说话人音色的语音；

（3）多对多的语音转换（any-to-any VC），也即输入可以是任意源说话人的语音，输出则是具有任意目标说话人的音色；只需给定任意一个目标说话人的若干句语音，即可从这些语音中抽取目标说话人的音色表征，进而生成具有该目标说话人音色特点的语音。

以上多对多的语音转换（any-to-any VC）是目前国内外研究的前沿热点，但需要较为复杂的模型、以及大规模的多说话人录音的训练数据。

本次实验的最终目标，在上述基础上进行了简化，希望实现**一个给定多个特定目标说话人集合（A Set of Many Target Speakers）的限定的多对多的语音转换（any-to-many VC）系统**，即输入可以是任意（any）源说话人的语音，输出语音的说话人音色则可以从若干目标说话人组成的集合（many）中进行选择，从而可以把任意源说话人的语音转换为目标说话人集合中的某个特定说话人的音色。

基本原理

语音转换 VC 的目标为在保持语音内容不变的情况下将源说话人的音色转换为新的目标说话人的音色。实现该目标的一个最基本的思路是先利用语音识别模型 ASR（Automatic Speech Recognition）把源说话人的语音（Input Speech）识别为文字（Text），然后再利用目标说话人数据训练的语音合成模型 TTS（Text-to-Speech）把文字转换为语音（Converted Speech）。

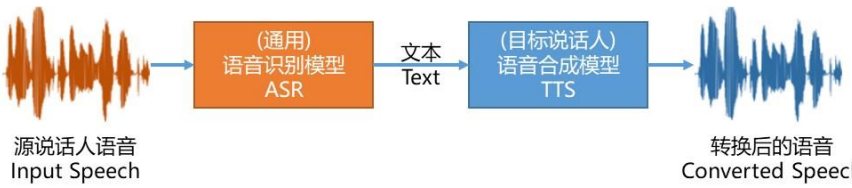

图 1：基于识别合成框架的语音转换基本思路

在实际模型实现时，并不需要将输入语音（Input Speech）识别为最终的文字序列，而

是可以将语音转换为表示内容的某种中间表征（Content Representation），比如和音素（Phoneme）序列相关的信息。理想情况下，每一个语音帧（Frame）都属于某个特定的音素（Phoneme）。但是在进行 ASR 识别的时候，ASR 作为分类任务，往往可以得到某一语音帧（Frame）属于*所有可能音素的后验概率*（Posterior of All Possible Phonemes），我们将其称之为**音素后验概率 PPG**（Phonetic PosteriorGram）。该音素后验概率 PPG 是一种较为理想的语音内容的中间表征。需要注意的是，因为语音识别使用的是通用 ASR 模型，因此可以支持任意源说话人（Any Source Speaker）的语音输入。

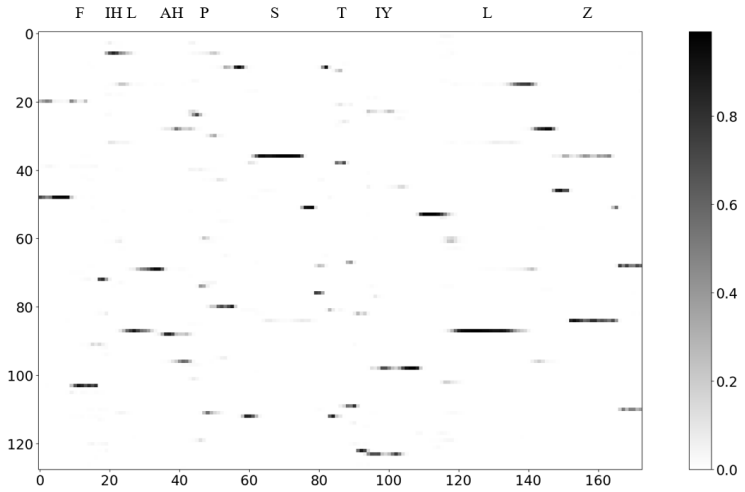

图 2：音素后验概率 PPG 示意图

但是，PPGs 的每一维具有较强的语言学意义，面对识别错误更加敏感。故而在实际应用中，我们也可以使用 PPGs 输出层前一层的瓶颈特征（Bottleneck Features，BNFs）作为代替，使得语音内容的中间表征具有更加鲁棒的结果。

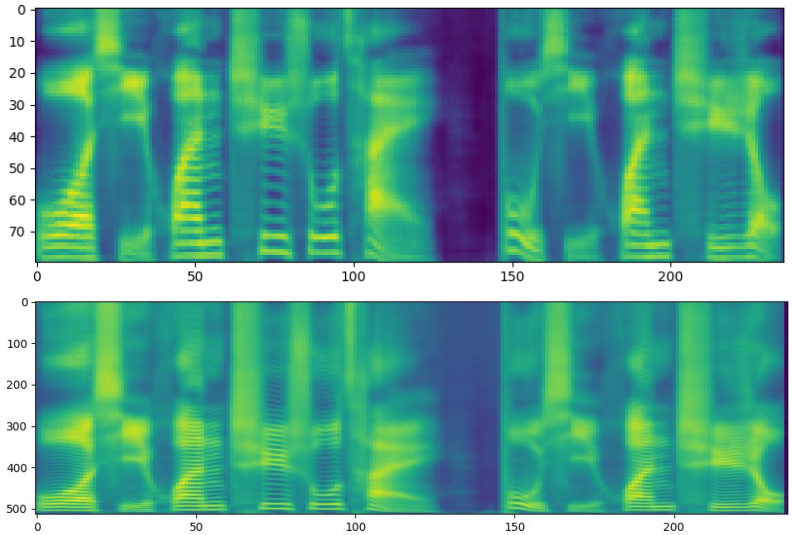

图 3：Mel 谱（80 维，上）与线性谱（512 维，下）对比
语音内容为：“我喜欢吃蔬菜，不喜欢吃肉”

为了将语音内容中间表征（即 PPG 或 BNF）转换为目标说话人的语音，需要训练一个**由语音内容中间表征到目标说话人语音梅尔谱（Mel Spectrogram）的映射模型**（Conversion Model），实现由 PPG 或 BNF 到目标说话人 Mel 谱的映射，并进而基于声码器（Vocoder）

把 Mel 谱恢复为语音波形（Speech Waveform），从而实现目标说话人的语音生成。需要注意的是，上述语音内容中间表征到目标说话人 Mel 谱的映射模型（Conversion Model）是通过使用目标说话人语音提取出来的语音内容中间表征（PPG 或 BNF）参数、Mel 谱参数作为模型的输入和输出来进行训练的。一般使用预测的 Mel 谱与真实的 Mel 谱之间的均方差 MSE（Mean Squared Error）loss 作为模型训练的损失函数。

经过上述过程，即可实现一个简单的基于 ASR 的多对一的说话人语音转换（any-to-one VC）系统，其基本流程图如下图 4 所示。

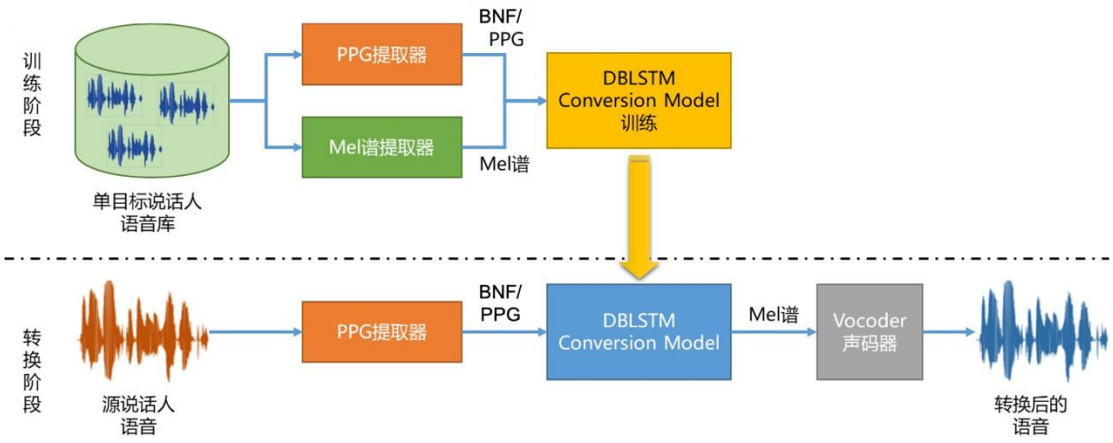

图 4：基于 PPG/BNF 的多对一说话人语音转换系统基本流程

另一方面，为了实现统一模型下多个目标说话人的语音生成，可以在模型中引入说话人相关的表征（比如说话人嵌入 Speaker Embedding）作为模型的额外条件输入。在训练时，可以使用多说话人的语音数据集来训练一个统一的多说话人语音内容中间表征（PPG 或 BNF）与 Mel 谱之间的映射模型（Multi-speaker Conversion Model）。因此可以实现基于语音识别合成框架的多对多的说话人转换（any-to-many VC）系统，其整个流程如下图 5 所示。

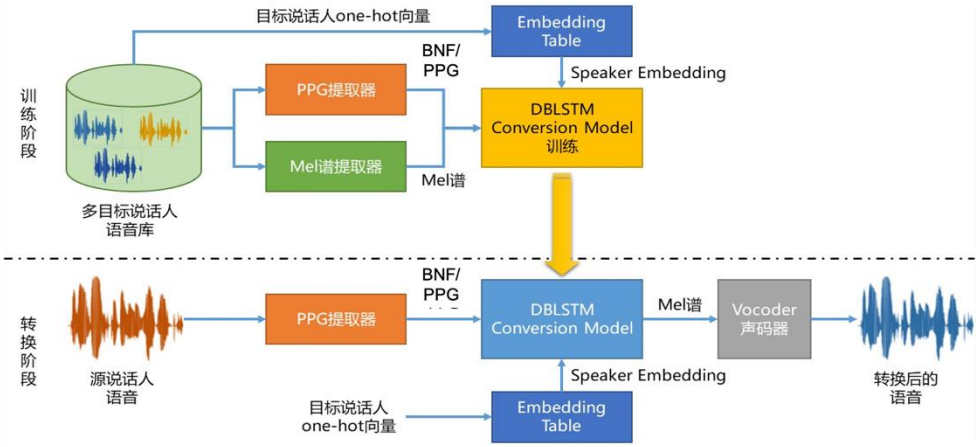

图 5：基于 PPG/BNF 的多对多说话人语音转换系统基本流程

文档简介

《语音信号数字处理》课程的第三次实验课为实现基于语音识别瓶颈特征（ASR

Bottleneck Features，BNFs）的语音转换 VC（Voice Conversion）系统。采用*课下自主实验为主与课上讲解为辅*的形式进行。

本次实验旨在帮助同学们了解 BNFs 和声学参数的提取与调用，熟悉神经网络的搭建与训练，理解 Griffin-Lim 声码器的原理，探索残差网络对性能的影响，掌握运用 BNFs 进行语音转换的基本流程，探究多说话人语音转换系统的实现思路。

本次实验需要在本学期结束之前完成，需撰写实验报告，完成实验任务书中所述实验任务并回答问题。最终的**实验报告、实验代码、以及有关生成的语音**需*打包*后，通过网络学堂提交（截止时间为 2025 年 1 月 13 日 23:59）（注意：不要提交模型或者任何中间结果数据）。

本文档为第三次实验课的指导文档。本文档对实验内容、评分标准进行了说明，提供了实验涉及到的算法的部分参考资料。实验所需要的环境等，请参考实验三对应的 GitHub repository（https://github.com/thuhcsi/dpss-exp3-VC-BNF）。

第 16 周 2024 年 12 月 26 日在国际一期 C504 进行期末汇报，展示实际效果，*要求可现场输入语音进行指定的说话人集合内的多说话人的语音转换*。

因此，请大家务必在 2024 年 12 月 25 日之前完成相关模型的构建、训练等工作。最终的实验报告、代码、生成的语音等，在 2025 年 1 月 13 日 23:59 截止时间之前通过网络学堂提交。

参考资料

1. 基于 PPG 的多对一语音转换系统：PPG-VC
2. Griffin-Lim (GL)声码器 Vocoder：Griffin-Lim Algorithm
3. Residual Network 残差网络结构（参考）：Residual network
4. Speaker Embedding 说话人嵌入：Speaker embedding

注：以上涉及到的 paper 在文件夹“参考文献”下亦可找到。以上资料仅供参考，帮助同学们对本实验中涉及到的算法有一个初步理解，请同学们在此基础上进行文献调研以及探索如何运用到本次实验中。

实验任务

任务内容

实验包括五个子任务：

（1）提取语音识别瓶颈特征（ASR BNFs）和声学参数梅尔谱（Mel Spectrogram）；

（2）搭建、训练、测试特定目标说话人的语音转换模型；

（3）实现**残差网络结构**（注意非 ResNet 残差网络）提升转换语音的音质；

（4）增加说话人嵌入网络，实现多目标说话人的语音转换；

（5）第 16 周实验课实验检查时需完成**“特殊”任务**。

任务说明

本次实验任务采取循序渐进的过程。首先通过任务一、二、三实现一个*多对一的语音转换（any-to-one VC）*模型（如图 4 所示），包括训练数据的准备（任务一）、基本模型的训练及测试（任务二）、改进模型以提升音质（任务三）。

在此基础上，任务四进一步引入说话人嵌入网络，实现给定目标说话人集合的*多对多的语音转换（any-to-many VC）*模型（如图 5 所示）。

任务准备

根据 GitHub repository（https://github.com/thuhcsi/dpss-exp3-VC-BNF）上的要求和建议，搭建环境（Setup Environment）、下载数据（Data Preparation）。

其中数据集可以直接从 GitHub 上所给的链接进行下载。本实验一共使用来自标贝、MST Challenge 的 3 个说话人，包括：bzn, mst-female, mst-male 三个说话人，其中 bzn, mst-female 为女声说话人，mst-male 为男声说话人。

在本次实验中，第一部分（任务一、二、三）实现多对一的语音转换（any-to-one VC），此时目标说话人固定为 bzn。即**任务一、二、三使用 bzn 话人的语音数据进行实验**。

第二部分（任务四、五）实现给定目标说话人集合的多对多的语音转换（any-to-many VC），此时目标说话人集合包括上述所有 3 个说话人。即**任务四、五使用全部说话人的语音数据进行实验**。

任务一：提取 BNF 与声学参数（15 分）

任务说明

课上提到，人耳对语音的感知主要是对不同频率成分的感知，主要体现在语音中不同频率成分的分布上，即语谱图（Spectrogram）；除此之外，还包括声源相关的特性，比如声带振动的频率，称之为基频（Fundamental Frequency，F0）；以及通过声带是否震动的特性来区分浊音（Voiced）还是清音（Unvoiced）等。

提取/计算相关声学参数时常采用短时分析的方法，将原始的语音波形序列进行加窗处理，得到一系列的语音帧（Frame），此过程称之为语音分帧。分帧需要考虑所用的窗长、窗移等。另外，常采用傅里叶变换（如 FFT）将时域信号转换为频域信号，并计算各频率分量的能量值，得到线性的语谱图，即线性谱（Linear Spectrogram）。再者，人耳对声音的感知是非线性的，对低频成分更加敏感，而对高频成分相对不敏感。为了使得提取的特征更加符合人耳的这种特性，常将线性 Linear 谱转换为对数域的梅尔谱（Mel Spectrogram）。

本任务准备模型训练用的数据。对所提供的语料库中的语音音频进行特征提取，提取出

ASR 瓶颈特征 BNFs 以及 Mel 谱等声学参数，并做归一化（normalization），进而进行数据集划分，将数据集划分为训练集、验证集、测试集。

其中，BNFs 提取器（即 ASR 模型）的具体结构请参考 dpss-exp3-VC-BNF/models/wenet/ 下的 py 文件，参数设置请参考 dpss-exp3-VC-BNF/config/asr_config.yaml。

任务要求

阅读上述任务说明中所提到的相关 Python 代码文件，了解特征提取的过程。

根据 GitHub repository 的说明文档中的要求（Any-to-One → Feature Extraction），运行相关命令进行特征提取。**请注意对提取的特征中的异常数据进行检查与去除，尤其需要注意提取的特征数据中是否出现 NaN，若出现需要将相应的数据条目从*.csv 中删掉。**

回答下列问题。

1. 提取音素后验概率瓶颈特征 BNFs（4 分）

（1）简要说明 BNFs 是什么，其在基于识别合成的语音转换框架中起到什么作用？

2. 理解 Mel 谱等声学参数的提取过程（3 分+3 分）

（2）理解声学参数提取的过程，为 hparam.py 中 class Audio 的主要参数添加注释，说明该参数的意义，可在实验报告中截图或者表格展示。

以下图片中有“#”行的参数必须添加注释（3 分），其他参数可选（最多 3 分加分）。

```
class Audio:
    num_mels = 80              #
    ppg_dim = 347              #
    num_freq = 1025            # cannot be too small
    min_mel_freq = 30.         #
    max_mel_freq = 7600.       #
    sample_rate = 16000        #
    frame_length_ms = 25       #
    frame_shift_ms = 10        #
    upper_f0 = 500.            #
    lower_f0 = 30.             #
    n_mfcc = 13
    preemphasize = 0.97
    min_level_db = -80.0
    ref_level_db = 20.0
    max_abs_value = 1.
    symmetric_specs = False
    griffin_lim_iters = 60
    power = 1.5
    center = True
```

图 6：Audio 类及其相关参数定义

3. 提取 Mel 谱等声学参数（6 分）

（3）Mel 谱和线性谱的提取过程有什么差异？它们之间是什么关系？

（4）提取出来的线性谱和 Mel 谱各是多少维的特征参数？Mel 谱的频率范围是多少？

（5）在语音转换中 BNFs 提供了内容信息，基频参数 F0 提供了什么信息？为什么要考虑基频参数 F0？

4. 数据集的划分（2 分）

（6）实验中将数据集划分为了训练集、验证集、测试集。它们之间的默认划分比例是多少？

任务二：训练并测试特定目标说话人的语音转换模型（40 分）

任务说明

基于任务一提取出的 BNFs 和 Mel 谱的数据，将其作为 BNFs 到 Mel 谱映射模型（Conversion Model）的输入和输出，训练该映射模型至收敛；并进而利用训练好的模型进行测试和语音转换。

训练过程详情可参考 train_to_one.py，VC 模型结构和参数设置详情参考 models/models.py 和 config/hparams.py，数据集的定义详情请参考 datasets/dataset.py，语音转换的具体过程请参考 inference_to_one.py。

任务要求

阅读上述任务说明中所提到的相关 Python 代码文件，了解数据集及数据加载模块的定义、Conversion Model 转换模型的定义、转换模型的训练过程（前向计算、反向传播）、损失函数的定义、模型验证、模型测试、以及语音转换等过程。

根据 GitHub repository 的说明文档中的要求，运行相关命令进行模型训练（Any-to-One → Train）、推理（Any-to-One → Inference，即目标说话人的语音转换）。

回答下列问题。

1. 数据集及数据加载模块定义（5 分）

（1）参考 datasets/dataset.py,了解 DataSet 及 VCDataSet 类的原理，熟悉 DataLoader 类调用时各参数的含义。在实验报告中回答 DataLoader 的调用参数 batch_size、shuffle、num_workers、collate_fn 分别是什么含义？

2. 转换模型定义（8 分）

（2）阅读 train_to_one.py，找出转换模型的定义，并根据 models/models.py 说明转换模型的网络结构，给出网络的基本结构图。

（3）进而说明转换模型的输入、输出参数分别是什么？输入、输出参数的维数是多少？

3. 转换模型训练（12 分）

（4）说明转换模型的训练过程，如何进行前向计算？使用了什么损失函数？损失函数

是怎么计算的？如何进行误差反向传播？

（5）给出模型训练时的损失函数曲线；模型训练了多少个 epoch？训练完成后，最终（对应最后一个 epoch）的平均训练 MSE loss 是多少？

（6）模型训练的结果保存在什么地方？

4. 转换模型验证（4 分）

（7）基于上述训练好的模型，在验证集上进行验证。在验证集上的 MSE loss 是多少？

5. 转换模型测试（5 分）

（8）基于上述训练好的模型，在测试集上进行测试。针对某句测试用例，得到该用例的真实语音以及预测的转换语音；给出该测试用例的真实 Mel 谱、预测 Mel 谱的图，简单分析它们之间的区别，并计算这两个 Mel 谱之间的均方差 MSE 距离。

（9）分析在 inference_to_one.py 中所调用的 inv_mel_spectrogram 函数，说明将 Mel 谱恢复为语音波形（Speech Waveform）的基本过程，理解声码器（Vocoder）的作用。在实验报告中只需给出 inv_mel_spectrogram 函数的流程，说明其所调用的各个子函数的基本功能，不需要深入到子函数中。

6. 进行语音转换（6 分）

（10）结合 inference_to_one.py，说明转换阶段由输入源说话人的语音到输出特定目标说话人语音的整体流程。

（11）选取上述训练好的模型所对应的 checkpoint 文件（ckpt 参数），给提供数据中某个特定说话人的某句语音作为输入（src_wav 参数），进行语音转换得到转换后的语音（save_dir 参数）。

（12）自己录制一段语音，并作为模型的输入（src_wav 参数），重复上述（11）的语音转换过程，得到转换后的语音。

（13）在实验报告中对（11）（12）中的转换后的语音的效果进行简单分析。

（14）简要说明多对一语音转换（any-to-one VC）为什么能够实现给定任意源说话人的“多”对一的语音转换，关键是什么？

任务三：探究残差网络结构对转换性能的影响（15 分）

任务说明

实验表明，基于 Mel 谱进一步预测其残差信息（Residual Information），并将该残差信息与原有 Mel 谱结合，能有效提升参数预测的性能，如下图所示。其中残差网络结构（ResidualNet）用来由 DBLSTM 生成的 Mel 谱进一步预测残差信息。

本任务将探究残差网络结构的作用，特别是其对 BNFs-Mel 谱转换网络性能的影响。

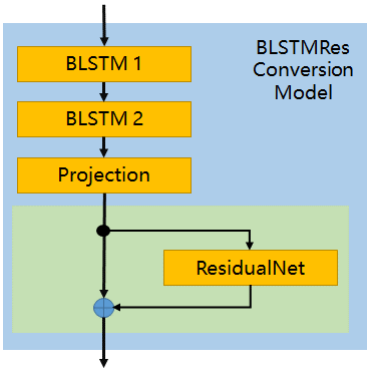

图 7：一种典型的残差网络结构及其在转换网络中的应用

任务要求

1. 实现残差网络结构（8 分）

（1）根据残差网络结构的有关原理，确定 ResidualNet 的某种特定结构（考虑到数据量的大小，建议不要使用很复杂的网络结构），实现 models/models.py 中 ResidualNet 类的具体代码；并对 models/models.py 中 BLSTMResConversionModel 类的相应部分进行修改。

（2）在实验报告中说明所实现的 ResidualNet 的具体结构，并给出网络的基本结构图。

2. 探究残差网络对转换模型性能的影响（7 分）

（3）进行消融实验（Ablation Study），探究上述任务二中无残差网络结构的转换模型、与本任务三中实现的有残差网络结构的转换模型的性能差别。

可以通过模型在训练过程中的 MSE loss 曲线的差异；模型训练完成后在验证集上的 MSE loss 值的差异；使用同样的测试用例的情况下，转换语音、预测的 Mel 谱及谱图的差异等来进行分析等等。

注意：

为了训练和测试具有残差网络结构的语音转换模型，你需要参考 train_to_one.py 定义自己的训练代码（如 train_to_one_res.py），以及参考 inference_to_one.py 中义自己的测试代码（如 inference_to_one_res.py），并在你定义的代码文件中对 BLSTMResConversionModel 模型进行调用。

任务四：增加说话人嵌入网络，实现多目标说话人的语音转换（20 分）

任务说明

以上三个任务实现了一个多对一的语音转换（any-to-one VC）模型，本任务将在此基础上，通过增加说话人嵌入网络的结构，构建包含说话人嵌入网络的多对多的语音转换（any-to-many VC）模型，从而实现支持多个目标音色的多目标说话人的语音转换系统。

为了实现说话人嵌入网络，一个简单的办法就是给定说话人的 one-hot 标签，通过一个

嵌入映射表（Embedding Table）得到该说话人对应的嵌入向量（Speaker Embedding），然后再通过一个全连接层，将全连接层输出与原有 DBLSTM 的两个 BLSTM 层的输入叠加，以此在原有的基于 DBLSTM 的转换模型中引入说话人信息，实现多说话人的转换模型，相关结构如下图所示。

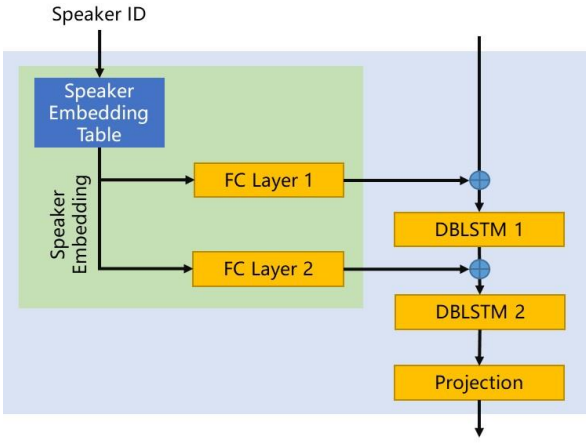

图 8：一种典型的说话人嵌入网络结构及其在转换网络中的应用

本任务将提取多说话人语音数据的声学特征参数，为训练多对多语音转换模型做好准备；实现说话人嵌入网络，并将其与上述多对一转换模型整合，实现多对多的语音转换模型；训练多对多的语音转换模型，并在验证集、测试集上观察模型的性能；并进而利用训练好的模型实现给定任意说话人的语音，将其转换为多说话人集合中的任意一个目标说话人音色的语音。

训练过程详情可参考 train_to_many.py，模型结构和参数设置详情见 models/models.py 和 config/hparams.py，数据集的定义详情请参考 datasets/dataset.py，语音生成的具体过程请参考 inference_to_many.py。

任务要求

1. 准备训练数据：提取多说话人的 BNFs 及 Mel 谱等声学参数（3 分）

（1）根据 GitHub repository 的说明文档中的要求（Any-to-Many → Feature Extraction），运行相关命令进行特征提取。**请注意对提取的特征中的异常数据进行检查与去除，尤其需要注意提取的特征数据中是否出现 NaN，若出现需要将相应的数据条目从*.csv 中删掉。**

2. 实现说话人嵌入网络（9 分）

（2）根据说话人嵌入网络的有关原理，确定说话人嵌入网络的某种特定结构（鼓励探索更好的模型结构），并实现 models/models.py 中 SPKEmbedding 类的具体代码；如果有需要，可进一步对 models/models.py 中 BLSTMToManyConversionModel 类的相应部分进行修改。

（3）在实验报告中说明所实现的 SPKEmbedding 的具体结构，并给出网络的基本结构图。

（4）若对 BLSTMToManyConversionModel 类进行了修改，也需要在实验报告中加以说明。

3. 多目标说话人转换模型训练、验证、测试（4 分）

（5）根据 GitHub repository 的说明文档中的要求（Any-to-Many → Train），运行相关命令进行模型训练。

（6）给出模型训练时的损失函数曲线；模型训练了多少个 epoch？训练完成后，最终（对应最后一个 epoch）的平均训练 MSE loss 是多少？

（7）基于上述训练好的模型，在验证集上进行验证。在验证集上的 MSE loss 是多少？

（8）基于上述训练好的模型，在测试集上进行测试。针对某句测试用例，通过给定不同的目标说话人的 Speaker ID，得到同一个源说话人的语音对应的不同目标说话人的转换语音；听辨转换语音的音色与目标说话人的音色，是否存在差异？为什么？

4. 进行多目标说话人语音转换（4 分）

（9）根据 GitHub repository 的说明文档中的要求（Any-to-Many → Inference），将某个特定源说话人的语音（src_wav 参数）转换为某个特定目标说话人（tgt_spk 参数）的语音。

（10）自己录制一段语音，并作为模型的输入（src_wav 参数），对提供的中的 3 个目标说话人（改变 tgt_spk 参数）进行语音转换，得到 3 个转换后的语音。

任务五：第 16 周实验课“惊喜”任务（10 分）

任务要求

1) 第 16 周实验课现场布置，“惊喜”等着你来现场揭开！(✿◡‿◡)

2) 请确保在 16 周实验课之前完成以上 4 个任务，**否则你将不能完成 16 周的实验课上的任务要求**。

====文档结束====