DOI: 10.16300/j.cnki.1000-3630.23031601 CSTR: 32055.14.sxjs.1000-3630.23031601
引用格式：马皓天, 洪峰, 毛海全, 等 . 用于提升聋哑人语音表现力的语音合成技术[J]. 声学技术, 2024, **43**(6): 843-853. [MA Haotian, HONG Feng, MAO Haiquan, et al. Study on text to speech improving the voice expression of deaf people[J]. Technical Acoustics, 2024, **43**(6): 843-853.]

用于提升聋哑人语音表现力的语音合成技术

马皓天[1,2]，洪 峰[1]，毛海全[1,2]，郑立通[1,2]，牟宏宇[1]，许伟杰[1]
(1. 中国科学院声学研究所东海研究站, 上海 201815；2. 中国科学院大学, 北京 100190)

摘要：目前，聋哑人主要通过手语的方式与健听人进行沟通，但这对未接受专业手语学习的健听人来说是一种挑战。因此，将手语转换为文本，再将文本转换成带有聋哑人音色的、健听人能理解的语音非常具有研究意义。为研究聋哑人语音合成的可行性，文章首先分析了聋哑人的语音特征，并根据分析的结论，提出了能合成高自然度、高清晰度且带有聋哑人自身声音特色的模型算法以及相应的评估体系。文章根据不同残疾程度的聋哑人语音特征，提出了面向轻度残疾聋哑人的语音转换和合成方法以及面向重度残疾聋哑人的语音克隆方法。根据分析结果，轻度残疾聋哑人语音与健听人语音具有一定的共性，因此使用 AdaIN-VC 语音转换模型转换出带有聋哑人音色、高可懂度的语音，并将转换好的语音结合 Tacotron2 语音合成模型进行文本到语音的映射。考虑到重度残疾聋哑人语音的不稳定性，文章基于 Zero-shot 的 SV2TTS 语音克隆框架，使用了 ECAPA-TDNN 作为重度残疾聋哑人音色表征的说话人编码器，以获取准确的聋哑人表征。此外，文章还引入基于基频情感分类的风格迁移模块，对合成语音进行风格上的迁移。实验结果表明，在保证一定相似度的情况下，实验中两位轻残聋哑人的自然度主观意见评分别从原来的 2.53 和 3.06 提高至 2.88 和 3.21，并且语音识别的错词率从 100% 分别降低至 80.77% 和 76.91%。同样，文中提出的主观错词率也有明显的下降。而在语音克隆的实验中，模型合成的重残聋哑人语音与其自身音色的相似度主观相似度意见评分达到 3，且聋哑人语音的自然度主观意见评分和情感表达能力均得到了提高。

关键词：语音合成；语音转换；语音克隆；风格迁移
中图分类号：H107 **文献标志码**：A **文章编号**：1000-3630(2024)-06-0843-11

Study on text to speech improving the voice expression of deaf people

MA Haotian[1,2], HONG Feng[1], MAO Haiquan[1,2], ZHENG Litong[1,2], MOU Hongyu[1], XU Weijie[1]
(1. *Shanghai Acoustics Laboratory*, *Chinese Academy of Sciences*, *Shanghai* 201815, *China*;
2. *University of Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: Currently, deaf people mainly use sign language to communicate with healthy people, however, most healthy people are untrained in sign language training. Therefore, it is of great importance to translate the sign language into spoken language using deaf accents that can be comprehended by the healthy people. To investigate the feasibility of text to speech (TTS) for the deaf people, the speech characteristics are analyzed firstly in this paper, and then, the TTS algorithms, which are capable of generating high naturalness and clarity speeches with deaf people's own voice characteristics, and the evaluation methods for these algorithms are developed. In this paper, a voice conversion and TTS method for mildly disabled deaf people and a voice cloning method for sever deaf people based on the characteristics of their speech are proposed. According to the analysis results, the voice of the mildly disabled deaf person has some similarities with the healthy voice, so the AdaIN-VC speech conversion model is used to convert the voice with the timbre and high understanding of the deaf person, and the converted voice is combined with the Tacotron2 speech synthesis model to map the text to the speech. Considering the instability of severely disabled deaf speech, the ECAPA-TDNN is used as the speaker coder for the tone representation of severely disabled deaf people to obtain accurate deaf representations. In addition, the style migration module based on the base frequency emotion classification is introduced to transfer the style of the synthetic speech. The experimental results show that under the

收稿日期：2023-03-16; **修回日期**：2023-04-21
基金项目：中国科学院声学研究所自主部署"前沿探索"项目 (QYTS202114)、中国科学院青年创新促进会 (2021022) 项目、上海市自然科学基金项目 (22ZR1475700)
作者简介：马皓天 (1996—)，男，广东广州人，硕士研究生，研究方向为声纹识别及语音合成方向。
通信作者：洪峰，E-mail: hongfeng@mail.ioa.ac.cn

condition of ensuring certain similarity, the subjective opinion scores of the two mild deaf people in the experiment increased from 2.53 and 3.06 to 2.88 and 3.21, respectively, and the misword rate of speech recognition is reduced from 100% to 80.77% and 76.91%, respectively. Similarly, the rate of subjective miswords proposed in the paper has also decreased significantly. However, in the experiment of speech cloning, the subjective similarity opinion score for the similarity of the severely disabled deaf speech and its own timbre reached 3, and the natural subjective opinion score and emotional expression ability of the deaf speech are improved.

Key words: text to speech (TTS); voice conversion; voice cloning; style transfer

0 引 言

声音是人类最重要的交流工具之一，也是人类性格中不可分割的一部分，人类的声音被 描述为“人类生命的本质之一”[1]。然而，聋哑人失去了倾听和说话的能力。在我国，聋哑人群基数庞大，据 2022 年 9 月 25 日“国际聋哑人节”的统计，我国聋哑人约有 2 780 万，仅次于视力残疾，占中国人口总数的 1.67%。目前，由于健听人一般缺乏系统性的手语学习和练习，因此难以准确地理解手语的意思，导致聋哑人在公共场合中常常面临沟通障碍的窘境。语音合成 (text to speech, TTS) 能将任意文字转化为流畅的语音输出。近年来，随着机器学习、深度神经网络等技术的高速发展，语音合成得到了深入的研究和广泛的应用。因此，合成具有聋哑人自身声音特色的高自然度、高可懂度的语音技术成为了可能。2016 年，Wang 等[2]提出 Tacotron 模型，即从音素序列中直接生成语音特征而不是语言特征，这可以看作是端到端语音合成的首次探索。Tacotron2 模型[3]在 2017 提出，对 Tacotron 模型做出了改进，并以 WaveNet[4]作为神经声码器，该方法能合成出较为自然的语音。Fastspeech2 模型能在模型推理时控制说话人的语速、音调等语音特征[5]。Neekhara 等[6]将语音合成模型与说话人识别技术相结合，提出语音克隆模型，该模型能合成训练集中不存在的说话人音色。

目前对聋哑人语音合成的研究还较少。谷歌团队提出了 Parrotron 模型，该模型能将聋哑人语音映射成特定健听说话人的语音[7]。本文直接对聋哑人本身的声音进行分析和研究，通过使用语音转换、语音合成以及语音克隆技术，输出与聋哑人音色相似度较高的、自然度和可懂度具有一定提升的语音。

合成富有情感的、自然的语音一直是语音合成领域最具有挑战性的问题。Mellotron 模型将基频作为刻画声调的特征，以此来决定多说话人语音合成的韵律[8]。Skerry-Ryan 等[9]用若干个初始化后的嵌入码，通过线性加权的方式学习语音的情感表达。无论是轻度残疾聋哑人 (简称为“轻残聋哑人”) 还是重度残疾聋哑人 (简称为“重残聋哑人”)，其语音均无法表达出情感。

为解决上述问题，本文提出可分别应用于轻度残疾聋哑人和重度残疾聋哑人的语音合成技术。针对轻残聋哑人的语音特性，本文使用 AdaIN-VC[10] 语音转换解耦聋哑人音色与语义特征，并与解耦的健听人语义特征进行重组，达到提高聋哑人语音可懂度和自然度的效果。而面向重残聋哑人的语音克隆技术则是在 SV2TTS[11]语音克隆框架的基础上，使用说话人识别更准确的 ECAPA-TDNN[12]提高克隆相似度，并引入基于基频的风格迁移模块提升聋哑人语音的情感表现力。除此之外，本文还提出主观错词率 (subjective world error rate, SWER) 用于对聋哑人语音主观可懂度评价。最后本文使用主观意见评分 (mean opinion score, MOS)、相似度主观意见评分 (similarity mean opinion score, SMOS)、错词率 (world, error, rate, WER) 和 SWER 作为合成结果的评价方法，以验证本文所提方法的有效性。

1 健听人与聋哑人语音特征分析

1.1 时域对比分析

对录制的聋哑人语音进行时域的对比分析，其中部分的轻残聋哑人能像健听人一样发出连续的语音，但无法控制清晰度和语调。部分重残聋哑人的语音具有间断性、声音幅度变化过大等特征。聋哑人和健听人语音的时域波形如图 1 所示。图 1(a)、1(b)、1(c) 分别为健听人、轻残聋哑人和重残聋哑人的语音波形图，其语料内容均为中文常用句“我觉得自己又胖了”。由图 1 可以得出健听人和聋哑人语音的差异：(1) 健听人和轻残聋哑人的语音较为连续，而重残聋哑人的语音连贯性不高；(2) 健听人的语音幅度有轻重之分，能控制语音表达的抑扬顿挫。而聋哑人语音的幅度变化较大，特别是重残聋哑人，发音时经常会出现破音的现象；(3) 相较于健听人而言，因为聋哑人发音困难，所以即使是相同的文本内容，他们需要更长的时间来发音。

1.2 健听人与聋哑人 log-Mel 分析

健听人与聋哑人的对数梅尔谱 (log Mel-spectogram, log-Mel) 如图 2 所示。图 2(a)、2(b)、2(c)

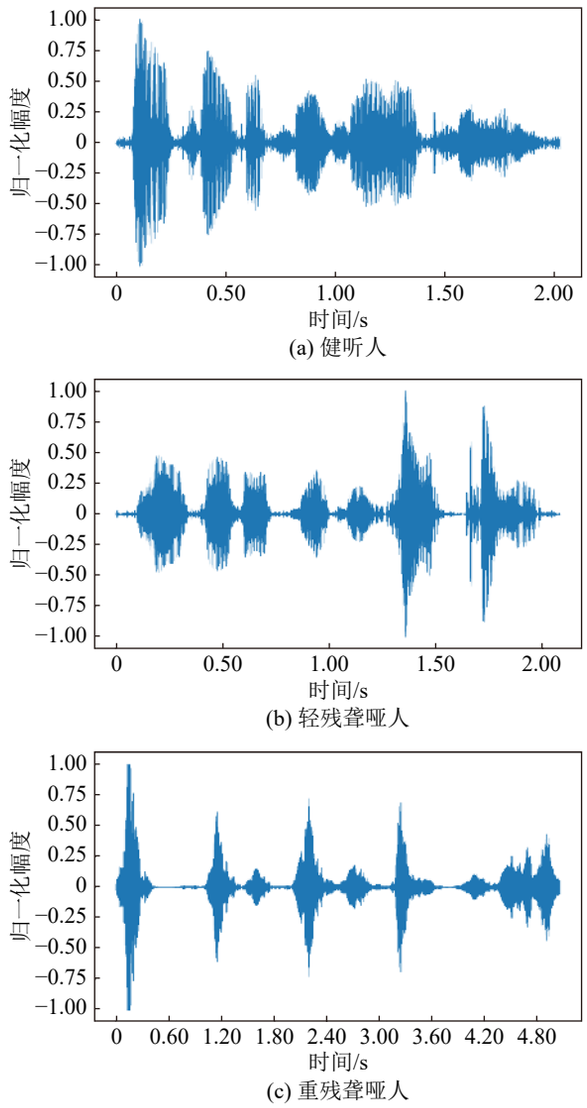

图 1　健听人与聋哑人语音的时域波形

Fig.1　Time domain waveforms of speech for healthy and deaf people

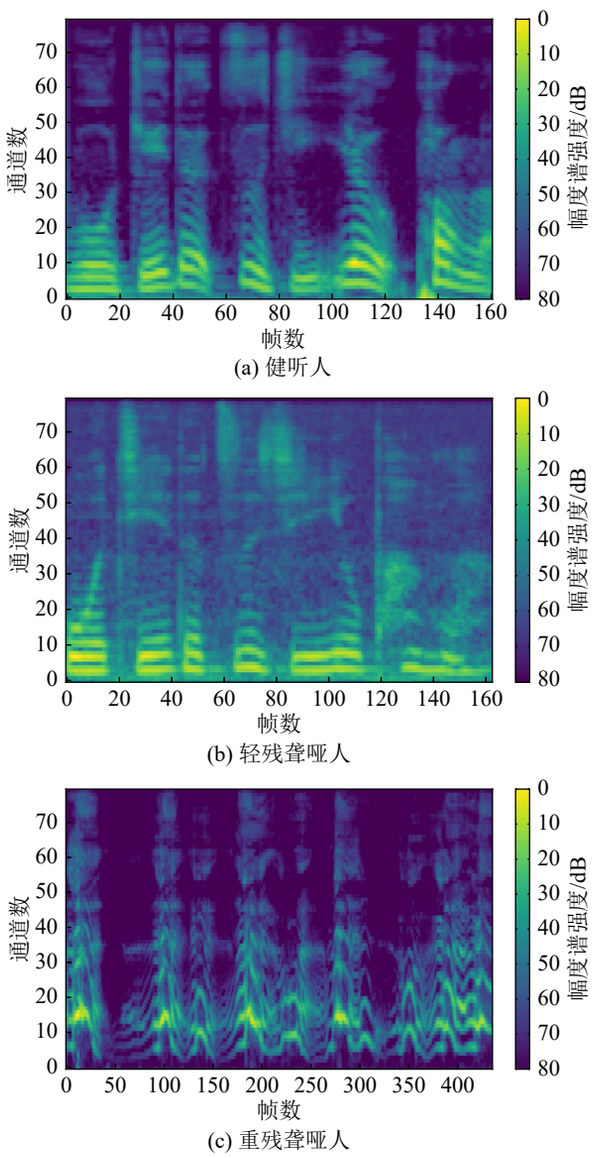

图 2　健听人与聋哑人 log-Mel 谱

Fig.2　The log-Mel spectra of healthy and deaf people

分别为健听人、轻残聋哑人和重残聋哑人的对数梅尔谱，语料内容均是中文常用句“我觉得自己又胖了”。由图 2 可知：(1) 健听人频谱图谐波分布较为紧密，能量主要集中在低频部分。聋哑人的各个谐波能量分布较为分散，其中重残聋哑人的 log-Mel 谱呈现不规则的形状；(2) 健听人与轻残聋哑人的谐波变化趋势较为平坦且连续性较高，而重残聋哑人的频谱谐波趋势变化较大，呈现剧增或剧减的趋势，并且具有间断性的特征；(3) 健听人频谱的高频信息和低频信息均比较完整，而重残聋哑人的高频与低频信息都较少。

由图 1 和图 2 可以得到以下结论：(1) 聋哑人无法像健听人一样发出连贯的声音；(2) 聋哑人语音的幅度起伏变化较大；(3) 聋哑人频谱图的低频能量分布较为分散且各谐波的变化趋势大。综上所述，聋哑人不仅难以发出可懂度和自然度高的语音，而且难以控制说话时的轻重音。健听人可以控制语音的音调、轻重音 (例如生气时声音较大且音调较高、沮丧时声音较小且音调较为低沉) 来表达情感，但聋哑人难以通过语音表达内容以及情感。

虽然轻残聋哑人的语音可懂度不高，但其连贯性、稳定性和自然度与健听人语音具有一定的相似之处，即轻残聋哑人在短时间内能稳定地发音。如图 2(a) 和 2(b) 可知，轻残聋哑人与健听人在低频处具有非常明显的声纹特性。因此可将聋哑人语音与健听人语音共同训练语音转换模型，从而提高聋哑人的可懂度。与轻残聋哑人与健听人的语音特征相比，重残聋哑人短时间内语音的音量、音高等特征变化过大，而且其声纹畸变严重。因此无法使用常规的语音转换方法进行音色提取。本文利用语音克隆的方法，通过说话人识别模型克隆与其音色相

似的语音。

2 面向轻残聋哑人的语音转换和合成方法

语音转换使用的模型是 AdaIN-VC，该模型将语音中的音色视为全局变量，而轻残聋哑人的语音具有一定的稳定性和连贯性，可以通过实例归一化 (instance normalization，IN)[13]从语音中提取全局信息作为音色，并解耦出语义信息。将解耦的聋哑人音色与健听人解耦的语义进行结合，聋哑人语音的可懂度和悦耳度均有明显地提高。但由于语音转换模型无法直接通过文本控制语音内容的输出，后续仍需要通过语音合成实现文本到转换后语音的映射。轻残聋哑人的语音转换和合成框架如图 3 所示。图 3 中左侧部分为语音转换的模型，其中说话人编码器用于提取输入语料的音色，用Z_s表示音色特征。内容编码器用于从语音中内容的分离，用Z_c表示内容特征；右侧部分是基本的 Tacotron2 语音合成模型。

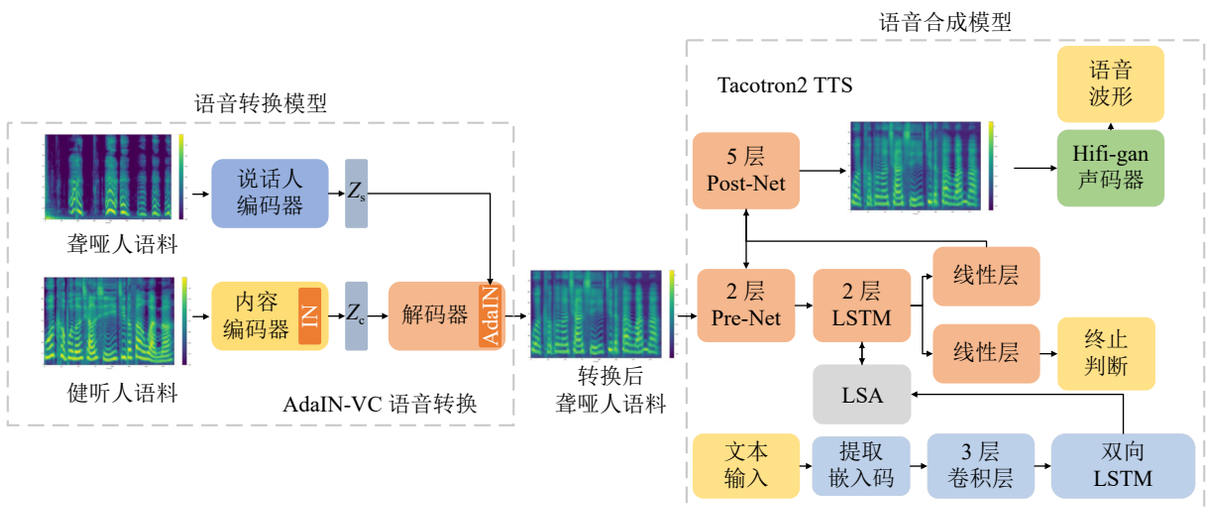

图 3 轻残聋哑人的语音转换和合成框架

Fig.3 A framework for voice conversion and synthesis of mild deaf people

2.1 AdaIN-VC 语音转换模型

AdaIN-VC 模型是非平行语料的转换模型 (即转换语料内容互不相同)，该模型通过变分自编码器[14]和 AdaIN[15]两种算法共同实现。AdaIN-VC 模型通过变分自编码器的损失函数学习语音中的特征分布，AdaIN 算法则用于说话人音色和语义信息的重构。由于健听人和轻残聋哑人的发音较为稳定，不会出现音色改变的情况，因此语音中的全局特征能代表音色。语音中的内容是不断变化的且与音色相互独立，因此可以通过实例归一化去除全局特征以达到解耦音色与语义内容的效果。实例归一化 (IN) 的原理如式 (1)~(3) 所示：

$$M_c'(w)=\frac{\boldsymbol{M}_c(w)-\mu_c}{\sigma_c} \tag{1}$$

$$\mu_c=\frac{1}{W}\sum_{w=1}^{W}\boldsymbol{M}_c(w) \tag{2}$$

$$\sigma_c=\sqrt{\frac{1}{W}\sum\nolimits_{w=1}^{W}\left[M_c(w)-\mu_c\right]^2} \tag{3}$$

其中：$\boldsymbol{M}_c(w)$表示输入的特征张量，$\boldsymbol{M}_c'(w)$表示归一化之后输出的特征张量。μ_c表示全局特征，通过计算每个张量通道特征维度的均值而得到。w表示第w层的特征，W表示总的通道数。σ_c表示每个通道的特征的标准差。一般来说，说话人的音色与内容是相对独立的，因而通过式 (1) 能去除输入声学特征 (80 维度的 log-Mel) 通道上的全局信息，从而解耦音色与语义信息。而用于语音重构的 AdaIN 算法的计算公式为

$$\boldsymbol{M}_c'(w)=\gamma_c\frac{\boldsymbol{M}_c(w)-\mu_c}{\sigma_c}+\beta_c \tag{4}$$

式中：γ_c和β_c是说话人编码器学习的说话人全局信息μ_c和方差信息σ_c。式 (4) 与式 (1) 是相反的过程。式 (1) 是对特征进行归一化，去除全局信息，式 (4) 则是对归一化之后的特征重新赋予新的全局信息，即对去除了说话人全局信息的特征，重新添加音色信息。

文中说话人编码器、内容编码器和解码器的结构与文献[3]中的结构一致，均是由若干个卷积模块和线性层模块之间的残差连接组成。每个卷积模块中包含两个一维卷积层以及两个 ReLU 激活函数。说话人编码器中的平均池化用于提取语音中的全局信息，即音色信息。内容编码器中则包含实例

归一化层，用于去除全局信息。解码器中的 AdaIN 层用于重组语义和音色，而像素重组 (pixel shuffle) 的作用则是类似于上采样的作用。

2.2　Tacotron2 语音合成模型

语音合成声学模型选用 Tacotron2 模型[3]，其网络结构图如图 4 所示。Tacotron2 是一个自回归的语音合成模型，能根据上一帧解码的结果生成下一帧的特征。图 4 中，蓝色部分为文本编码器，由 3 层卷积层和双向长短期记忆网络 (long short-term memory , LSTM)[16]组成。灰色部分为位置敏感注意力 (location sensitive attention, LSA)，其通过累加历史的注意力权重信息使模型在解码的过程中保持对齐的稳定性，以减少自回归模型推理时出现的漏字、多字等问题；橙色部分是解码器，声学特征先通过 2 层的 Pre-Net，随后送入 2 层 LSTM，与 LSA 输出的上下文特征拼接后共同训练，用于学习文本与音频特征的映射关系。最后将特征输入至 Post-Net 以微调 log-Mel 特征，合成出更逼真的 log-Mel 特征。终止判断用于生成结束符。

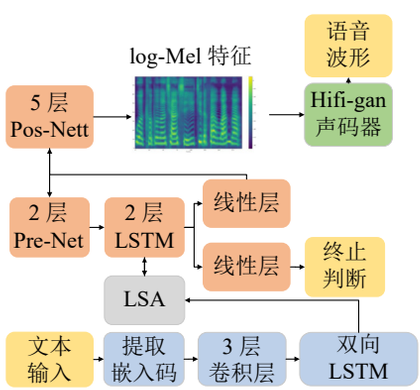

图 4　Tacotron2 的基本网络结构

Fig.4　The basic network structure of Tacotron2

2.3　HiFi-GAN 声码器

虽然 WaveNet 具有较好的合成效果，但是其合成速度较慢，难以满足实时性的要求[4]。本文使用 HiFi-GAN[17]声码器将输出的 log-Mel 谱转换成语音音频，该声码器不仅能合成自然度高的语音，还能提高合成语音的速度。该声码器基于生成对抗网络 (generative adversarial network, GAN)，由一个生成器和两个判别器组成。其中多感受野融合生成器由若干个残差链接模块构成，其主要作用是对输入的 log-Mel 特征进行上采样，输出对应的语音信号。两个判别器则是由多周期判别器和多尺度判别器组成。然后通过判别器与生成器对生成的语音进行对抗训练。本文 HiFi-GAN 声码器模型使用开源的预训练模型，来源于 GitHub 网站[18]。

3　面向重残聋哑人的语音克隆方法

为提高聋哑人语音的可懂度，本文训练时直接对健听人语音进行训练。说话人识别模块能提取待克隆说话人的声纹表征。风格迁移模块能提高聋哑人情感表达力。因此，模型不仅能克隆出与重残聋哑人相仿的音色，还能合成出具有不同风格的语音。

3.1　基于 ECAPA-TDNN 和 Tacotron2 的语音克隆模型

基于标签的风格迁移语音克隆训练框架如图 5 所示，模型的主体结构由说话人编码器、合成器和声码器共同组成。与文献[11]相比，本文使用识别性能更好的 ECAPA-TDNN 替代 LSTM 作为说话人编码器。ECAPA-TDNN 模型由 SE-Res2Net[19-20]、多尺度特征融合 [21]和注意力统计池化 [22]组成，SE-Res2Net 不仅能学习全局信道中重要的通道特征，并且通过对 Res2Net 引入残差连接的层次化结构提高性能。由于时延神经网络的结构具有层次性，多尺度特征融合机制将多个 Res2Blocks 的输出特征在特征维度上聚合，以丰富说话人信息。注意力统计池化能有效地表示说话人高阶统计量，学习帧级特征中重要的时间的信息。为了克隆训练集中未出现的说话人音色，需将 Tacotron2 模型生成的本文特征与说话人识别模块的说话人嵌入码进行拼接。此说话人识别的模型需要在大型的数据集中训练，因此模型能在充足的样本中学习绝大多数说话人的特

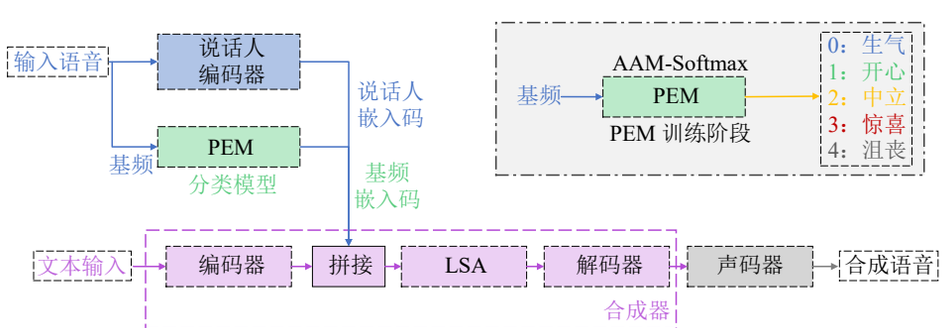

图 5　基于标签的风格迁移语音克隆训练框架

Fig.5　Training framework for pitch-based style transfer voice cloning

征。合成器模块和声码器模块分别使用 Tacotron2 模型和 HiFi-GAN 模型，网络结构与 2.2 节和 2.3 节所描述的一致。

3.2 引入基频的风格迁移语音克隆

基频特征能反映说话人的情感特征，文献[8]和文献[9]均说明基频能作为情感的重要特征。由于聋哑人的音高不能准确表达其情感，本文通过输入参考基频，对合成的语音进行情感风格上的迁移。并引入基频嵌入向量提取器模型 (pitch embedding model, PEM) 用于对输入的语音基频特征进行情感风格分类的训练。从训练完成的 PEM 中提取固定长度的风格嵌入码作为风格迁移的特征。参照多说话人语音合成的原理，将风格嵌入码、合成器输出的文本特征和说话人编码器输出的说话人嵌入码拼接后共同训练。模型在学习说话人特征的同时学习对应的风格特征，从而实现风格迁移的语音克隆具体的方法如下：

(1) 基频特征提取。在提取基频特征之前，对语音进行降采样、语音活动性检测 (voice activity detection, VAD) 和预加重处理。基频特征使用 WORLD 声码器[23]中的 DIO 算法进行提取。该算法原理是通过不同截止滤波器提取出置信度最高的正弦波作为基频。为保证提取的基频帧数与 log-Mel 谱的帧数保持一致，提取时设置与提取 log-Mel 谱相同的帧长和帧移。

(2) 基频特征情感分类模型。PEM 网络结构如图 6 所示，每一个 TDNN 模块中包含一个一维卷积层 (Conv-1d)、批归一化 (batch normalization, BN)[24]和激活函数 Leaky ReLU。图 6 中 X 代表输入的基频特征，B 表示批大小 (batch size)，T 表示帧数，⊕表示在特征维度上串联，Embedding 代表输出的基频嵌入向量。每个卷积层的输入与输出在时间维度上保持一致，但通道数则成倍增加。批归一化能对每一个批次的数据进行归一化，使网络中每层输入数据的分布相对稳定，同时能缓解过拟合的现象。激活函数选用 Leaky ReLU，该函数允许输出负值的梯度更新。在输入至池化层之前，根据 TDNN 的层次性，在时间帧维度上串联各个 TDNN 模块的输出，得到 496 维的向量。这种串联的方式能增加基频信息相互之间的联系，丰富基频信息。池化方法使用 ASP，用 256 维的向量来表示基频嵌入向量，训练时的损失函数选择 AAM-Softmax[25]。该损失函数相比于 Softmax，能使区分说话人的边界最大化并加强类内紧密度和类间差异。最后将 256 维的嵌入码通过 5 维的线性层，用于 5 种风格的分类。

(3) 引入基频特征的风格嵌入语音克隆。将 PEM 模型的基频嵌入码作为情感特征。嵌入的方式与多说话人 TTS 的方法一致，在文本输出的特征处和说话人嵌入码进行拼接。

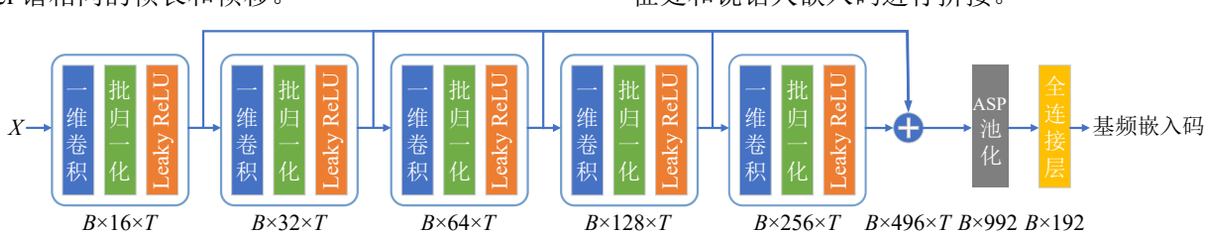

图 6 PEM 的网络结构

Fig.6 The network architecture of PEM

4 两种实现方法的有效性实验评估

本文的实验均在单个 RTX 3090 显卡上进行训练，使用 Python 3.8 进行程序代码编写，框架为 PyTorch。

4.1 面向轻残聋哑人的语音转换和合成方法

4.1.1 轻残聋哑人语音转换的数据集

本文中每一位轻残聋哑人需要训练专属的模型才能合成其可懂度高的语音。训练时使用了 8 位说话人的语料，包含 1 位轻残聋哑人和 7 位健听人。本文将详细分析编号 dm002(男性) 与编号 dm004 (女性) 的轻残聋哑人合成结果。训练 dm002 的转换模型时，聋哑人语料为 300 句，时长约为 1 h，语音采样率为 48 kHz。其余 7 位说话人均来自 AISHELL3[26]数据集中男性的健听人，每位健听人的数据约有 350 句短语料，时长约为 2.7 h，语音的采样率为 44.1 kHz；而在训练 dm004 时，聋哑人语料同样有 300 句，时长约为 1 h，语音采样率为 48 kHz。与 dm002 不同的是，7 位健听人中有 6 位健听人语料来自 AISHELL3 数据集中的女性说话人，每位说话人平均有 350 句。剩下 1 位健听人为标贝科技公开的标准女声数据集，数据集共有 10 000 条语料 (训练语音转换时仅使用 800 条语料)，时长为 12 h(作为转换训练的语料时长约为

2 h)，采样率为 48 kHz。训练好的模型对于闭集的轻残聋哑人语音有较好的转换质量，且转换后的语音质量能达到语音合成的质量要求。

由于转换后的聋哑人语料数量较少 (仅 400~800 句)，因此需要训练一个健听人的语音合成模型作为预训练的模型，并在此模型下进行转换后语音的训练，此过程被称作 Few-shot 说话人自适应，属于个性化 TTS 的方法之一。本文使用的预训练数据集为标贝科技公开标准女声数据集。选择该数据集的原因在于：(1) 该数据集的说话人具有较为标准的中文发音；(2) 该数据集没有杂音且信噪比高，训练时能排除噪声对合成语音的干扰，从而能合成出较高质量的语音。

4.1.2　模型配置

对输入语音进行 VAD、预加重等操作，并统一将采样率降至 16 kHz。随后提取帧长为 25 ms、帧移为 10 ms 的 80 维 log-Mel 作为语音转换和语音合成训练的声学特征。AdaIN-VC 模型与 Tacotron2 模型均使用 Adam 优化器，设置的初始学习率分别为 0.01 和 0.001，迁移学习时均为 0.0001)，设置优化器参数β_1和β_2均为 0.9 和 0.99。

在进行语音转换训练时，AdaIN-VC 模型的总体结构均是由卷积模块 (每个模块由 2 个卷积层和 2 个 ReLU 激活函数，部分模块会增加其他功能) 和线性层模块 (每个模块由 2 个线性层和 2 个 ReLU 激活函数) 组成。绝大部分卷积层的通道数为 128，卷积核大小为 3，步长为 1。其余的卷积层步长为 2。所有的线性层的神经元个数为 128。训练时批大小 (batch size) 为 64，共迭代 20 000 次，每迭代 2 000 次学习率衰减至原学习率的 90%。

在 Tacotron2 模型训练中，文本编码器由三层卷积层 (卷积核大小为 3) 和一个双向 LSTM 层组成，每一个音素被编码成相应数字后映射成 256 维度的嵌入码。解码器由 Pre-Net、LSTM 和 Post-Net 组成。Pre-Net 中 2 层的卷积层和 Post-Net 中 5 层卷积层的卷积核大小均为 3。对上述数据集对应的文本数据均进行归一化、分词等操作，将汉字拆解成声母和韵母后，对每一个发音单元使用字母和数字代替。例如，“wo3”拆解成“w”“o”和“3”后，根据字典对每一个音素编码，并使用 256 维的嵌入码作为文本编码器的特征输入。

HiFi-GAN 声码器的网络结构与文献[17]中 V1 基本一致，不同的是上采样层的上采样比例从 8、8、2 和 2 修改为 5、5、4 和 2，同时卷积核大小修改为 10、10、8 和 4。本文使用的声码器模型来自网上的开源预训练模型[18]，并非自主训练。

4.1.3　实验结果与分析

本文使用 MOS 分作为语音合成质量的评判指标，MOS 是一种主观评估方法，广泛地应用于语音合成质量的评价，也是现阶段最具代表性的评价方法之一。自然度 MOS 中的 5 分表示合成语音质量较好，1 分表示合成语音失真严重。相似度 SMOS 最高分同样为 5 分，表示与原说话人语音高度相似。本文中的主观 MOS 评分和 SMOS 评分均是由 20 位评测人主观评分的 95% 置信空间表示。

如表 1 所示，本文首先对健听人和聋哑人真值 (ground truth, GT) 以及直接合成的语音进行了 MOS 的评分。表 1 中，“合成”表示使用 Tacotron2 作为声学模型，HiFi-GAN 作为声码器直接合成语音的结果。健听人 GT 的 MOS 为 4.49±0.09，合成语音 MOS 为 4.22±0.17，说明在健听人的实验中合成语音的质量较好。而轻残聋哑人 GT 的自然度 MOS 仅为 2.94±0.19，这是由于轻残聋哑人本身的发音不清晰、不流畅导致的。对于重残聋哑人来说，其语音与轻残聋哑人相比不仅更难听懂，而且悦耳度较差，因此其 GT 的 MOS 仅为 2.40±0.25。同时由于重残聋哑人的发音不稳定，导致无法使用其语音进行直接训练。

表 1　健听人与聋哑人的真值 (GT) 及直接合成语音的质量评判指标 (MOS)

Table 1　MOS of GT and direct synthesis of the healthy and deaf people

说话人	类型	语音自然度MOS
健听人	GT	4.49±0.09
	合成	4.22±0.17
轻残聋哑人	GT	2.94±0.19
	合成	2.77±0.29
重残聋哑人	GT	2.40±0.25
	合成	-

本文把自然度 MOS、音色相似度 SMOS、WER 以及 SWER 作为评价指标，评价该方法是否具有实际的可行性。实验选取男性 dm002 和女性 dm004 的语音进行评估。由于语音转换模型仅需说话人提供 1 句即可进行转换，本文首先对 dm002 与 dm004 聋哑人的语音进行主观的选取，选取语音的标准为发音最稳定、最悦耳且连续性高的语音。语音转换和合成后的评价结果如表 2 所示，表 2 中“转换”表示语音转换后的 log-Mel 使用 HiFi-GAN 声码器作为语音的输出。表 2 中“转换+合成”表示语音转换后的频谱作为 Tacotron2 的训练输入，同样使用 HiFi-GAN 声码器作为语音

表 2 健听人和聋哑人的 GT、语音转换和合成结果的自然度、可懂度和相似度

Table 2 GT and the naturalness, intelligibility and similarity of voice conversion and synthesis of the healthy and deaf people

说话人	类型	语音自然度MOS	音色相似度SMOS	WER/%	SWER/%
健听人	GT	4.48±0.09	-	5.42	0.00
	转换	3.35±0.31	3.37±0.28	25.56	15.75
dm002聋哑人	GT	2.53±0.35	-	100.00	94.00
	转换	2.94±0.19	3.13±0.31	72.50	58.44
	转换+合成	2.88±0.17	2.98±0.27	80.77	56.77
dm004聋哑人	GT	3.06±0.46	-	100.00	56.42
	转换	3.32±0.27	3.24±0.22	61.25	40.22
	转换+合成	3.21±0.30	3.11±0.40	76.91	45.14

的输出。

由表 2 可得，用于语音转换前的健听说话人 GT 的平均 MOS 为 4.48±0.09，且语义的可懂度较为清晰，WER 为 5.4%(通过对 AISHELL-1、标贝标准女声以及 AISHELL-3[26]三个数据集的 100 个字测试，表明在健听人的语音识别中具有较好的性能)，SWER 为 0%(全部的字主观可懂)。转换之后的语音自然度 MOS 和可懂度均有略微的下降，MOS 下降的主要原因是转换后的对数梅尔谱，而可懂度下降主要是由于合成音质质量的下降导致的。从轻残聋哑人 dm002 以及 dm004 的 GT 评价指标可知，虽然轻残聋哑人的语音较为连贯且 SWER 不为 100%，表明在没有文本的提示下，部分测试人员能略微理解聋哑人语义。但是从语音识别的结果可知，其表达内容在客观的评价中是不具有识别能力的，WER 均为 100%。转换之后的语音 MOS 得到了较高地提升，而且 WER 和 SWER 均得到了大幅降低。SMOS 分能达到 3 左右的评分，表示转换的音色仍然具有聋哑人的特征。而在 Tacotron2 语音合成之后，dm002 和 dm004 的自然度 MOS、音色 SMOS、WER 和 SWER 均有略微地下降。与 GT 相比，合成的语音在可懂度、悦耳度上有明显的提升，表明该方法能较好地解耦说话人音色以及内容，并可以应用在轻残聋哑人的语音转换与合成中。

4.2 面向重残聋哑人的语音克隆方法

4.2.1 重残聋哑人语音克隆数据集

(1) 说话人编码器数据集：使用 King-ASR-459 数据集作为说话人编码器的训练数据集。训练时，共选取 1 800 位说话人的 185 930 条语料进行训练。采用 80 维的 log-Mel 作为训练的语音特征，提取时帧长和帧移分别设置为 25 ms 和 10 ms。在提取之前，对音频进行降采样 (采样率降至 16 kHz)、VAD、归一化和预加重等处理。训练时采用开源的噪声数据库 MUSAN[27]以及 RIR[28]数据集分别对训练语音进行加噪和加混响处理。

(2) 合成器数据集包括 ESD 情感数据集和 aidatatang_200hz 数据集。本文使用 ESD 作为情感语音合成的数据集。该数据集包含生气 (angry)、开心 (happy)、中立 (neutral)、沮丧 (sad) 和惊喜 (surprise) 共 5 种不同的情感风格。数据集中有 20 位不同的说话人，其中 0001 号~0010 号说话人是母语为中文的说话人，由 5 位女性说话人和 5 位男性说话人组成。而 0011~0020 号说话人是母语为英语的说话人。训练时仅选择中文说话人的语料，每位说话人的每种感情风格有 1 500 条语料，10 位说话人共 15 000 条语料，时长约为 31 h。而 Aidatatang_200hz 大型多说话人数据集作为补充数据集，Aidatatang_200hz 录制总人数约为 600，总共约 12 000 条语料。录制的采样率为 16 kHz。由于多说话人 ESD 数据集语料有限，且易提取语音中的基频特征，因此引入的 Aidatatang_200hz 多说话人数据集能使模型更充分地学习更多说话人的音色信息和风格信息。

4.2.2 模型配置

ECAPA-TDNN 的网络结构与文献[13]中一致，最后输入 192 维的向量作为说话人嵌入码。训练时，批大小为 64，优化器使用 Adam，共训练 80 轮。初始学习率为 0.001，每迭代 5 次学习率减小至原来 90%。Tacotron2 和 HiFi-GAN 的结构如 2.2 和 2.3 节所示。风格嵌入码与合成器文本特征的输出均为 256 维，本文拼接了说话人嵌入码和风格嵌入码后，特征输出的维度增加至 704(256+256+192)。后经过线性层映射成 128 维作为解码器的输入。

4.2.3 实验结果与分析

(1) 说话人识别结果。本文使用闭集 King-ASR-459 以及开集 AISHELL-1[29]、轻残聋哑人数据集 (5 人) 和重残聋哑人数据集 (10 人) 进行说话人识别的模型评估。AISHELL-1 共 178h 的语音，

音频采样率为 16 kHz，共包含 400 位说话人，其中训练集、测试集和验证集分别包含 340、20 以及 40 位说话人的语料。轻残聋哑人的数据集有 5 位说话人，每人 300 句语料。重残聋哑人有 10 位说话人，每人同样 300 句语料。表 3 采用等错误率 (equal error rate, EER) 和最小检测代价函数 (minimum detection cost function, minDCF) 来评估 ECAPA-TDNN 与模型 LSTM 的说话人识别性能以及对聋哑人语音与健听人语音数据集的适配性。

表 3　不同说话人编码器的识别性能

Table 3　Speaker recognition performances of different speaker encoders

模型	训练数据集	验证数据集	EER	MinDCF
LSTM(基线)	KingASR-459	AISHELL-1	1.120	0.098
		轻残聋哑人	2.678	0.107
		重残聋哑人	33.622	0.763
ECAPA-TDNN	KingASR-459	AISHELL-1	0.601	0.043
		轻残聋哑人	2.069	0.058
		重残聋哑人	30.494	0.649

测试时，挑选各说话人最长的语料作为注册语料。从模型的性能上分析，相比于 LSTM（基线）模型，无论是健听人的数据集还是聋哑人的数据集，ECAPA-TDNN 模型在各验证数据集的效果均优于基线模型。AISHELL-1 数据集 ECAPA-TDNN 模型的 EER 比基线模型相对降低了 46%，而对于轻残聋哑人与重残聋哑人的数据集，分别相对降低了 8% 和 9%。从数据集上看，AISHELL-1 数据集的 EER 最低，识别性能最好，虽然轻残聋哑人的识别性能略微地降低，但仍然有较好的识别性能。重残聋哑人的说话人识别性能较差。可能是因为重残聋哑人的语音幅值变化过大且发音不稳定，因此无法直接使用健听人语料训练的说话人识别系统获取重残聋哑人单个语句的音色嵌入码。在健听人的语音克隆中，较少使用平均嵌入码的方式进行克隆。即使健听人的说话人识别性能较好，但同一说话人不同语料之间的嵌入码的分布也会存在差异，例如性能较优模型的余弦相似度仅达到 0.7。若对其作平均处理，会出现音色不稳定的情况。本文提出使用重残聋哑人的多个余弦相似度较高的语料平均嵌入码 (5 句) 作为其说话人的表征，该方法能改善重残聋哑人音色不稳定的问题，从而获得更稳定的音色嵌入码。

(2) 基频情感风格分类结果。基于基频特征的情感语音克隆的实验中，本文首先使用 ESD 情感数据集进行基频情感分类。分类结果如图 7 所示，总的分类结果的准确率为 64.8%。生气、开心、中立、沮丧和惊喜五种情感的测试数量分别为：999、998、998、1 000 和 997。其中，识别性能最好的是中立和沮丧情感，这两种的准确率均能达到 70.0%。而识别最差的为开心情感。从图 7 中可以看出，开心和惊喜两种情感的识别结果相比于其他情感，比较容易混淆。上述结果的原因可能为：① ESD 数据集中五种情感均是人为表演，其情感表达能力与实际情感有一定差距。② 开心和惊喜两种情感均是表达说话人积极向上的情感。因此，此两种情感本身也有比较高的相似度，所以识别结果相比于其他的情感略差。

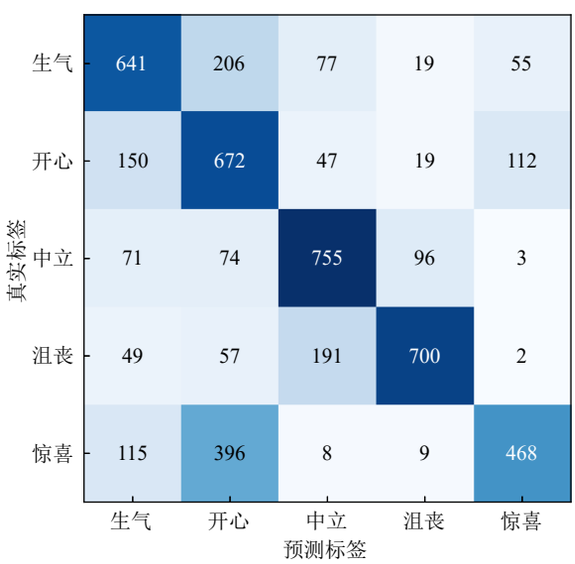

图 7　情感分类结果

Fig.7　Result of emotion classification for pitch

(3) 语音克隆模型性能分析。本文对每种类型的说话人随机选取 2 位作为合成语音的测试对象，并对每位说话人各风格的 2 条语料进行评分 (每位说话人共 10 条)，结果如表 4 所示。由表 4 可知，闭集说话人的自然度 MOS 和音色 SMOS 均较高，均能达到 3.5 分以上。在风格 SMOS 方面，情感表现力最好的是中立，其次是沮丧，而生气仅有 2.63 的平均分。从表 4 可以看出，模型在闭集的情况下，生成中立语音质量最优，无论是 MOS 和 SMOS 均能达到 4 分以上，而合成风格相似度最差的是生气。至于开集健听人，除了沮丧风格之外，其他风格合成语音的自然度 MOS 没有明显的下降。音色 SMOS 最高的为中立，其他风格的相似度有较明显的下降，这是由于开集说话人在注册时仅有中立风格的语音。风格 SMOS 中表现最好的是沮丧风格。从开集重残聋哑人的结果中可以发现，合成的 MOS 分比聋哑人原音有较大地提升，其原因是语音克隆是在健听人的大型数据集中训练的，因此合成的是比较正常的语音。而各风格的音

表 4 基于基频的风格迁移语音克隆效果
Table 4 Results of pitch-based style transfer voice cloning

说话人类型	情感	GT	语音自然度MOS	音色相似度SMOS	风格相似度SMOS
闭集健听人	生气	4.71±0.13	3.75±0.38	3.67±0.32	2.63±0.33
	开心	4.64±0.18	3.94±0.41	3.71±0.38	3.13±0.36
	中立	4.54±0.16	4.11±0.40	4.25±0.24	4.21±0.17
	沮丧	4.53±0.16	3.94±0.38	3.93±0.31	3.56±0.28
	惊喜	4.67±0.14	3.75±0.39	3.56±0.35	3.17±0.33
开集健听人	生气	-	3.59±0.24	3.31±0.50	2.88±0.59
	开心	-	3.34±0.29	3.12±0.53	2.69±0.58
	中立	4.81±0.13	3.78±0.24	3.61±0.52	3.75±0.52
	沮丧	-	3.12±0.33	3.25±0.52	4.05±0.51
	惊喜	-	3.56±0.24	3.31±0.55	3.31±0.66
开集重残聋哑人	生气	-	3.44±0.31	2.96±0.50	2.56±0.37
	开心	-	3.81±0.27	3.21±0.43	2.43±0.42
	中立	2.81±0.57	3.75±0.28	3.21±0.47	3.25±0.39
	沮丧	-	3.56±0.25	3.39±0.35	4.00±0.49
	惊喜	-	3.31±0.29	3.01±0.41	2.44±0.36

色相似度 SMOS 除生气风格以外，均能达到 3 分，表明合成音色与原聋哑人音色有一定的相似度。虽然风格相似度 SMOS 上表现力不足，但仍然能满足输出不同的风格语音的需求，合成出带有不同风格特征的语音。

5 结 论

针对目前国内外对于聋哑人语音研究较少的问题，本文分析了不同程度聋哑人的语音特征，并创新性地对不同残疾程度的聋哑人语音提出两种方法，提高了可懂度、自然度。本文提出可用于提高聋哑人自身语音可懂度和自然度的中文语音合成系统。实验结果表明，面向轻残聋哑人的语音转换和合成方法能大幅提高聋哑人语音的可懂度，且合成的语音具有一定的相似度。面向重残聋哑人的语音克隆方法能通过说话人编码器 ECAPA-TDNN 提取准确的说话人表征，并根据重残聋哑人的平均嵌入码合成出与该说话人相似的音色。除此之外，引入的风格迁移模块能丰富输出语音的风格，使输出语音情感表现力更好。虽然模型在开集说话人的风格相似度上有待提高，但相比于传统的端到端语音合成，本文所提出的模型仍然能合成出多种不同风格的语音，提高了语音的表现力。

参 考 文 献

[1] LIGHT J. "Communication is the essence of human life": reflections on communicative competence[J]. Augmentative and Alternative Communication, 1997, **13**(2): 61-70.

[2] WANG Y X, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[EB/OL]. 2017: arXiv: 1703.10135. https://arxiv.org/abs/1703.10135.

[3] SHEN J, PANG R M, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 4779-4783.

[4] Van DEN OORD A, DIELEMAN S, ZEN H G, et al. WaveNet: a generative model for raw audio[EB/OL]. 2016: arXiv: 1609.03499. https://arxiv.org/abs/1609.03499.

[5] REN Y, HU C X, TAN X, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[EB/OL]. 2020: arXiv: 2006.04558. https://arxiv.org/abs/2006.04558.

[6] NEEKHARA P, HUSSAIN S, DUBNOV S, et al. Expressive neural voice cloning[EB/OL]. 2021: arXiv: 2102.00151. https://arxiv.org/abs/2102.00151.

[7] BIADSY F, WEISS R J, MORENO P J, et al. Parrotron: an end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation [EB/OL]. 2019: arXiv: 1904.04169. https://arxiv.org/abs/1904.04169.

[8] VALLE R, LI J, PRENGER R, et al. Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. IEEE, 2020: 6189-6193.

[9] SKERRY-RYAN R, BATTENBERG E, XIAO Y, et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron[EB/OL]. 2018: arXiv: 1803.09047. https://arxiv.org/abs/1803.09047.

[10] CHEN Y H, WU D Y, WU T H, et al. Again-VC: a one-shot voice conversion using activation guidance and adaptive instance normalization[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada. IEEE, 2021: 5954-5958.

[11] JIA Y, ZHANG Y, WEISS R J, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada. New York: ACM, 2018: 4485-4495.

[12] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: emphasized channel attention, propagation

and aggregation in TDNN based speaker verification[EB/OL]. 2020: arXiv: 2005.07143. https://arxiv.org/abs/2005.07143.

[13] ULYANOV D, VEDALDI A, LEMPITSKY V. Instance normalization: the missing ingredient for fast stylization [EB/OL]. 2016: arXiv: 1607.08022. https://arxiv.org/abs/1607.08022.

[14] KINGMA D P, WELLING M. Auto-encoding variational Bayes[EB/OL]. 2013: arXiv: 1312.6114. https://arxiv.org/abs/1312.6114.

[15] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. IEEE, 2017: 1510-1519.

[16] GRAVES A. Supervised sequence labelling with recurrent neural networks: long short-term memory[M]. Berlin: Springer, 2012: 37-45.

[17] KONG J, KIM J, BAE J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis [EB/OL]. 2020: arXiv: 2010.05646. https://arxiv.org/abs/2010.05646.

[18] BABYSOR. MockingBird: realtime voice clone for Chinese[EB/OL]. (2021-10-23)[2023-4-3]. https://github.com/babysor/MockingBird/tree/v0.0.1.

[19] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, **43**(2): 652-662.

[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 7132-7141.

[21] YU L, GAO Y, ZHOU J, et al. Multi-layer feature aggregation for deep scene parsing models[J]. arXiv: 2011.02572v1 [cs. CV] 4 Nov 2020.

[22] OKABE K, KOSHINAKA T, SHINODA K. Attentive statistics pooling for deep speaker embedding[EB/OL]. 2018: arXiv: 1803.10963. https://arxiv.org/abs/1803.10963.

[23] MORISE M, YOKOMORI F, OZAWA K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications[J]. IEICE Transactions on Information and Systems, 2016, **E99.D**(7): 1877-1884.

[24] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. 2015: arXiv: 1502.03167. https://arxiv.org/abs/1502.03167.

[25] 胡月文, 李丹. 基于 ArcFace 损失函数的监护安全人脸识别[J]. 现代信息科技, 2020, **4**(17): 132-135.
HU Yuewen, LI Dan. Face recognition for guardianship security based on ArcFace loss function[J]. Modern Information Technology, 2020, **4**(17): 132-135.

[26] SHI Y, BU H, XU X, et al. AISHELL-3: a multi-speaker mandarin TTS corpus and the baselines[EB/OL]. 2020: arXiv: 2010.11567. https://arxiv.org/abs/2010.11567.

[27] SNYDER D, CHEN G G, POVEY D. MUSAN: a music, speech, and noise corpus[EB/OL]. 2015: arXiv: 1510.08484. https://arxiv.org/abs/1510.08484.

[28] ALLEN J B, BERKLEY D A. Image method for efficiently simulating small-room acoustics[J]. The Journal of the Acoustical Society of America, 1979, **65**(4): 943-950.

[29] BU H, DU J Y, NA X Y, et al. AISHELL-1: an open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul, Korea (South). IEEE, 2018: 1-5.