湖州师范学院

2023 届毕业设计(论文)

课 题 名 称: 基于 CycleGAN 的非平行语音转换系统研发

课 题 名 称（英文）: **Research and development of non-parallel voice conversion system based on CycleGAN**

学 生 姓 名: 王会元 学 号: 2019082201

专 业 名 称: 物联网工程

指 导 教 师: 王泽峰 职 称: 教授

所 在 学 院: 信息工程学院

完 成 日 期: 2023 年 3 月 26 日

教务处制表

基于 CycleGAN 的非平行语音转换系统研发

摘要：语音转换技术是指在在不改变源说话人说话内容的前提之下，改变这句话的音色、音调，使之听上去像是目标说话人说出的话的技术。得益于科技水平的进步与发展，目前能够实现语音转换的技术有许多种，例如 CycleGAN、自编码器、变分自编码器以及 GAN 都能够完成该功能，此外还有另辟蹊径的方法，比如先将源说话人的说话内容读取出来保存为文本，后续再使用 tts 技术通过文本生成出相应的语音以达到语音转换的目的。本文将研究基于 CycleGAN 的非平行语音转换系统。在实现语音转化的任务后，本文发现转换后的音频虽然能够实现语音转换的基础要求，但是其质量不好参杂噪音。因此本文认为可以采用滤波器或者降噪的手段对输出的音频进行处理，以达到提高输出音频的质量，增强自然度的效果。本文采用并对比了高斯滤波器、带通滤波器、中值滤波、小波去噪和非线性谱减法 4 种方法对语音音频进行去噪且记录了其 MSE 损失和差分两项数值作为客观评价，当然，MSE 和差分这两项数值虽然能够体现出降噪前后音频的变动程度，但是并无法表达出降噪后音频表现出的真正效果，因此本文结论的得出还结合本文笔者的主观评价。结合主客观评价，上述五种方式之中使用非线性谱减法并且设置阈值为 0.3，幂次方为 1.1 时输出的音频能够拥有最好的效果。

关键词：语音转换，深度学习，非平行语料，CycleGAN

Research and development of non-parallel voice conversion system based on CycleGAN

Abstract: Voice conversion is a technique that changes the tone and pitch of a sentence to make it sound like the target speaker's voice, without changing the content of the source speaker's speech. Thanks to the progress and development of technology, there are many techniques to achieve speech conversion, such as CycleGAN, self-encoder, variational self-encoder and GAN, and there are also alternative methods, such as reading out the source speaker's speech and saving it as text, and then using tts technology to generate the corresponding speech from the text to achieve the purpose of voice conversion. This paper will study the CycleGAN-based approach. In this paper, we study a non-parallel voice conversion system based on CycleGAN. After achieving the task of voice conversion, this paper finds that the converted audio can achieve the basic requirements of voice conversion, but its quality is not good with noise. Therefore, this paper argues that the output audio can be processed by means of filters or noise reduction in order to improve the quality of the output audio and enhance the naturalness. This paper uses and compares four methods to denoise speech audio: Gaussian filter, band-pass filter, median filter,wavelet denoising and nonlinear spectral subtraction, and records two values of MSE loss and difference as objective evaluation, of course, the two values of MSE and difference can reflect the change of audio before and after noise reduction, but cannot express the real effect of audio performance after noise reduction. Therefore, the conclusion of this paper also combines the author's subjective evaluation. Combining the subjective and objective evaluations, the best results are obtained by using the nonlinear spectral subtraction method with a threshold of 0.3 and a power of 1.1 among the above four methods.

Keywords: Voice conversion, deep learning, non-parallel, CycleGAN

目 录

第一章 绪论

本章说明了语音转换技术的任务，分析了语音转换对相同领域的其他技术的意义及其社会意义，深度阅读了相关方向的文献并对其总结归纳，确定了本文实现语音转换任务的方式。

1.1 选题的意义

语音转换指的是，输入源说话人说话的一段音频，能够在不改变其说话内容的前提之下，将其声音特征例如音色、音调给改变为目标说话人的声音特征，使输出的音频像是目标说话人说出来的一项技术。

随着科学技术的发展，语音转换技术的出现及其进步无疑给声音领域的其他技术，如语音合成、语音增强、语音识别等技术带来了强大的推进力量。在语音合成的领域，这种技术的实现方法使通过输入文本的方式合成出相应的语音音频，其缺点是合成出的语音波动较小，没有情感起伏且间隔过于平整，使其合成出的语音整体不自然，而语音转换的技术则补足了其没有感情起伏的缺点，使其输出的音频更加自然可信。在语音增强领域，可以先使用语音转换的技术将较低质量的语音转换为高质量语音从而是语音增强的起到更好的效果。在语音识别的领域，这项技术对音频本身的要求较高，否则会难以识别出较为准确的音频内容，而语音转换的出现可以改变其说话人特征，填补了语音识别在不同说话人之间识别的精准度不同的问题，提高了语音识别的准确度和速度。

此外，语音转换技术在社会意义上也有着很大的意义。这项技术可以运用在语音疾病诊断、语音情感识别、语音语调转换等领域，为医疗、心理学等领域的研究和应用提供支持。语音疾病诊断实际上是语音识别在医疗方面的运用，这项技术可以通过声音的一些特征来判断对象是否有特定的病症，但是这项技术需要大量的样本输入，而语音转换可以在只有较少样本的情况下转换出相应对象的声音作为样本，从而弥补样本不足的问题。语音转换技术本身就会设计语音情感识别和语音语调转换的内容，因此也可以在这些领域的发展中有所贡献。语音转换技术在医疗和心理学的领域的应用主要是在精神疾病的心理治疗中，患者本身是否愿意配合治疗会对治疗的效果产生很大的影响，患者越是有配合意愿，治疗也就越容易见效，在这方面，可以使用语音转换的技术将医师的声音转换称为患者亲近的人的声音，使其提高配合度，增强治疗的效力。

1.2 国内外研究现状

语音转换技术有两方面的工作，一方面是使用提取输入语音样本的声音特征，并将声音特征重新合成为音频，另一方面工作就是将提取出来的声音特征通过算法将其从源改为目标的声音特征并输出用于合成。这两项工作分别是由声码器和算法实现的，此处也将提到几种声码器和可以实现语音转换的算法。

1.2.1 声码器的研究

张小峰等人归纳道语音合成中的声码器被分为自回归式和并行式，其中自回归式包括 wavenet、SampleRNN、waveRNN 编码器，并行式包括 Parallel Wavenet、WaveGlow、FloWavenet 编码器。其中自回归式编码器是在较早期被广泛使用的声码器，它有两种网络模型，分别是基于卷积神经网络的和基于循环神经网络，通过卷积神经网络构建的声码器为了达到按时序生成语音的目的利用了因果卷积、带洞卷积等卷积方式。基于循环神经网络构建的声码器则是通过长短时记忆网络和门循环单元等方式实现按时序生成[1]。

在语音合成的项目中尤其关注声码器的使用，上述综述提到了声码器的分类，接下来本文会根据上

述所提到的分类讨论针对声码器的研究。

凌震华等人对 wavenet 声码器进行了研究，他们认为，wavenet 声码器采用了非线性的变换，能够重构音源输入的基频和频谱特征的波形，从而改善音源的质量，也可以因此避免其他声码器可能会存在的丢失频谱细节及相位的问题。在语音输入方面，他们分别了输入自然声学特征和预测声学特征，这两种方式在输出语音的质量中并没有明显的区别。他们还尝试了减少输入量，发现其自适应的特征能够较好的客服语料不足的问题，生成出质量相对较好的音频。他们总结道现阶段，wavenet 声码器相对与其他的声码器，依旧存在这生成语音时复杂度高以及效率低的问题[2]。

MEHRI S 等人设计了一种 SampleRNN 声码器，他们使用了循环神经网络来搭建声码器的网络模型，这种声码器是由多层循环神经网络组成的，其中最低的一层作为采样层，而其他的层都被称为帧层。其特点是能够同时在不同的时钟速率下学习不同抽样级别样本的特征。这其中帧层能够在一个时间步下学习多个帧样本，而采样层在同样的一个时间步内只能够学习一帧样本。与 wavenet 声码器相比，SampleRNN 声码器拥有更快的语音合成速度和更简单的模型构成，且生成质量也能够和 wavenet 声码器所生成的音频不相上下，但是其缺点是需要用较多的技巧加速模型训练，因而会浪费计算资源[3]。

NAL K 等人则针对语音合成耗时大的问题设计了一种 waveRNN 声码器，这种声码器改变了自身的结构，减少了神经网络的层数，使用了一层循环神经网络和两层 softmax 层的模型，并且还运用了裁剪法来使神经网络的每一层计算耗时减少，这种结构及方法的改变极大的加快了其生成语音的速度，甚至在计算资源很少的设备上也能够成功的运行，生成出目的音频[4]。

HAO Y 等人提出了能够实时音频合成的 Parallel Wavenet 声码器，这种声码器需要经过两个训练网络，且对其中的教师网络质量要求较高。它与原始的 wavenet 的不同之处有以下几点，它在训练过程中并没有使用因果卷积，而是使用了非因果卷积进行替代，并且输入其中的噪声是在高斯分布中随机提取出的，最大的不同就是 Parallel Wavenet 声码器是非自回归的[5]。

安鑫等人在使用 waveGlow 声码器的过程中总结道，这种声码器以梅尔频谱图的输入为前提能够输出质量较高的音频，它结合了 Glow 和 wavenet 的特点，不仅拥有了高效推理预测的特点，还拥有快速合成高质量音频的优点，这几个特质使 waveGlow 声码器无需进行自回归[6]。

王研等人在设计语音合成模型时使用了 WORLD 声码器，这种声码器是一种基于频域分析的语音分析与合成技术，解决了传统基于时域信号分析的声码器难以处理高音和噪音的问题。其核心思想是将语音信号转换为频域的声道特征和基音周期，然后利用这些特征来合成语音信号。WORLD 声码器拥有合成语音质量高、自然度好，能够更好地处理高音、噪声等问题的优点[7]。

刘畅等人从声码器、语料对齐以及迁移模型 3 个重要影响因素的角度对研究现状进行分析，基于参数语音合成的传统编码器可以解决从一个说话人到另一个说话人转换的问题，其思路是将共振峰作为声道特点，使机器学习目标共振峰，将材料进行变换使其共振峰轮廓相重合，并将其音高相匹配，以合成出符合目标特征的语音，但其缺点是在迁移过后，语音的质量容易下降。完全自回归的 wavenet 声码器可以在端对端文本-语音任务中产生高质量的语波形。如果能获取平行语料，那么对语料进行处理的难度将会很小，但大多情况下，平行语料都很难或是不可能获取，对于非平行语料的处理可以在训练时使用基于线性预测编码模型和基于谐波加噪声模型的方法进行处理，也可以采用平行语料与非平行语料的训练相结合的方法进行处理。传统的迁移模型是基于高斯混合模型的迁移模型，此种模型会导致风格迁移后的语音过于平滑而失去自然感。基于人工神经网络的迁移模型在情绪表达方面目前具有最好的性能[8]。

1.2.2 算法的研究

实际上语音转换技术能够以许多方式实现，一种是传统的语音转换思路，将源说话人和目标说话人的声音特征和说话内容区分开，计算出源说话人和目标说话人声音特征之间的映射，通过映射改变音频的声音特征，保留说话内容从而实现语音转换的目的。另一种思路则是将源说话人说的话通过语音识别技术给提取成文字，再通过语音合成的技术将其转换成目标说话人的声音。两者皆可以实现语音转换的目标，此处将先讨论语音合成的研究。

以下为在语音合成方向的的研究。

在这方面，唐浩彬等人对近年来基于情感及韵律的表现性语音合成进行了全面的总结、比较和分析，首先是普通的语音合成，被分为三类：自回归语音合成模型、非自回归语音合成模型、基于神经网络的语音合成。这类用深度学习合成出的语音有明显的缺陷，即音调平滑、无节奏、无表现力。上述这类语音合成都只关注到了语音合成三要素中的“说什么”，而没有考虑到“谁说”、“如何说”。表现性语音转换则是在不影响“说什么”的正确性的前提下，重点关注“谁说”与“如何说”，此处的信息输入会被分为两类，一类为显式信息，包含说话人、音高、持续时间等，另一类则为隐式信息，包含全参考编码器、全局风格标记、变分自动编码器等。此处使用由生成数据的生成器和判别数据真实性的判别器组成的生成性对抗网络可以有效提高建模能力[9]。

张冠萍针对英语翻译机器人合成后的音频过于不自然，从而导致人机交互效果不好的问题设计了基于 HMM 的语音合成模型，在其使用相同数据集为输入的情况下，用 HMM 实现的语音合成模型相对于英语翻译机器人原本的合成模型表现出的结果有更高的 MOS 评分和相似度，该模型能够显著的提高翻译机器人合成出语音的真实度和自然度[10]。

李乃寒对准了语音合成目前存在的自然度不高、文本音频的对准以及鲁棒性三个问题入手，对该技术进行了研究，在这项研究中，他使用了序列到序列转换的模型，其代表模型有 Seq2Seq 和 Transformer，在使用了 Transformer 的基础上，他还使用了 Tacotron2 声谱预测网络，这个网络能够生成出非常接近人类声音的波形，这种结合方式提高了生成出的音频的自然度。同时他在分析了鲁棒性受限的原因之后，提出了 Robu Trans 模型，这种模型能够在保持前面模型的自然度的前提下，有效的提升整个算法的鲁棒性。除此之外，他还提出了 Mobo Aligner 模型，这种模型是基于神经网络提出的全新的注意力机制，可以用于音频和文本的对齐，测试结果显示，这种对齐方式非常精准，同时还能减少 45%参数输入量和 30%的训练时间[11]。

王智等人针对语音合成技术普遍忽略的情感不足的问题，研究了情感语音的合成方法，他也使用了 Tacotron2 模型作为语音合成的工具，此外他选择了音频种的基频和能量这两个参数作为判断情感特征的参数进行分析，在此处他借用了 PRAAT 声学分析软件用于提取参数，他统计了 6 种情感的 400 局语音，从中记录了每种情感能量和基频的最大最小值和均值作为后续合成的依据。测试结果显示，通过引入能量和基频作为情感特征参数确实可以达到合成出情感语音的预期，但是该种方法在预测频谱时会严重受到文本长度的影响[12]。

帕丽旦 •木合塔尔等人针对 HMM 模型合成出的模型虽然稳定性高但是自然度较低的问题进行了研究，他们的研究以维吾尔族语音为输入，用 HMM 模型作为前端获取其语音特征，并构建了多终基于神经网络的自回归模型作为后端进行对比实验。测试结果表明，在以 HMM 模型为前端的前提下，用 BiLSTM 网络作为后端可以有效的提高合成后语音的自然度[13]。

王瑞针对汉藏双语的语音转换语音质量不佳，自然度不高的问题做出了研究，他在使用 HMM 模型

结合 DNN 网络完成语音转换目的的基础上使用了一种时延神经网络（TDNN）来训练模型提取汉藏双语的特征。并分别通过语音识别、机器翻译、语音合成、音色转换四个项目来完成语音转换任务，其中语音识别部分采用了深度全序列卷积神经网络（DFCNN）结合连接时序分类（CTC）的方法来实现。以有注意力机制的序列到序列模型实现机器翻译的功能。在语音合成方面则结合使用了 Tacotron 模型和 Fastspeech 模型。最后再使用了 StarGAN 的方法实现了音色的转换。此种方法相较于原来的双语转换的质量有了明显的提高[14]。

以下为语音转化的算法研究。

谭智元使用了自编码器的算法实现语音转换，其提出的方法的最大特点是可以实现零样本语音转换，这种方法不需要将后续需要转化的源说话人以及目标说话人的语言，只需要大量收集非平行语料，在训练过程中，算法会捕捉每个音频的身份编码，在后续使用转换功能时，模型也会捕捉源说话人及目标说话人的身份编码并对其进行转换从而实现目标[15]。

康筱在研究语音转换系统的过程中使用了变分自编码器的算法，他使用非平行语料作为输入，将源说话人的声音改变为目标说话人声音。这个过程中，他发现由于缺乏语言监督，内容编码器很难将纯净的内容信息提取出来，这会导致后续数据在声学特征的分析中混乱，从而使转换出的语言效果有较大噪声且质量不高[16]。

李涛认为当前多数语音转换的任务都需要使用平行语料，而平行语料在实际场景中的获取难度很大，因此他认为在画风迁移方面较为出色的 CycleGAN 算法能够在使用非平行语料的前提下相较于其他算法拥有更大的优势，他针对性的修改了算法的网络和损失函数，将该方法的实验结果与使用平行语料的 GMM 算法进行对比，该算法虽然在质量与相似度上均弱于 GMM 算法，但是考虑到该算法使用的是非平行语料，而 GMM 算法使用的是平行语料，说明了 CycleGAN 算法在语音转换方面确实是可行的，不过依旧有待改进[17]。

朱雅楠采用了基于表示分离生成对抗网络的语音转换算法，在该文章中，作者使用了实例归一化来使语音语义信息与说话人特征分离，再使用解码器对需要转换的语音进行语义提取后再与目标说话人特征结合，从而达成了语音的转换。该实验发现在归一化处理后提取出的语音语义仍会包含部分说话人的语音特征，便引入了矢量量化技术，提出了基于矢量量化表示分离生成对抗网络。进一步分离语音的语义信息与说话人个性特征。实验表明,使用矢量量化技术后,得到的语音语义信息的说话人分类准确度显著降低,并且转换后语音质量也有所提升[18]。

于杰发现大多数情感语音转换的方法都更加注重于对频谱特征进行转换,而对于基频特征,仅通过对数高斯归一化函数进行转换。他以 Star GAN 模型为基础，对这一问题进行研究和改进。该作者先是提出一种基于 StyleGAN-EVC 模型的情感语音转换方法，用情感风格编码器提取语音的情感风格特征，这一方法相比于使用 Star GAN 模型中的 one-hot 向量能够体现出更加丰富的情感信息。在这一基础上再使用自适应实例归一化，使情感特征与语义相结合。为了进一步增强情感饱和度，,该作者还提出一种基于基频差异补偿的 StyleGAN-EVC 模型对语音进行转换。其研究表明，在开集情况下，在不损伤语音转换质量的情况下通过使用情感风格编码器和基频差异补偿向量可以增加情感的饱和度[19]。

邱祥天针对 Star GAN 在语音转换中语音的质量以及个性的相似度等几个方面做出了优化，首先他将 ESR 模型添加到生成器网络中提取语音频谱的深层特征，并将提取出的深层特征于生成器网络中提取到的浅层特征相结合，这种做法能够明显改善转换后的语音质量。再此基础上为了改进生成器网络模型还添加了 DSNet，这能够在改善语音质量的基础上提高了转换后语音的个性相似度。除此之外，他还去除了 StarGAN 中的 one-hot 标签，转而使用风格编码器来代替此标签以达到提取说话人声音特征的目

的，在优化了损失函数后，该网络转换出的语音个性相似度得到了进一步提高[20]。

戴少梁针对跨语种语音转换的开集问题和算法运行效率两方面进行探讨并提出一系列改进工作，他提出基于激活指导的跨语种语音转换模型，该模型采用 U 型连接的编码器-解码器结构，通过实例归一化和激活指导来实现语音语义和说话人声音特征的分离，且这种方法不受语种和说话人数量的限制。其中实例归一化作为分离语音语义与人声特征的工具，激活指导则可以突破语种和说话人数的限制。另外，为了提升转换效率，该作者还在上述基础上提出基于激活指导和内卷积的跨语种语音转换模型。使用内卷积部分代替标准卷积,减少原模型的参数量和计算量,提升了算法的运行效率。实验仿真结果说明，该种策略在可以实现跨语种、多说话人的语音转换能够取得较好转换效果的前提下，还能够有较快的运行速率[21]。

徐伶俐针对目前语音转换的主流研究方向的难以实现跨语种转换、需要大量数据支撑、为了提高质量导致网络结构复杂从而对设备产生极高的要求三个问题展开了研究。对跨语种的转换问题，她将两个编码器进行结合，两个不同的编码器分别对应一种语言，对输入的声音特征进行分析编码，后续再对其进行表征并解码，达到跨语种转换的目的。对于需要大量数据作为训练支撑的问题，模型中的说话人编码器可以动态编码说话人信息,通过解耦思想表征说话人信息，可以使模型再训练阶段不需要说话人的标签作为辅助。为了降低对设备性能的需求，她还使用了深度可分离的卷积来替换掉模型中原本的常规卷积，使其总体训练时长减少了 26%[22]。

胡雨婷认为当前对语音转换的研究主要集中于干净无噪声的环境下，但大多数现实情况是用户无法避免噪声的干扰，因此她提出了在噪声环境下基于深度卷积网络的多模态语音转换模型。这种模型通过两个卷积神经网络分别提取唇部图片特征以及语音的特征，将其融合并送入全连接层来建立说话人视听特征与声学特征的对应。这项研究在测试时构造了 42 种噪音环境，证明了该种方法确实能够降低环境噪音对输出的影响[23]。

1.3 本章总结

上述编码器中，考虑到 wavenet 声码器声码器结构复杂，效率低下，SampleRNN 会浪费计算资源，Parallel Wavenet 需要严格的训练网络，以上几个声码器将不会在本文中用于合成语音，另外考虑到 WORLD 声码器拥有较好的鲁棒性，在提取基频及分析声学特征时也便于操作的优点，本文将会使用 WORLD 声码器达到将声音特征转换为语音的目的。

上述算法中，考虑到众多语音合成方面的研究难以输出拥有感情饱和度的音频，虽然其中有方法能够输出具有情感特征的音频，但其复杂度过高，在文中难以再现，因此不再考虑通过语音到文本，文本到语音的形式来完成任务。传统语音转换的方法中，多数方法都对语料有严格的要求，虽然部分方法中能够实现零样本的语音转换，但其实现过于困难，因此本文将采用较为简单的 CycleGAN 算法实现语音转换功能并对其进行研究。

第二章 语音转换基础

本章介绍了在语音转换任务中比较重要的三个声音成分，并初步介绍了 WORLD 声码器。

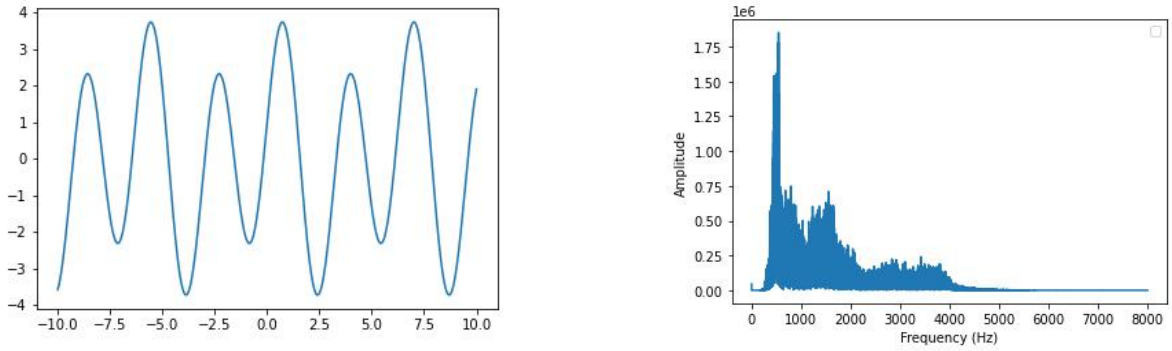
2.1 语音的特征信号

在语音转换这项技术中，主要研究的语音特征信号是语音的声学特征，这部分的特征是由基频、频谱包络等信号数据中提取出来的，下文中将会对这部分信号进行介绍

2.1.1 基频

语音信号可以被看作是复杂的音频信号，是由许多频率的成分组合而成的，其中最低的频率分量被认为是这段音频信号的基频。基频的频率分量取决于发声源的振动频率，当音频信号为一段语音信号时，基频也就可以被理解为是说话人说话时，声带的振动频率，可以用于判断说话人的音高。基频的转换可以通过对数高斯归一化完成。

声音的基频是会随着时间变动的，所以提取基频的方法一般是先把声音分帧，再在逐一的计算每一帧的基频。这其中的提取方法被分为时域法和频域法两类。时域法是将声音信号的波形图输入，波形的最小周期就是该语音的基频。频域法则是寻找频谱图中的峰值，基频的整数倍数在频谱图上会以峰值的形式表现出来，如图 2-1。



（a） 时域图　　（b） 频域图

图 2-1 声音的时域及频域图

基频的提取需要先计算出音频信号的自相关函数，自相关函数的公式如下：

$$C(t) = \sum[x(n) * x(n-t)] \tag{2.1}$$

自相关函数的计算方法是将信号与自身在不同时间点的移位进行乘积运算，并将结果求和。音频的自相关函数可以体现出信号的重复性和周期性，当音频信号本身具有周期性时，自相关函数也会体现出周期性的特征，因此可以用自相关函数来求取音频信号的基频。基频的取值为自相关函数最大值所对应横坐标，即：

$$f_0=\text{argmax}(C(t)) \tag{2.2}$$

2.1.2 频谱包络

频谱包络信息可以体现出一段音频在随着时间变化的过程中，其幅度的整体变化趋势，是音频整体能量变化的一种体现。能够在一定程度上体现出音频的音量以及音调。除此之外，人在说话时，每个发

音的特征都能够在频谱中表现出来，例如元音的发音是由频谱中的前三个共振峰来决定的。在语音信号中，频谱包络通常被认为是声道特性的一种反映，它可以反映出声音在声道中被放大或衰减的情况。因此在语音转换的技术中，对频谱包络的分析是至关重要的。

频谱包络线实际上是频谱图上每个频率中的最高点连接而成的线，频谱包络线是紧贴着频域图的平滑曲线。

频谱包络的提取的公式如下：

$$H(f)=X(f)/G(f) \tag{2.3}$$

式中，H(f)为频谱包络，X(f)为音频信号经过傅里叶变换处理后获得的频域信号，G(f)为 X(f)经过高斯滤波处理后获得的平滑曲线，将 X(f)和 G(f)作商，即可获得频谱包络。

频谱包络可以用来提取 Mel 系数，这可以更好地模拟人耳对声音的感知。人耳对低频率的声音感知能力较好，而对高频则分辨能力较差，Mel 的非线性尺度使声音更好的被表示出来。同时可以提高音频特征的匹配度，增加鲁棒性。Mel 频率的求取公式如下：

$$f_{Mel} = 2529 * \log(1 + f/700) \tag{2.4}$$

其中 f 为频率，f_{Mel} 为 Mel 频率。

提取 Mel 倒谱系数特征向量的流程图如下：

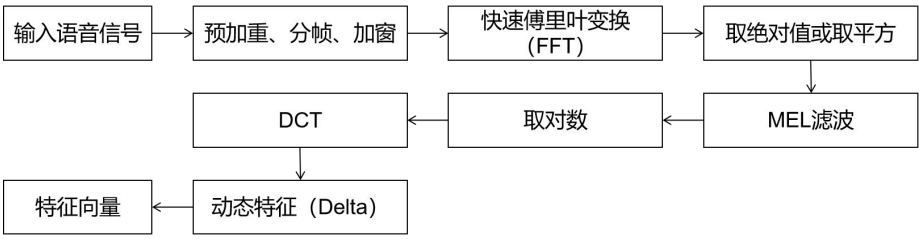


图 2-2 Mel 倒谱系数提取

其中 DCT 指的是 DCT 变换，它的全称是离散余弦变换，离散余弦变换相当于一个长度大概是它两倍的离散傅里叶变换，这个离散傅里叶变换是对一个实偶函数进行的。其含义是使信号函数成为偶函数，去掉频谱函数的虚部。动态特征步骤则是因为标准的倒谱参数 MFCC 只反映了语音参数的静态特性，语音的动态特性可以用这些静态特征的差分谱来描述。

2.1.3 非周期信号

非周期信号中有着说话人共振峰和嘶音的特征。共振腔体的大小、形状和口型的不同，共振峰都会有所变化，因此不同的说话人会有着不同的共振峰，这也是决定音频音色和质量的因素之一。嘶音是说话人在发出清辅音和浊辅音是会产生的，对于音频信号的识别和理解较为重要，除此之外，非周期信号还可以用于转换过程中时域和频域信号的对齐。

非周期信号的值取决与当前频率及临近的两个频率所对应的值，在三个值中取最小值作为当前频率的非周期信号值，具体公式如下：

$$D4C(f)=\min\{X(f-1),X(f),X(f+1)\} \tag{2.5}$$

2.2 WORLD 声码器

WORLD 声码器在本文的实现中将作为语音信号特征的提取、分析工具和将特征信号合成为音频信号的工具。如 1.2.1 中所说，WORLD 有着合成语音质量高、自然度好，能够更好地处理高音、噪声等优点，除此之外，WORLD 声码器还可以将音频信号分解为基频、频谱和非周期三个成分，WORLD 声码器在文中功能实现中的作用如图 2-3。

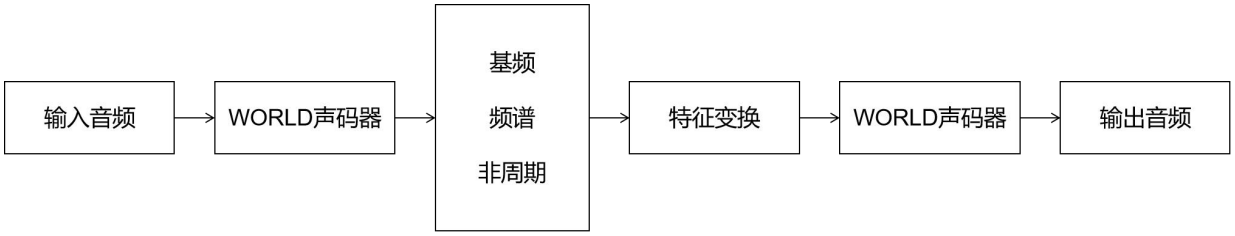


图 2-3 WORLD 声码器在系统中的运用

其中的将变换后的特征输入声码器后，声码器会将变换过后的基频和 Mel 频谱信号以及原非周期信号相乘从而形成新的音频信号，其公式如下：

$$s(t)=P(t)*E(t)*N(t) \tag{2.6}$$

其中，s(t)为音频信号，P(t)为变换后的基频信号，E(t)为变换过后的 Mel 频谱信号，N(t)为原始的非周期信号。

2.3 本章总结

本章对音频信号的特征提取作了介绍，并且说明了在不使用模型的前提下该如何对基频的特征进行改变，另外还简单的说明了 WOELD 声码器在该系统中运用在哪些部分，如何实现从声音特征信号到音频信号的转换。

第三章 深度学习相关介绍

本章初步介绍了深度学习及其几种算法，重点介绍了 GAN 及其分支 CycleGAN 的工作流程及其对其损失函数的理解。

3.1 深度学习

深度学习是机器学习技术的特殊分支，以将机器的计算能力与神经网络中的连接模式相结合的方法，构建出分层的人工神经网络，再从原始数据中抽象出数据特征，并通过隐含层学习样本数据之间的内在关系，这种学习会使其在之后的场景中对学习到的关系进行映射，这种映射能够用于对数据进行预测。

相比与机器学习，深度学习拥有有更强的拟合能力，在不对数据进行过多预处理且未设计复杂的特征提取器的前提之下，构建出的神经网络依旧能够学到不同层次的特征，我们往往将高层次的特征看作使低层次特征的抽象，而高层次特征比起手动的提取低层次特征能够拥有更好的代表特征的能力以及鲁棒性。

目前较为主流的深度学习模型有以下几种：CNN、DBN、RNN、自动编码器、GAN 。

其中卷积神经网络 CNN 一般被用于图像处理、图像辨别与分类，深度置信网路 DBN 适用于高维具有不确定性数据的分类，递归神经网络 RNN 适合运用于处理时间序列的问题，自动编码器能将高维信息以低微方式表现出来，常用于领域适应方面的研究，生成对抗网络 GAN 有无监督、生成质量高的特点，常被用于图像生成、画风迁移任务领域。

如 1.2.2 中所说，本文将使用 CycleGAN 方法实现语音特征的转换任务。在介绍 CycleGAN 之前，本文将先介绍 GAN。

3.2 GAN 网络

生成对抗网络主要有生成器和判别器两部分组成，生成器不断的通过函数输出样本数据，而判别器则会将输出的样本数据和真实数据进行比对判别真假，生成器和判别器在这个过程中会不断的优化和提升能力，最后达到平衡完成训练。下图为生成对抗网络的训练流程图。

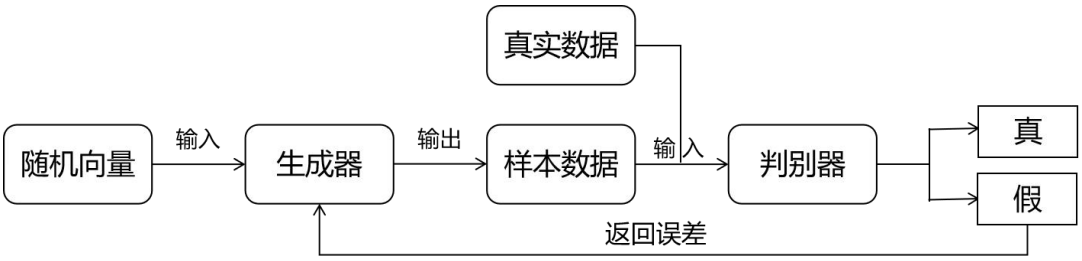


图 3-1　GAN 网络训练流程

以下是 GAN 网络的目标函数，同时也是 GAN 网络的对抗损失函数：

$$\min \max V(D,G) = E_{x\sim P_{data}}[\log D(x)] + E_{Z\sim P_{Z(z)}}[\log(1 - D(G(z)))] \tag{3.1}$$

其中的 min 是对生成器 G 而言的目标，在训练过程中，G 的目标函数值越小。此公式的解读如下，其中 V 代表的是该网络的交叉熵，同时对于生成器 G 来说是损失，其值越小越好，对于判别器 D 而言则是一个类似与激励函数的存在，其值越大越好。在此损失函数中，所有 log 底数均为 e。$E_{x\sim P_{data}}[logD(x)]$的含义为，当所有 x 都为真实的目标声音特征 x 时，[logD(x)]的期望，这是在动态训练的角度上写出的式子，但实际上在单次训练中或者在训练结束后，生成网络和判别网络的映射固定时，该期望就等同于均值。$E_{Z\sim P_{Z(z)}}[\log(1-D(G(z)))]$同理，其中 z 代表的是为了生成出虚假的声音特征所生成出的随机数输入。其中 D(x)代表的是，当真实数据被输入判别器后，判别器的输出为真的概率，而 D(G(z))则代表当虚假的数据被输入判别器后，判别器输出为真的概率。因为判别器的限制，两边的 log 实际上的定义域只在(0,1]。

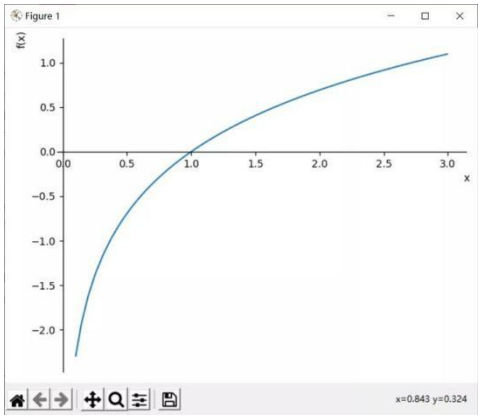


图 3-2　log 以 e 为底的曲线

那么从判别器 D 的角度来考虑，最好的结果是判别器能 100%的识别真假数据，也就是 D（x）输出为 1，而 D(G(z))输出为 0，那么对于交叉熵 V 而言就能够达到最大值 0。此时若将值带入公式：

$$V(D,G)=E_{x\sim P_{data(x)}}[\log 1]+E_{Z\sim P_{Z(z)}}[\log(1-0)]$$
$$=0 \tag{3.2}$$

而对于生成器而已，其最好的结果是生成出的虚假数据能够使判别器无法判断为假，即 D（G（z））的输出为 1，而 D（x）的输出则为一个常数 a，不论 a 取何值，定义域都在(0,1]，loga 始终＜0。那么此时，交叉熵将取到最小值-∞。若将值带入公式：

$$V(D,G)=E_{x\sim P_{data(x)}}[\log a]+E_{Z\sim P_{Z(z)}}[\log(1-1)]$$
$$=-\infty \tag{3.3}$$

3.3 CycleGAN 网络

CycleGAN 的核心是 Cycle，即让输入音频与输出音频之间实现循环，因此与传统 GAN 不同的是，CycleGAN 需要创建两对生成器和判别器从而来实现输入与输出的循环，将两组不同风格的语音随机取一条做配对，作为训练神经网络的输入，训练的结果在理论上，除了音色有所改变，内容也许也会跟着有所变化，因此需要将产品与素材进行损失对比，循环训练的要求即是降低这个损失，从而达到模型能够只改变音色而不改变内容的效果。下图为 CycleGAN 网络的训练流程图：

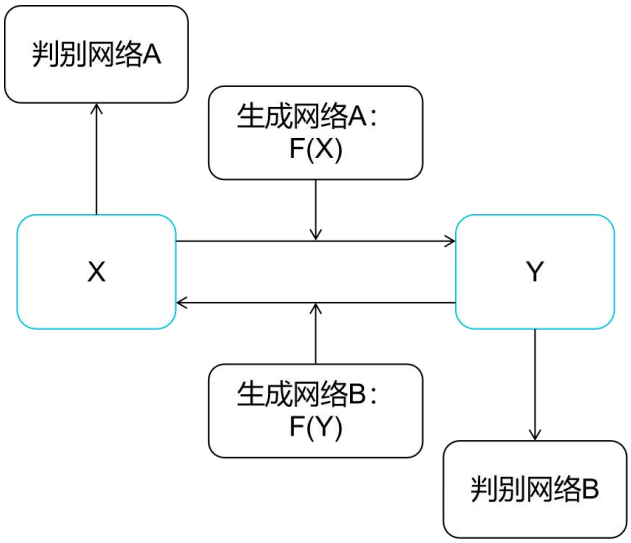


图 3-3 CycleGAN 网络训练流程

图中的 X 为源说话人音频的声音特征，Y 为目标说话人的声音特征，理论上的训练流程是从 X 源说话人音频的声音特征开始，通过生成网络 A 的映射使源说话人的声音特征 X 变为合成出的目标说话人声音特征 Y'，随后判别网络 B 会对生成出的目标说话人声音特征 Y'进行判断是否符合数据集中目标说话人声音特征 Y，并为生成网络 A 返回 Loss。在这之后网络会将合成出的目标说话人声音特征 Y'通过生成网络 B 合成出源说话人的声音特征 X'，并给判别网络 A 进行判断并为生成网络 B 返回 Loss。

但是在具体实施过程中，流程并不会像上述理论一样形成一个循环往复的 Cycle，而是 X 通过生成网络 A 生成出 Y'，Y'再到 X'。或者，Y 通过生成网络 B 生成出 X'，X'再到 Y'。即实际上训练的进行是一个个单独的 Cycle。

此外 CycleGAN 网络除去和 GAN 网络的几乎一致对抗损失函数之外，还有两个损失函数，分别是循环不变性损失（Cycle-consistency loss）以及映射一致性损失（Identity-mapping loss）。

(1)对抗损失：该损失用于训练生成器，使得生成器能够生成与目标说话人类别相同但声音特征与源说话人不同的声音。该损失由判别器的输出计算得到，目标是最小化生成器和判别器之间的损失。公式如下：

$$L_{adv}(G_{X\to Y},D_Y)=E_{y\sim Pdata(y)}[\log D_B(y)] + E_{x\sim Pdata(x)}[\log(1-D_B(G(x)))] \tag{3.4}$$

此处的对抗损失与 3.2 中提到的目标函数理论是一致的，不过 CycleGAN 有两对判别网络和生成网络，所以实际上的对抗损失函数是两条对称的式子，上述的这条式子是用于训练判别网络 D_B 和生成网络 A 的。其中与 3.2 所提的不同的地方在于，原本的 x 为目标的声音特征，而此处的目标声音特征为 y，而原本用于输入生成数据的随机数 z 在此处为源说话人的声音特征 x。

(2)循环一致性损失：该损失用于保持循环一致性，即保证声音转换的可逆性，使得经过循环转换后的声音能够与原始声音尽可能地接近。该损失通过计算原始声音和经过循环转换后的声音之间的近似 MSE 损失计算得到。公式如下：

$$\begin{aligned} L_{Cyc}(G_{X\to Y},G_{Y\to X}) = {} & E_{x\sim Pdata(x)}[\|G_{Y\to X}(G_{X\to Y}(x))-x\|_1] \\ & + E_{y\sim Pdata(y)}[\|G_{X\to Y}(G_{Y\to X}(y))-y\|_1] \end{aligned} \tag{3.5}$$

这段式子的中，$G_{X\to Y}(x)$的含义为，将源说话人的声音特征输入到 X→Y 的生成器中，生成出虚假的目标说话人声音特征 y'，因此 $G_{Y\to X}(G_{X\to Y}(x))$的含义就是将生成出的虚假目标说话人声音特征再输入

到 Y→X 的生成器，生成出虚假的源说话人声音特征 x'。同理 $G_{X\to Y}(G_{Y\to X}(y))$的含义就是 y→x'→y'的过程。这个式子将生成出的虚假声音特征与源声音特征一一做差并取绝对值相加并求取期望，在模型固定或在单词训练中，求取期望等同于求取均值。之所以说是近似 MSE 损失是因为 MSE 损失的计算方法是求平方差的和并取均值，而此处没有作平方处理而是取出其绝对值。

这个公式的用途是使循环生成的虚假声音特征和源声音特征间会保持一部分内容，而此内容就是声音转换任务中要求不作改变的说话内容，因此该损失是该项目中的重点。

(3)身份映射损失：该损失用于保持说话人的身份特征不变，使得源说话人的声音特征保持不变。该损失通过计算源说话人的声音和经过源到目标再到源的循环转换后的声音之间的近似 MSE 损失计算得到。公式如下：

$$L_{id}(G_{X\to Y},G_{Y\to X}) = E_{y\sim PY(y)}[\|G_{X\to Y}(y)-y\|_1]+E_{x\sim PX(x)}[\|G_{Y\to X}(x)-x\|_1] \quad (3.6)$$

该公式的与循环一致性损失的公式结构一致，其中 $G_{X\to Y}(y)$的含义是将目标声音特征 y 放入 X→Y 的生成器中，$G_{Y\to X}(x)$含义是将目标声音特征 x 放入 Y→X 的生成器中，并将生成的虚假声音特征与其原特征一一做差取绝对值求均值。

这么作的意义是，当输入的声音特征原本就为目标的声音特征时，生成器不应该再对输入的目标声音特征进行改变，这是许多时候都会被忽略的一个损失。

3.4 本章总结

本章初步介绍了深度学习、GAN 网络的训练流程和其损失函数，重点介绍了 CycleGAN 的训练流程以及其三个损失函数，分析了损失函数的计算方法及其意义。

第四章 语音转换的实现

这个章节介绍了从录制音频文件开始到预处理音频形成训练集，再到构建生成网络判别网络与将数据输入网络的过程。

4.1 音频的采集

该项目使用了自制的训练和测试集，对于音频的采集流程如图 4-1：

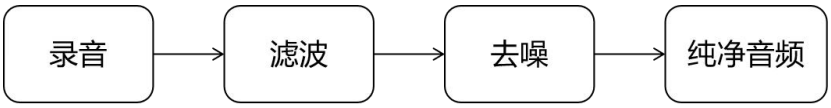


图 4-1 音频的采集流程

采集的要求如下：

1.音频的采集格式统一为 16000hz，16bit 位深度。其原因是这个格式的音频拥有较高的保真度，能够保证语音转换任务不会因为数据集的质量过低而影响语音转换的实现，在此前提之下能够缩小数据占用的空间、在训练时也能够降低运算的复杂度提高训练速度，此外 16000hz 是比较常用的音频采样频率支持的平台和设备广泛。

2.录音时采用内容丰富的素材，避免使用古诗词等素材。本文采用的朗读素材是高中政治教材中的知识点，相比于古诗词其内容更加密集，可以把有效的计算资源利用起来，若是使用古诗词作为素材，其中大段的空白内容会浪费资源，且同时可能导致输入的数据量过少而影响训练的结果。

3.在尽量安静的环境下进行录制，减少噪音的复杂程度，方便后续对音频进行处理。

4.在采集完成后要对音频进行滤波和去噪操作，以免录制过程中过多的噪音导致训练结果不佳。

5.说话人定为两位男性。跨性别的语音转换理论上也可以实现，但是转换后的音频质量会因此有所降低。

4.2 音频的预处理

从输入音频开始，提取出音频信号的基频、频谱包络和非周期信号，求取出基频的平均值和标准差并以.npz 的格式进行保存方便后续调用，于此同时会对频谱包络进行压缩并将其转换为 Mel 系数，即图中的 Mel 系数，对 Mel 系数进行标准化处理并以.npz 格式进行保存，最后将源说话人和目标说话人的 Mel 系数分别保存为.pickel 格式的文件。下图为预处理的流程图：

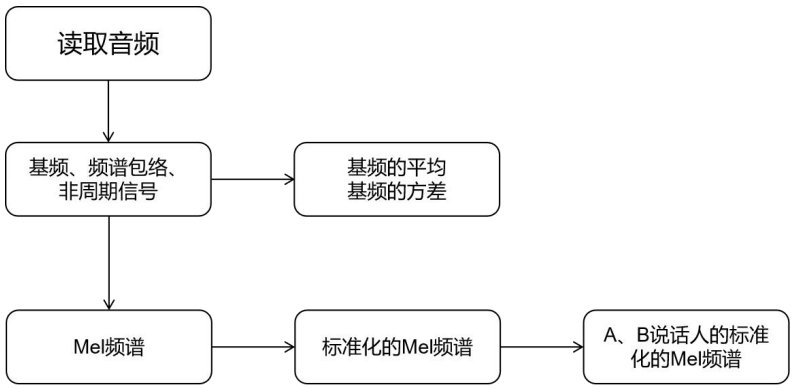


图 4-3 音频预处理流程

这其中读取音频的操作是依靠 soundfile.read()方法实现的。对 f0、sp、ap 的提取则在 2.1 的各小章节介绍的方法。基频的平均值和标准差可以通过直接建立数学公式进行计算，也可以调用 mean()和 std()方法对列表直接进行计算，本文中使用的是后者。提取 Mel 系数也使用到了 WORLD 声码器中的 pyworld.code_spectral_envelope()方法。对 Mel 系数的标准化是先计算出 Mel 系数的平均值和标准差，再使用 Mel 值减去平均值，以结果除以标准差，公式如下：

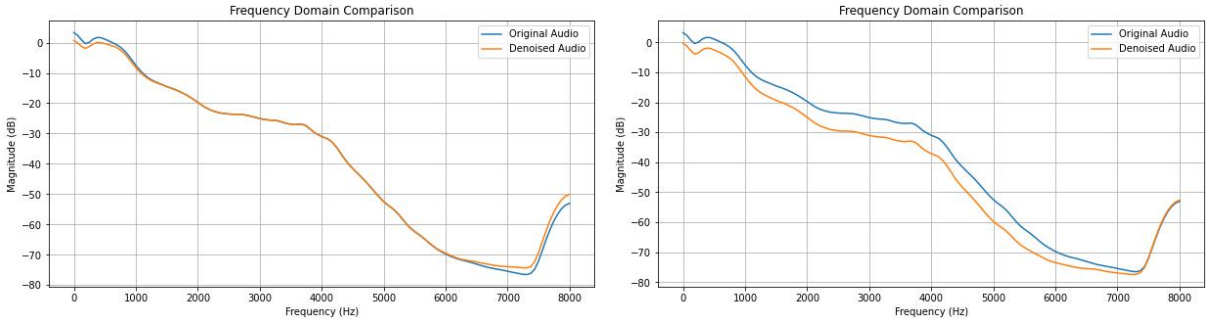
$$coded\_sps\_normalized=\frac{Mel-Mel\_mean}{Mel\_std} \tag{4.1}$$

其中 coded_sps_normalized 是标准化的 Mel 频谱，Mel 代表的是 Mel 频谱本身，Mel_mean 指的是 Mel 频谱的平均值，Mel_std 是 Mel 频谱的标准差。最后再对源说话人和目标说话人的标准化后的 Mel 系数分别进行保存。

4.2.1 音频的滤波及去噪

考虑到在安静环境中，高频噪声较少，所以此处录制出的音频中的噪音大多为低频噪音，因此设计一个 4 阶高通滤波器。过滤去频率低于 300Hz 的部分。此处使用的是括巴特沃斯（Butterworth）滤波器，其原理是将将输入的音频信号与一组滤波器系数进行卷积运算，得到输出音频信号。在使用 Butterworth 滤波器时可以选择滤波器的阶数，阶数的选择会影响滤波器的过滤带，若阶数较高则过渡带较窄，阶数低则过渡带宽。有过渡带一说是因为高频滤波器的实现并不是简单的将低于阈值的频率信号全部裁剪掉，而是对低频的声音信号进行减弱，使低于阈值的声音信号接近于 0。而过渡带就是在对低频声音信号进行减弱时，减弱的程度未达到最大值的那一段频率。即阶数越大，对声音信号的减弱处理就越精准，但同时，越高的阶数也会造成越大的延迟。

此处的去噪方法为非线性谱减法，其优势是在比较稳定的噪声环境下能够有效的去除噪声对音频信号的影响且不会对音频信号本身的频率成分造成损失和扭曲。但这种方法也有其劣势，非线性谱减法需要根据具体情况变换非线性参数、平滑参数和调整参数。在信噪比过低的情况下可能会在非线性处理的过程中增加伪声，在高噪音环境下可能会导致音频信号失真和畸变。非线性谱减的流程是：设定非线性、平滑、调整参数，读取噪声信号并估计功率谱，计算音频信号本身和噪音信号的幅度谱，采用非线性方法对音频信号进行幂次变换来保护音频的频率成分，用平滑方法消除由二值掩蔽带来的不连续性，用调整方法对平滑后的信号进行进一步的优化进一步降低噪音水平，将使用调整方法方法后得到的音频信号与先前计算出的音频信号幅度谱相乘得出增强语音幅度谱，通过增强语言幅度谱重建出去噪后的音频信号。图 4-2 为音频滤波及去噪前后的频谱对比：



（a）高通滤波前后频谱图　　（b）谱减去噪前后频谱图

图 4-2 滤波去噪处理前后的音频频谱

4.2.2 预处理生成文件

在音频预处理过程中，会形成两中类型的文件，一种为.npz 格式，另一种为.pickle 格式的文件，接下来将对这两种文件进行介绍。

.npz 格式的文件是 numpy 的一种保存格式，可保存一个或多个数组，并为每个数组命名，使用 numpy 库的函数可以直接加载读取被保存的数组。可以使用 numpy.savez()方法将多个定义了名字的数组保存在同一个文件中，后续使用时可以通过 numpy.load()方法直接把整个文件调用出来。示例代码如下：

```
np.savez(os.path.join('./data/', 'data.npz'),
                A=a,
                B=b)
x=np.load('./data/data.npz')
C=x['A']
D=x['B']
```

该段代码中将数组 a 和数组 b 分别命名为 A 和 B，并保存为在路径'./data/'下 data.npz 的文件，将文件中的数据保存在变量 x 中，并给 C 赋值文件中名为 A 的数组，即将数组 a 赋给 C，给 D 幅值文件中名为 B 的数组，即将数组 b 幅值给 D。

.pickle 文件是一种 Python 中的序列化文件格式，可以将 Python 对象（如列表、字典、类等）保存到文件中，以便后续在程序中进行加载和使用。pickle 模块是 Python 标准库中的一个模块，可以实现数据的序列化和反序列化。使用 pickle，可以将 Python 对象转换为字节流，然后将字节流写入文件，或者从文件中读取字节流并将其反序列化为 Python 对象。

```
save_pickle(variable=a,fileName='./data/A.pickel')
x=pickel.load('./data/A.pickel')
```

该段代码是将 a 变量的值保存为在./data/路径下的 A.pickel 文件，并将该文件中的内容读取到变量 x 中。

4.3 CycleGAN 网络的构建

CycleGAN 是一种特数的 GAN 网络，它拥有两对结构相同但用处不同的生成网络和判别网络，接下来将会分别介绍该项目中生成网络和判别网络的结构。CycleGAN 的网络结构如下：

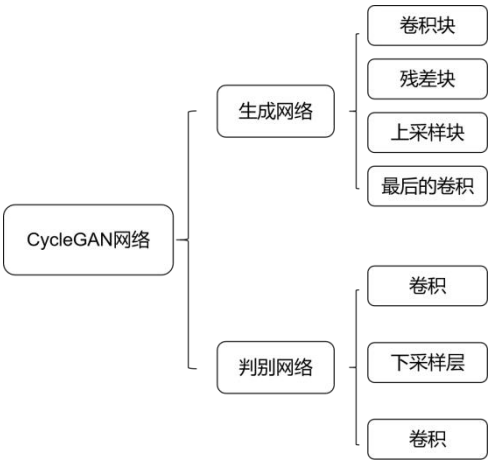


图 4-4 CycleGAN 网络结构

4.3.1 生成网络的构建

CycleGAN 的生成网络架构是一种自编码网络，输入一段音频，输出另一段音频。它由一个编码网络和一个解码网络组成。编码网络中包含能进行下采样的卷积层，编码网络中包换两个上采样块和一个最终的卷积层。总的来说，生成网络由以下几个块组成：卷积块、残差块、上采样块、最终的卷积层。

(1)卷积块：卷积块的结构如下：获取 Mel 频谱图输入、卷积、通过门控单元选通随后进入下采样环节。其中下采样包含两组卷积层、实例化、门控单元，随后对 Mel 频谱图作二维到一维的转换。

图 4-5 卷积块结构

此处以 Mel 频谱图作为输入，输入时其维度为[B,36,T]，其中 B 代表一次训练中输入的样本数量，36 代表在音频信号的每个时间步上使用 36 个 Mel 频率滤波器提取出的特征数量，T 表示在该步长下的时间步数。在进入第一个卷积层 conv 之前给数据作升维处理，使数据维度变为[B,1,36,T]，其中增加的维度用于表示通道数。这样做的好处是可以让神经网络更好地学习音频数据中不同通道之间的关系，提高模型的泛化能力和准确率。

进入卷积 conv 后，增加通道数使数据的维度由[B,1,36,T]变为[B,128,36,T]。

其中门控单元 GLU 以 Mel 频谱图和 gated_Mel 作为输入，用 gated_Mel 对 Mel 频谱图进行有选择的数据流通。也因此，卷积层中会存在两个相似的卷积，一个卷积输出 Mel 频谱图，另一个用于输出 gated_Mel。

下采样过程中为了减少数据的时间步数，从而降低模型的计算复杂度，同时保留数据的重要特征。数据维度经过两层下采样层从[B,128,36,T]变为[B,256,18,T/2]再变为[B,256,9,T/4]。

另外下采样层中的实例化步骤采用 instance normalization 的方法，这个方法与 batch normalization 相比较起来，它会对每一个声音样本都进行归一化处理，因此会更加适合处理小批量音频数据，可以使语音转换模型的训练效果以及泛化能力提高。

Mel 频谱图 2D→1D 的转化过程中合并了通道数与 Mel 特征数量，通过 reshape 操作，维度由[B,256,9,T/4]变为[B,2304,1,T/4]，再使用 squeeze()方法减少一个维度使之变为[B,2304,T/4]，再经过一次卷积核为 1×1 的 conv 层后 Mel 频谱图维度为[B,256,T/4]，从而达到 Mel 频谱图降维的目的。

(2)残差块：残差块是由 6 个残差连接层组成的，残差连接层的结构如下：卷积层、实例归一化层、门控单元、卷积层、实例归一化层、加法层。随后再对 Mel 频谱图作一维到二维的转换。

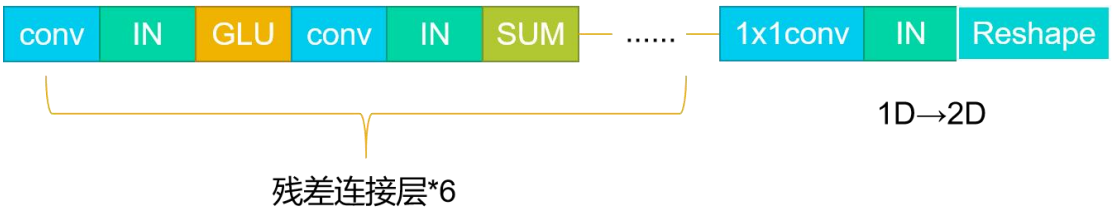


图 4-6 残差块结构

一个残差连接层中有 3 个卷积，其中两个是卷积层，一个卷积用于输出门控信号 gated_Mel，第一个卷积层输入为 256，输出为 512，第二个卷积层输入为 512，输出为 256，即在一个残差连接层内，Mel 频谱图的维度会从[B,256,T/4]变为[B,512,T/4]再变回[B,256,T/4]。但是若将残差连接层看做是一个整体，那么 Mel 频谱图的维度就可以看做是没有改变的。每个残差层最后通过加法层使输入输出的 Mel 频谱图相加，随后再进入下一个残差连接层，这样的做法可以使网络更容易学习到源说话人 Mel 频谱图到目标说话人 Mel 频谱图的映射关系。

残差块的最后依旧使用了卷积核为 1*1 的卷积层使 Mel 频谱图的维度由[B,256,T/4]变为[B,2304,T/4]，再通过 unsqueeze()方法增加一个维度变为[B,2304,1,T/4]，再通过 reshape 操作变为[B,256,9,T/4]，使 Mel 频谱图维度转换回二维。

(3)上采样块：上采样块由两个上采样层组成，上采样层结构如下：卷积层、PS 层、实例归一化层、门控单元。

图 4-7 上采样块结构

上采样过程中第一次通过 conv 层使 Mel 频谱图维度由[B,256,9,T/4]到[B,1024,9,T/4]再通过 PixelShuffle 操作对数据进行重排列，将通道数减少，增加数据的高和宽，从而达到将低分辨率的语音信号映射到高分辨率的语音信号，Mel 频谱图维度因此改变为[B,256,18,T/2]，再一次通过 conv 层 Mel 频谱图维度变为[B,512,18,T/2]，PixelShuffle 操作使之变为[B,128,36,T]。

(4)最后的卷积层：该部分只有一个卷积层，通过这个卷积层 Mel 频谱图维度由[B,128,36,T]变为[B,1,36,T]，这个卷积层的作用就是将之前一系列操作生成的由语音信号各种特征组成的特征图通过卷积操作将其转换为音频信号的样本。随后还会进行一次 squeeze()操作使 Mel 频谱图维度变为[B,36,T]这个维度与输入时相同，可以直接输出。

图 4-8 最后的卷积层

4.3.2 判别网络的构建

相比于生成网络架构，判别网络架构的的组成要简单很多，从输入开始，它由一个卷积层、一个门控单元、四层下采样层、一层卷积层，最后输出判断。其中下采样层由卷积层、实例归一化层以及门控单元组成。

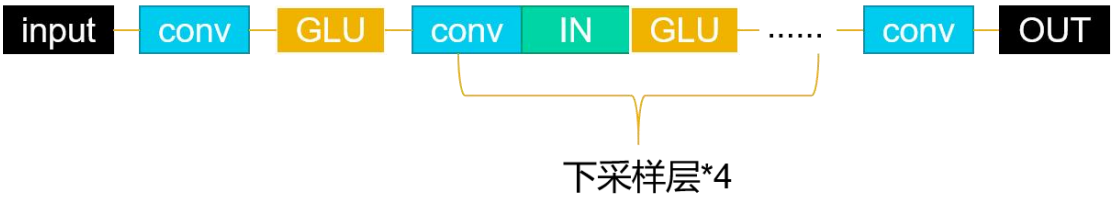


图 4-9 判别网络架构

此处输入时 Mel 频谱图维度为[B,36,T]，经过 unsqueeze()操作增加通道维度为[B,1,36,T]。通过 conv 操作将通道数增加为[B,128,36,T]。四层下采样层中每次进入 conv 层分别作以下变化，第一层中增加通道数，减少 Mel 特征数量与时间步数为[B,256,18,T/2]，第二层同上使维度变为[B,512,9,T/4]，第三层同上，维度为[B,1024,5,T/8]，这其中 Mel 特征数量无法整除 2，因此取 5，第四层不对数据维度作改变。通过最后一个 conv 减少通道数为[B,1,5,T/8]并输出。

4.4 本章总结

本章介绍了如何建立数据集，即音频的采集和预处理流程，还介绍了转换 Mel 频谱所需的 CycleGAN 网络的框架和搭建流程，除此之外还介绍了 Mel 频谱在进入 CycleGAN 框架中时，它的维度变动及其意义，这部分是实现语音转换功能的核心部分。

第五章 界面设计与问题的分析处理

本章介绍了对语音转换系统的界面设计及对转换后的音频音质不高的处理与研究方法，最后得出结论。

5.1 界面的设计

本文使用了 PySimpleGUI 实现了简单的人机交互界面，分别设计了登录，注册、主界面以及模型训练界面。界面设计流程图如 5-1：

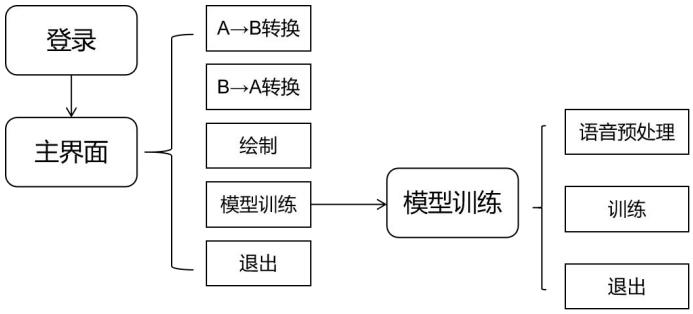


图 5-1 界面流程图

此流程图中，圆角矩形框为界面，矩形框为功能性按钮。图中不包含注册界面，其原因是考虑到该系统的注册可能需要受到管理，因此注册界面并没有在任何界面中设置跳转的入口。图 5-2 为登录界面与注册界面：

（a） 登录界面

（b） 注册界面

图 5-2 登录界面和注册界面

本文的系统需要先在注册页面进行注册，点击注册界面的确定按钮后，填进输入框中的内容就会被读取并输入到通过 sqlite3 库创建的 user.db 数据库中的 user 表中。在登录界面中登录时，会在 user.db 数据库的 user 表中进行查询，是否有一项内容的账号与密码与被填入输入框中的数据完全相同，若有则登录成功，跳转至主界面，若没有与之匹配的项则会跳出账号或密码错误的提示框，点击确定后依旧会在登录界面，但是账号与密码的输入框会被清空。另外，登录与注册界面的取消按钮在点击后都会直接将页面关闭。

图 5-3 为登录成功后显示的主界面：

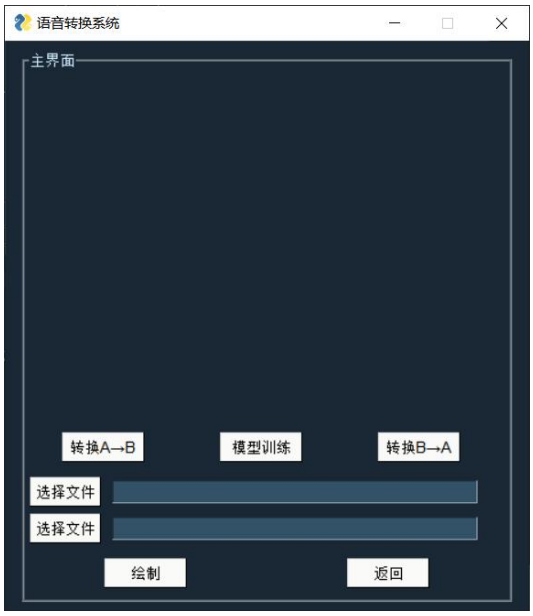


图 5-3 主界面

在该主界面中，除了两个选择文件的按钮框，当光标在每个按钮框上停留时都会弹出悬浮提示框，对该按钮的功能进行提示。当鼠标停留在“转换 A→B”按钮上时，会提示要将需要转换的音频放在该文件的 A 文件夹下，生成的转换后音频会出现在 turnedA 文件夹下。“转换 B→A”按钮同理。当光标停留在“模型训练”按钮上时会提示接下来界面将会跳转至模型训练界面。在“绘制”按钮上停留时，会提示将上方选择的两个文件的频域图进行绘制，方便用户对转换前后的音频进行对比。在“退出”按钮上停留时则会提示该按钮会返回登录界面。图 5-4 为使用绘制功能绘制出一对转换前后音频的频域图：

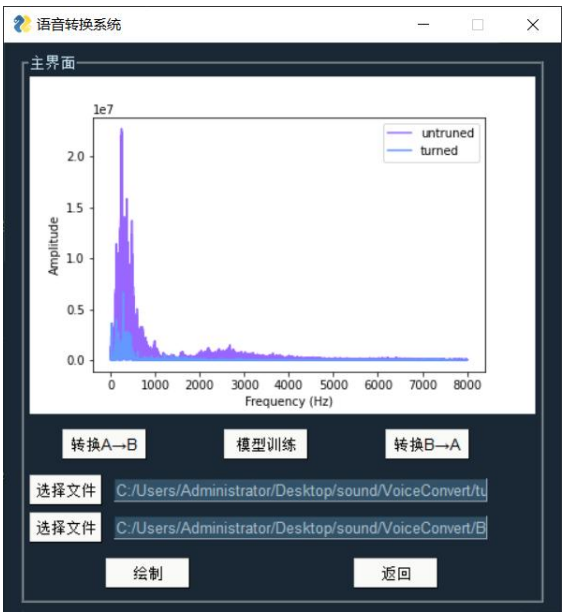


图 5-4 绘制功能的表现

点击主页面中的“模型训练”按钮会跳转至模型训练的界面，图 5-5 为模型训练界面：

图 5-5 模型训练界面

在该界面中，两个文件夹选择按钮会要求使用者指定源说话人和目标说话人的音频文件夹，当鼠标停留在“语音预处理”按钮上时，便会提示需要将对选择的文件夹中的音频进行预处理，并将结果保存在哪个文件夹中。按下该按钮后，系统将会对音频文件进行如本文章节 4.2 中的预处理操作，并在操作完成时跳出操作成功的弹窗。当在“训练”按钮上停留时，将会提示该过程需要大量时间，训练时间越久模型的最后体现出的效果就越好，因为训练过程中不会主动中断，所以此处在点击后不会跳出弹窗。停留在“返回”按钮后，将会提示点击该按钮后将会退回至主界面。

5.2 问题的分析处理

为了确保该系统的语音转换功能确实实现了声音特征的改变，本文计算了转换后与目标的音高与音调并进行对比，音高的计算需要先对音频信号进行傅里叶变换，获取其中的峰值位置与峰值，对峰值进行排列并以最高峰值对应的频率为音高。

音调的计算需要先对音频信号进行傅里叶变化，并取其中位数作为代表该段声音频率的特征，但因为中位数并不能完全反映音调，所以此处还使用了半音容差进行弥补，最后再判断该段音频信号的基准音符，音调计算的公式如下：

$$tune=med*tol+ freqs[med] \tag{5.1}$$

这个式子中，tune 就是所需的音调，med 是在频谱中，中位数所在的横坐标，tol 是半音容差是一个常数，具体的值是$2^{\frac{1}{12}}$，freqs[med]就是中位数的值。

若转换后音频的音高与音调的差距小于 5%，那么就认为确实达到了语音转换的目的，本文的测试选择了男 A→男 B、男 C→男 D、男 A→女 A 三组对四条音频进行了测量对比，其中由于该系统采用的非平行的语料，因此目标音高与音调为数据集中全部音频的平均值，测试所得的结果如表 5-1。

表 5-1 转换后与目标的音高音调对比分析表

组别	序号	目标音高	转换后音高	目标音调	转换后音调	是否符合标准
男 A→男 B	1	266	269	758	742	是
	2		283		740	是
	3		277		764	是
	4		270		832	否
	5		258		755	是
男 C→男 D	1	232	240	587	558	是
	2		228		610	是

	3		236		603	是
	4		245		574	否
	5		237		569	是
男 A→女 A	1	557	539	630	653	是
	2		565		595	否
	3		546		660	是
	4		596		642	否
	5		561		635	是

测试结果中可以看出，在同性之中对语音进行转换有较高的成功率，异性之间的转换更容易出现不符合标准的情况，但结合主观感受，转换后的音频与目标都有较高的相似度。

在完成语音转换的目标之后，本文在研究中发现，转换后输出的音频较为模糊，从体感上能感受到不是由一个“人”直接说出来或是录制后被播放出来的，更像是对讲机中发出的有杂音的音频。

由于转换后的音频表现出的质量不高的原因是参杂噪音，因此本文推测，可以用两种方式对音频进行处理，即使用滤波器或对音频降噪。在降噪后将计算降噪前后的 MSE 损失和差分，用于评价音频去噪的效果。其中 MSE 损失在 3.3 中循环一致性损失的介绍中有提到过，差分的计算方法是在降噪前后的两端音频的相同采样点进行做差求出差分值，并在最后累加求均值。MSE 损失和差分都可以在一定程度上反映音频在降噪前后的变化程度，因此 MSE 损失和差分值越大代表其降噪程度越大。因此此处的期望是在 MSE 损失和差分较大的情况下，音频能够进行有效的去噪，但不至于失真。

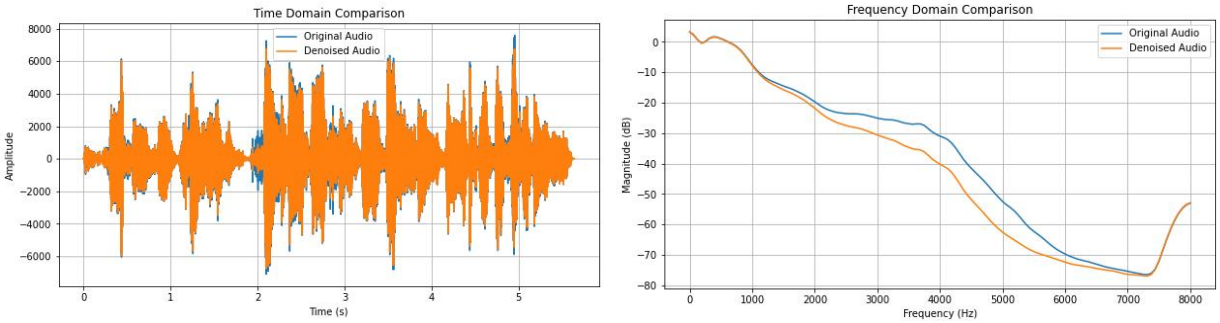
接下来将取男 A→男 B 转换的 5 对转换后的音频进行滤波和降噪处理并记录降噪前后音频对比得到的 MSE 损失和差分的平均。各方法的处理所得得 MSE 损失和差分值如表 5-1。

表 5-2 滤波和去噪方法的分析对比表

方法\数值	MSE 损失	差分
高斯滤波（sigma=2）	3.02e-5	2.97e-3
高斯滤波（sigma=1）	5.51e-6	1.17e-3
高斯滤波（sigma=0.5）	3.73e-7	2.96e-4
带通滤波（100，4000）	4.05e-7	2.30e-4
带通滤波（100，5000）	3.15e-9	2.85e-5
带通滤波（200，5000）	5.40e-7	4.92e-4
中值滤波（5）	1.25e-5	1.02e-3
中值滤波（9）	4.12e-5	2.37e-3
中值滤波（3）	1.74e-6	2.62e-4
小波去噪（db4,3,3）	7.68e-6	1.99e-3
小波去噪（db4,5,3）	1.14e-5	2.59e-3
小波去噪（db4,5,5）	2.42e-5	3.75e-3
非线性谱减法（0.5，1.5）	9.08e-4	1.99e-2
非线性谱减法（0.3，1.5）	9.08e-4	1.99e-2
非线性谱减法（0.3，1.1）	1.35e-4	7.90e-3

高斯滤波器：高斯滤波器常用于处理包含高频噪声的音频，此滤波器能够在处理音频噪声后输出比

较柔和平滑的音频信号，在需要保留原始音色的音频处理中常常能够起到作用，因此此处适用了高斯滤波器进行尝试。在使用高斯滤波器时需要设置 sigma 参数，该参数会影响到以下几个因素：输出音频的信号平滑程度，输出音频信号的模糊程度，去噪效果，计算复杂度。文中先将 sigma 值设置为 2 进行测试，测试结果音频中的人声部分音量被明显削弱，所以此处将 sigma 值降为 1 后进行尝试，尝试后发现人声部几乎未被削弱，且噪音问题有明显的改善。此处又将 sigma 设置为 0.5 进行尝试，MSE 损失和差分值相比 sigma 为 1 时更小，此处音频中的噪音依旧明显。因此高斯滤波组中，sigma 为 1 时能取得相对而言最好的效果。图 5-6 为 sigma 为 1 时去噪前后的时域和频域图：

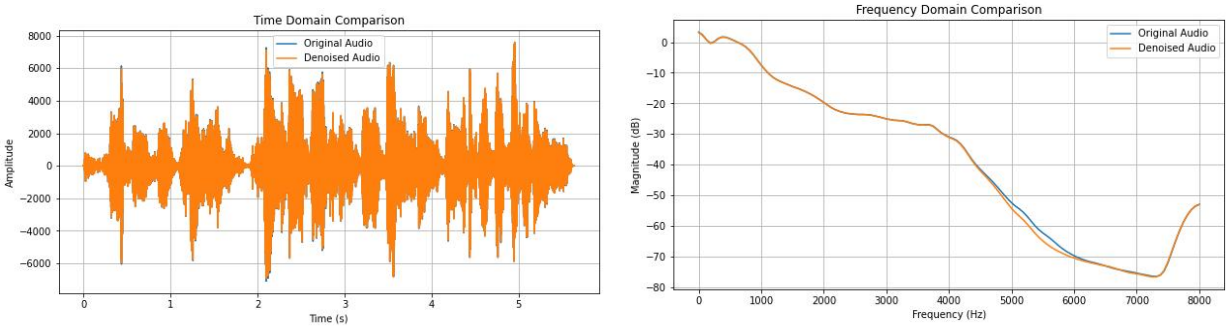


（a）高斯时域图　　（b）高斯频域图

图 5-6 高斯滤波器的时域及频域图

此处为了更加方便对比，绘制频域图时使用了折线图。

带通滤波器：带通滤波器可以去除部分频率的噪声，在某种程度上可以认为是高通滤波器和低通滤波器的结合使用，其参数影响的是能够通过滤波器的音频范围。由于该处是对人声的处理，所以先尝试（100，4000）为参数，尝试后发现人声部分受到了较大的削弱，考虑到人耳听到的音量中，很大部分是受到声音的谐波影响，所谓谐波是基频整数倍数的频率的声音波形，因此在后续的尝试中提高了高频的取值。进行参数为（100，5000）的尝试，尝试中发现人声部分的音量正常，但是音频中参杂着较多噪音，此噪音在参数（100，4000）时也有出现，不过因为音频声音受到了削弱所以显得并不明显，因此此处提高了低频部分的取值。尝试参数为（200，5000）的带通滤波效果，MSE 损失和差分值比取值为（100，5000）时明显要大，对噪音的处理相对其他参数更有效。总的来说，带通滤波器组在参数为（200，5000）时取得了最好的效果，但是与其他组相比，其效果并不明显。本文认为，这种现象出现的原因是在转换后的音频中，人耳听到的嘈杂声的频率分布很人声的频率分布有大部分的交叉，所以用带通滤波器很难取得有效的成果。图 5-7 为带通滤波器参数为（200，5000）时去噪前后的时域图与频域图：

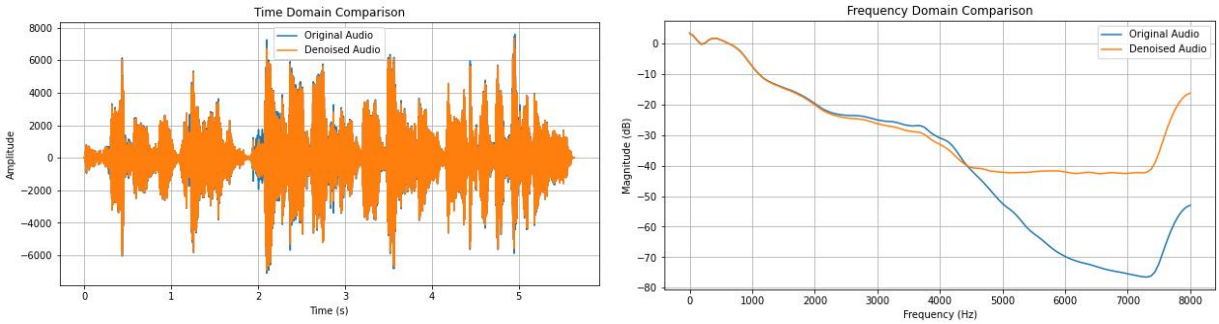


（a）带通时域图　　（b）带通频域图

图 5-7 带通滤波器的时域及频域图

中值滤波：中值滤波器在使用时需要对窗口大小进行设定，所谓窗口大小就是该滤波器在每次滤波

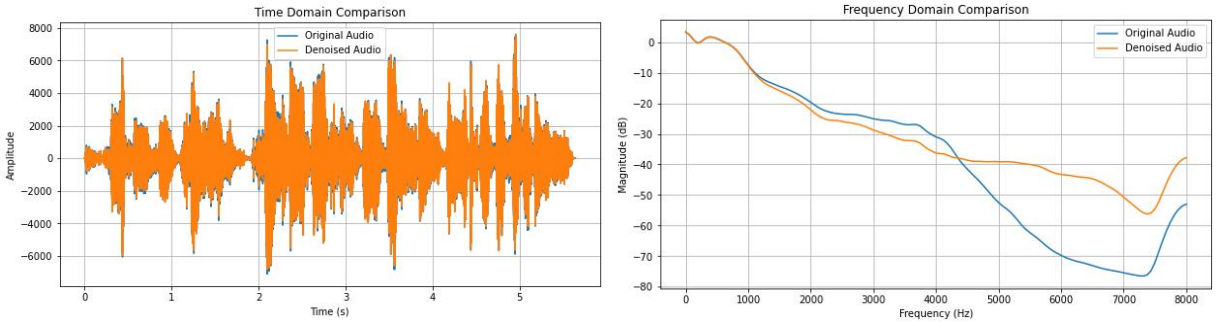
时会选择多少个采样点尺寸的邻域并计算中值。窗口的大小会影响去噪的效果和输出音频的平滑度，较大的窗口适合处理低频噪声较多的音频信号，较小的窗口适合处理高频噪声较多的信号，与此同时，过大的窗口可能会导致音频信号的畸变。窗口大小应被设定为奇数，因为该算法中，一个采样点的取值将由以该采样点为中心的数个值的中值决定，若窗口大小为偶数时，窗口的中心是不存在的。此处将窗口选择为 5 进行测试，测试输出的音频噪声依旧明显，继续加大窗口为 9 进行测试，测试结果相较与窗口为 5 时并无太大区别，此处进行推测，转换后输出的音频噪声以高频噪声为主，因此此处减小窗口为 3 进行测试，测试输出的音频相较窗口为 5 时，噪音的影响有所改善，但依旧影响观感。此处取窗口大小为 5 的去噪前后音频时域频域图作展示：



（a）中值时域图　　（b）中值频域图

图 5-8 中值滤波器的时域及频域图

小波变换：小波变换需要选择小波基，由于人声的频率范围主要集中在低频部分，因此本文打算使用具有良好低频性质的小波基 db4，这种小波基在语音去噪中有广泛的应用。小波变化中还有两个参数分别是小波变换的级数和阈值。级数越高捕获到的信号细节就越多，但是过高时可能会导致信号的失真，过低时则会导致噪音无法被有效去除，阈值会决定小波系数的保留或丢弃，过高的阈值可能会导致关键信息被丢弃而导致失真，过低会导致噪音无法去除。此处将参数设置为（db4，3，3）进行尝试，尝试发现噪音明显，因此此处先调高级数进行尝试，将参数设置为（db4，5，3），发现噪音依旧明显，与参数（db4，3，3）时并无明显区别，因此此处在将参数设置为（db4，5，5）发现效果与其他参数时相似，但其 SME 损失与差分都有变化，所以此处认为不是操作问题，不过在音频的表现上没有变化。该组的效果都并不好，取参数为（db4，3，3）的小波去噪前后音频的时域频域图作展示：

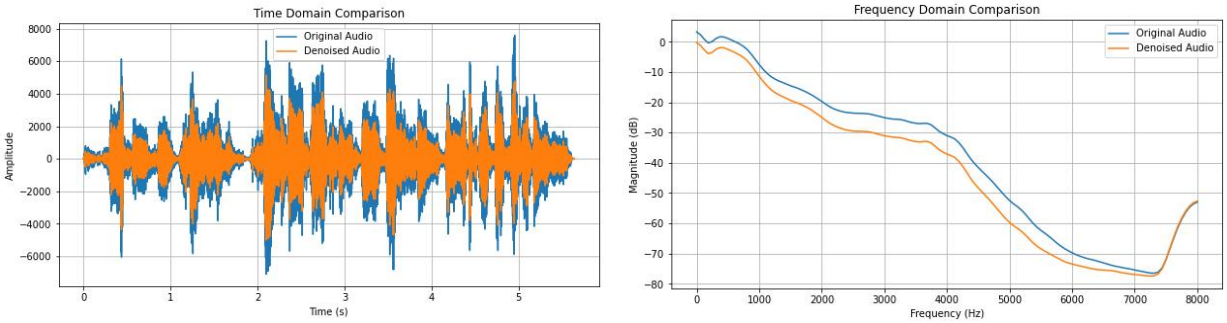


（a）小波时域图　　（b）小波频域图

图 5-9 小波去噪的时域及频域图

非线性谱减法：非线性谱减法在噪声环境简单的情况下一般有比较好的效果，其中有两个参数，分别是阈值和幂次方，阈值用于判断哪些信号部分是噪声，哪些是声音主体，较低的阈值能够保留较多语音信号，但也可能导致噪音去除不干净，幂次方则是用于增强语音信号，抑制噪音信号，但是过高的幂次方可能会引入新的噪声。此处从参数（0.5，1.5）开始进行尝试，发现人声部分受到了较大的削弱，

此处认为是幂次方不够高，导致语音信号未被增强，因此将参数调整至（0.5，3），发现音频中没有声音，因此未记录此项尝试的 MSE 损失和差分值，推断其原因是阈值为 0.5 时，人声部分被判断为是噪声，因此此处将参数调整为（0.3，1.5），发现人声依旧受到了明显的削弱，此处尝试降低幂次方，在参数为（0.3，1.2）时声音问题相对而言明显改善，且未感受到噪音的影响，再次尝试了参数（0.3，1.1）并与（0.3，1.2）进行对比，发现幂次方为 1.1 时声音更清晰，且没感受到噪音，因此此表中记录（0.3，1.1）的 MSE 损失与差分。图 5-9 为非线性谱减法去噪前后的时域图与频域图。



（a）谱减时域图　　（b）谱减频域图

图 5-10 非线性谱减法的时域及频域图

除去对男 A→男 B 转换后的 5 条音频进行测试之外，本文为了证明其结果并不是偶然性的得出，还对男 B→男 A，男 C→男 D，男 D→男 C，男 A→女 A，女 A→男 A5 组输出后的 5 条音频进行了上述 5 中方法中效果较好的参数组进行了测试并取 MSE 损失，其结果如下：

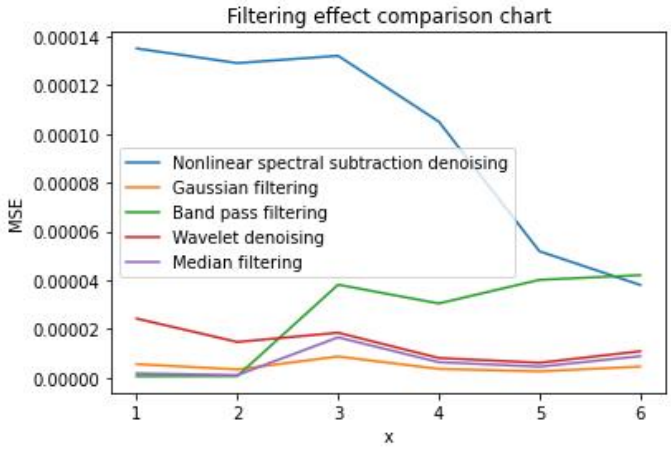


图 5-11 不同组别在去噪时的 MSE 损失

5.3 本章总结

本章对语音转换系统进行了界面设计，其中有四个界面，分别是登录界面、注册界面、主界面和模型训练界面。在这四个界面之中，考虑到该系统的使用可能需要受到管控，所以注册界面是独立的，不能通过其他界面进行跳转。另外三个界面中，在登录界面登录成功后就可以自动跳转到主界面。主界面可以使用现有模型对音频进行说话人 A 到 B 和 B 到 A 的双向转换，在转换后还可以选择两个音频绘制出频域图，进行对比，此外主界面中还有按钮可以进行页面跳转，模型训练按钮可以跳转至模型训练界面，退出按钮可以返回至登录界面。模型训练界面中可以新建模型并对其进行训练，点击返回按钮即可退回主界面。

在完成语音转换的功能后，本文发现在对音频进行转换之后，输出的音频中会参杂着些许噪音，本

文认为这会降低音频的自然度和使用者的舒适度，因此本文分析对比地先后使用了高斯滤波、带通滤波中值滤波、小波去噪以及非线性谱减法去噪对输出地音频进行处理，这其中使用的每种方法，本文都对其进行了分析，尽可能的找到了其效果最好的一项参数并选择了寻找和分析过程中的三项参数进行记录。最后本文发现在这四个方法当中，当高斯滤波 sigma 为 1 时，谱减法参数为阈值为 0.3，幂函数为 1.1 时对转换后音频的处理都能有比较好的效果，其中，高斯滤波器输出后人声相对谱减法更大，但是谱减法输出后，噪声部分对音频造成的影响更小，因此此处认为，适用非线性谱减法调整阈值为 0.3，幂函数为 1.1 能够取得最好的效果。

第六章　总结与展望

6.1 全文总结

本文分析了语音转换技术的发展对其相关领域技术的意义及其本身的社会意义，研读相关文献后确定了用 CycleGAN 算法结合 WORLD 声码器来实现语音转换的功能。在通过对语音转换技术的深度了解之后，发现了语音转换的本质就是提取出音频中能够体现出说话人声音特征的基频、频谱包络，并对其进行分析和转换最终达到实现语音转换的功能。

了解并学习了深度学习中 GAN 的基础知识，理解并分析了 GAN 网络中的损失函数，在此基础上学习并理解了 CycleGAN 网络的构建过程以及其 3 个用于训练的损失函数的意义。

在得到了转换音频的输出结果后，发现了音频质量不高的问题，分析并提出了四种可能能够解决该问题的方式，分别是高斯滤波、带通滤波、小波去噪和非线性谱减法去噪，并通过测试获取了其 MSE 损失和差分，对比分析客观数据并结合了对音频的主观感受得出结论。

本文的工作内容如下：

1.概述了语音转换的选题背景及其意义，指出了语音转换技术的发展能对语音合成、语音增强、语音识别等多个技术产生推力，并分析预测了语音转换技术在医学和心理学领域的作用和必要性。分析查阅资料，确定了完成语音转换目的的方法。

2.介绍了语音音频的几个重要组成部分及其能够体现表达出的意义，并对 WORLD 声码器进行了初步的介绍。

3.介绍了几种深度学习方法及其适用范围，了解和分析了 GAN 及其分支 CycleGAN 的训练流程和损失函数。

4.采集声音并进行预处理形成训练集，构建网络实现了语音转换的功能。

5.通过计算转换后音频的音高与音调确认了语音转换有较高的成功率，确认了功能的实现。再次之后发现音频质量不高的问题，提出了通过滤波和降噪的方式来改善问题，测试了高斯滤波、带通滤波、中值滤波、小波去噪和非线性谱减法，通过计算对比 MSE 损失及差分为客观评价方式，以体感听觉作为主观评价方式，此处认为适用非线性谱减法并将阈值设置为 0.3，幂次方设置为 1.1 时能够在保留输出音频原有音量的前提下最大程度的进行去噪。

6.2 下一步研究工作

目前对该系统的开发只停留在了语音转换的实现，绘制音频文件的频谱图，以及登录和注册的功能。但是由于时间和资源的限制，本设计只停留于实验阶段，并没有在实际场景中投入运用过，还存在很多问题与不足。若是需要实际应用，还需要进一步地修改和深化。

今后可以从以下方面进行研究完善：

1.系统中并没有添加录制音频，去噪的功能，这会导致用户体验不佳，因此完善系统的基本功能是下一步研究与开发的重点。

2.由于时间的限制，对系统界面的开发还停留在初期阶段，人机交互不够流畅，系统界面不够精美，以及未投入过实际应用，这些缺陷是一个系统所要尽力避免的，因此这也会是下一步工作的重点。

致谢

时光飞逝，转眼已经临近了四年大学生活的终点。

从开始选题，到完成论文的写作，这将近 6 个月的时间里，诚挚地感谢王泽峰导师、胡连信老师以及林贤伟师兄为我提供地帮助和指导。

首先要感谢我的导师，王泽峰教授，王老师博学多识、敬业严谨，总是能够在我迷茫的时候为我提供设计系统的方向和思路，在我停滞不前时教导我，让我明白了何为科研，如何做才是面对科研的正确态度。让我懂得了在面对问题时如何解决、如何分析。在论文的编写过程中，对论文的质量精益求精，端正了我的学习态度。至此向王泽峰教授致以诚挚的敬意与感谢。

同时感谢胡连信老师，在论文中为我提供了许多细致的建议，使我能够在论文的编写过程中少走弯路，让我了解到了如何掌握论文的主要思想。至此感谢胡连信老师辛勤的付出。

其次还要感谢林贤伟师兄，小至音频采集的注意事项，大至代码的编写与理解，师兄为我解决了许多技术性问题，再次真诚的感谢师兄。

于此同时还要感谢我的家人，在我的求学生涯中向我提供了经济与精神的支持，在我完成学业的过程中作我坚实的后盾。

最后感谢论文评阅组和答辩组的老师们，感谢百忙之中阅读我的论文，感谢各位老师的指导。

参考文献

[1]张小峰,谢钧,罗健欣等.深度学习语音合成技术综述[J].计算机工程与应用,2021,57(09):50-59.

[2]凌震华,伍宏传.基于WaveNet的语音合成声码器研究[J].人工智能,2018,No.2(01):83-91.DOI:10.16453/j.cnki.issn2096-5036.2018.01.008.

[3]MEHRI S,KUMAR K,GULRAJANI I,et al.Samplernn:an unconditional end-to-end neural audio generation model[J].arXiv:1612.07837,2016.

[4]NAL K,ERICH E,KAREN S,et al.Efficient neural audio synthesis[C]//Proceedings of the 35th International Conference on Machine Learning,Stockholm,Sweden,2018.

[5]HAO Y,DONG L,WEI F,et al.Visualizing and understanding the effectiveness of BERT[J].arXiv:1908.05620,2019.

[6]安鑫,代子彪,李阳,孙晓,任福继.基于BERT的端到端语音合成方法[J].计算机科学,2022,49(04):221-226.

[7]王研,吴怡之.基于变分自编码的语气语音合成模型[J].计算机科学与应用,2020,10(12):2159-2167

[8]刘畅,魏为民,孟繁星,才智.语音风格迁移研究进展[J].计算机科学,2022,49(S1):301-308+362.

[9]唐浩彬,张旭龙,王健宗,程宁,肖京.表现性语音合成综述[J/OL].大数据:1-23[2023-01-05].http://kns.cnki.net/kcms/detail/10.1321.G2.20221108.1439.006.html

[10]张冠萍.基于语音合成的英语机器翻译机器人设计[J].自动化与仪器仪表,2023,No.280(02):247-252.DOI:10.14016/j.cnki.1001-9227.2023.02.247.

[11]李乃寒.面向语音合成的深度学习算法研究与应用[D].电子科技大学,2021.DOI:10.27005/d.cnki.gdzku.2021.000031.

[12]王智,刘银华.基于深度学习的中文情感语音合成方法[J].自动化与仪器仪表,2022,No.275(09):10-15.DOI:10.14016/j.cnki.1001-9227.2022.09.010.

[13]帕丽旦·木合塔尔,吾守尔·斯拉木,买买提阿依甫.HMM与神经网络相融合的低资源语音合成方法[J].计算机仿真,2021,38(12):203-211.

[14]王瑞.基于自动语种识别的汉藏双语跨语言语音转换研究[D].西北师范大学,2022.DOI:10.27410/d.cnki.gxbfu.2022.000692.

[15]谭智元.基于自编码器的零样本语音转换系统研究[D].天津大学,2020.DOI:10.27356/d.cnki.gtjdu.2020.004546.

[16]康筱.基于矢量量化-变分自编码器的语音转换系统研究[D].新疆大学,2021.DOI:10.27429/d.cnki.gxjdu.2021.000428.

[17]李涛. 基于CycleGAN网络实现非平行语料库条件下的语音转换[D].大连理工大学,2018.

[18]朱雅楠.基于表示分离的语音转换方法[D].杭州电子科技大学,2022.DOI:10.27075/d.cnki.ghzdc.2022.000843.

[19]于杰.基于基频差异补偿的StyleGAN情感语音转换研究[D].南京邮电大学,2022.DOI:10.27251/d.cnki.gnjdc.2022.000669.

[20]邱祥天.融合DSNet与ESR网络的StyleGAN语音转换研究[D].南京邮电大学,2022.DOI:10.27251/d.cnki.gnjdc.2022.001351.

[21]戴少梁.基于激活指导和内卷积的跨语种语音转换研究[D].南京邮电大学,2022.DOI:10.27251/d.cnki.gnjdc.2022.001228.

[22]徐玲俐.基于双编码器的快速one-shot跨语种语音转换方法[D].南京邮电大学,2021.DOI:10.27251/d.cnki.gnjdc.2021.001015.

[23]胡雨婷. 噪声环境下基于深度卷积神经网络的多模态语音转换研究[D]. 安徽大学,2020.DOI:10.26917/d.cnki.ganhu.2020.001009.