面向非平行语料的语音转换技术综述

李鹏程[1,2]，张旭龙[1]，王健宗[1]，程宁[1]，肖京[1]
1. 平安科技（深圳）有限公司，广东 深圳 518063；
2. 中国科学技术大学，安徽 合肥 230026

摘要
语音转换是语音及人工智能领域的一项研究课题，其目标是在保持源语音内容不变的情况下改变语音的音色，使其听上去像是由另一个目标说话人说出的，同时还需保证语音的质量和自然度。面向非平行语料的语音转换技术是当下的热门研究内容，其使用非平行的多说话人语音数据集进行模型训练，能完成多对多以及任意对任意的语音转换。对近年来面向非平行语料的语音转换进行了全面的总结和分析。首先概述了早期面向平行语料的语音转换及其缺陷，然后对当下面向非平行语料的语音转换的各类实现方法进行介绍和对比分析，最后对语音转换技术进行了总结和展望。

关键词
语音转换；人工智能；深度学习

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024011

A survey of voice conversion based on non-parallel data

LI Pengcheng[1,2], ZHANG Xulong[1], WANG Jianzong[1], CHENG Ning[1], XIAO Jing[1]
1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518063, China
2. University of Science and Technology of China, Hefei 230026, China

Abstract
Voice conversion is a research topic in the fields of speech and artificial intelligence. The goal of voice conversion is to change the timbre of speech while preserving the content of the source speech, making it sounds like spoken by the target speaker. It is essential to ensure both the quality and naturalness of the converted speech. Voice conversion based on non-parallel data gains much attention currently, where models are trained using non-parallel multilingual speaker datasets, enabling many-to-many and any-to-any voice conversions. This paper provides a comprehensive summary and analysis of recent developments in non-parallel voice conversion. Firstly, we outline the early voice conversion techniques based on parallel corpus and their limitations. Then, we introduce and compare various approaches to voice conversion based on non-parallel data, providing a thorough analysis. Finally, a summary and outlook on voice conversion technology is provided.

Key words
voice conversion, artificial intelligence, deep learning

0 引言

语音转换研究在不改变语音内容的情况下将一名说话人的语音片段（源语音）转换为与另一说话人（目标说话人）发音相似的说话片段。语音转换可视作语音领域的风格迁移，它将目标说话人的音色迁移到源语音上，使输入语音与输出语音间的音色发生改变而内容保持一致。语音转换曾被归为语音合成的一个分支，语音合成主要任务有文本到语音的合成以及改变语音属性（如说话人身份、说话情感等）的合成，而语音转换的研究重点便是改变语音的说话人身份信息。

语音转换能应用于许多不同领域，如语音身份隐私的保护[1-2]、语音伪装和模仿[3-5]、语音增强技术[6-7]等；或用于实现风格化语音的合成，如实现个性化语音助手[8-10]和改变语音的情感或韵律[11-13]。也有研究在医疗领域的语言障碍患者的语音辅助设备中引入语音转换技术[14-15]。近年来随着深度学习技术的快速发展，神经网络建模技术在语音领域各类任务（如语音识别、语音合成、说话人验证）中取得了可观的成果，随之大量研究将神经网络用于语音转换任务，基于神经网络的语音转换技术在自然度、相似度等方面效果提升明显。

优秀的语音转换方法应该具备高质量的语音合成能力，生成清晰度高、自然度高、失真少的转换语音。同时应该有优秀的说话人特征和内容特征保持能力，即转换后的语音应该与目标说话人的语音片段的音色保持一致，并且转换后的语音与源语音内容相同。随着语音转换技术的不断进步，研究者对这项任务提出了更高的要求，如低延迟和高效性，使语音转换技术能应用于实时转换的场景，并具备硬件资源方面使用的高效性，以便于在各种环境中运行；语音转换的鲁棒性则要求模型能应对场景噪声、不同录制设备、不同语速带来的影响，以适应语音转换的实际应用场景。此外，一种优秀的神经网络语音转换方法不应该依赖于平行语料进行设计和训练。由于平行语料数据集的制作难度高、耗时大，数据集的体量难以扩充，训练过程一般需要复杂的对齐操作，平行语料的语音转换方法难以推广，并且这类方法通常只能对数据集中出现的说话人进行语音转换，无法面对多说话人对多说话人或任意说话人到任意说话人的语音转换场景。近年来，大部分语音转换研究工作致力于面向非平行语料的零样本语音转换研究，使用普通的非平行多说话人数据集进行模型训练，摆脱对平行语料的依赖，并能完成任意对任意的语音转换任务，在推理时只需要一段目标说话人的语音片段完成转换，且训练数据中不存在该目标说话人的语料数据。

本文将对语音转换的发展以及面向非平行语料的语音转换技术进行概述。面向非平行语料的语音转换方法相较于平行语料的语音转换在数据依赖、灵活性、可扩展性等方面有明显的优势，仍是未来语音转换技术的发展方向。本文在最后对语音转换技术相关工作进行了总结和展望。

1 语音转换

1.1 语音转换系统

一般的语音转换系统由3个模块组成[16]：语音分析模块、映射模块和语音重构模块，如**图1**所示。其中语音分析模块将源语音分解为能表示分段信息的特征；映

射模块将这些特征向目标说话人的音色进行改变和迁移；最后通过语音重构模块将映射得到的特征重构为转换语音。说话人风格迁移是语音转换的本质，映射模块是语音转换模型训练的核心，同时也是众多相关工作的研究重点。

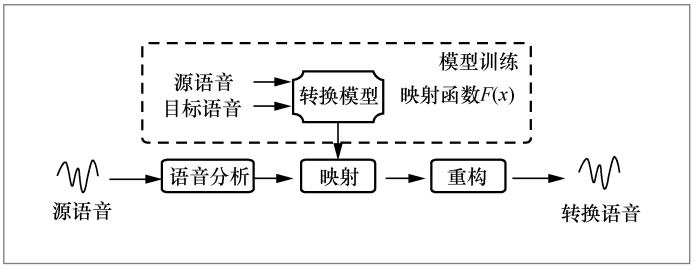

图1 语音转换系统一般构成

1.2 平行语料的语音转换

早期的语音转换方法依赖于平行语料数据集，平行语料也就是由多个说话人录制的相同话语的数据，具有内容相同而音色不同的特点，易于从中学习不同说话人话语间的映射关系。这些平行语料的语音转换方法通常需要进行语料的对齐操作，一般使用动态时间变换[17]（dynamic time warping）技术实现对齐。早期面向平行语料的语音转换方法可分为两类：参数统计方法和非参数统计方法。参数统计方法对数据的分布做出明确的假设，模型通过估计这些分布的参数来拟合数据，这类方法高度依赖训练数据的数量，更多的训练数据能带来性能的有效提升。这类方法的代表工作有基于高斯混合模型（gaussian mixture model，GMM）的语音转换[18-20]、偏最小二乘回归（partial least square regression，PLSR）语音转换[21]和动态核偏最小二乘回归（dynamic kernel partial least square regression，DK-PLSR）语音转换[22]。上述回归方法通过建立输入和输出间的关系模型来进行语音转换，其中动态核偏最小二乘回归通过引入核技巧来处理非线性关系。另一类使用非参数统计方法的代表工作为基于非负矩阵分解（non-negative matrix factorization，NMF）的方法[23]，非参数统计方法在参数统计方法的基础上降低了对训练数据数量的要求和依赖。基于统计方法的语音转换为基于深度学习的语音转换技术提供了基础。

但这些依赖于平行语料数据集的语音转换方法存在较大的局限性。一方面，由于平行语料数据集的制作成本高昂，扩充数据集中数据的数量会带来远高于非平行语料数据集的成本。另一方面，数据集中说话人数量有限，这些方法大多只能用于单说话人到单说话人的语音转换或多说话人到单说话人的语音转换，无法满足现实中对语音转换的要求，即任意说话人到任意说话人的转换场景。

1.3 深度学习下的语音转换

基于神经网络的语音转换方法可追溯到20世纪[24]，近年来神经网络在各领域广泛应用，大量研究将神经网络应用于各式语音任务，如语音合成、语音识别等。同时在语音转换领域，神经网络也发挥了至关重要的作用。深度学习极大地提升了合成语音的质量，如基于深度学习的声码器[25-29]能高效地将声学特征（如梅尔谱图）还原为声音波形，语音转换模型的语音重构模块能由此大幅提升转换语音的质量。另外，深度学习能使映射模块从大量的数据中学习，极大的数据量使模型能有效提升转换语音质量以及与目标说话者的相似度。同时，这也在一定程度上反映了大数据驱动的深度学习下的语音转换对数据量的依

赖，因此用到的数据集往往是非平行语料数据集，这样的数据集相较于平行数据集更易于制作和扩充，非平行语料数据集一般由多个说话人的大量话语构成，每个说话人的话语内容可以不一致。深度学习极大促进了语音合成及语音转换的发展，大量的端到端语音合成模型被提出[30-32]，端到端的语音转换系统[33]也由此出现。端到端的语音转换不再遵循一般的语音转换系统框架，神经网络声码器与整个模型一同训练，不会使用单独的声码器进行语音的重构。深度学习为面向非平行语料的语音转换技术的出现和发展提供了强大的推动力。

2 面向非平行语料的语音转换

依赖于大量数据的神经网络语音转换模型通常使用非平行语料数据集进行训练，这样的数据集制作成本较平行语料数据集更低。此外，图1中语音分析和语音重构过程中推断低阶细节不应该是语音转换任务的工作[16]，因为这些工作可以通过训练神经网络声码器来完成[34]，而语音转换模型只需将工作重点放在更高阶的语音特征学习，也就是如何学习说话人表示特征、不同说话人之间的映射。神经网络语音转换若要在非平行语料中学习映射，需要从语音中提取出充足的中间表示。深度学习中引入了嵌入（embedding）作为中间表示的概念，比如在基于解耦的语音转换方法中，通常从语音中提取出表示语音内容信息的内容嵌入和表示说话者身份信息的说话者嵌入然后进行语音的重构。

2.1 早期面向非平行语料的语音转换

早期面向非平行语料的语音转换主要对语料对齐的方法进行改进，以达到非平行语料对齐的目的[35-36]。这一时期中另一个面向非平行语料的语音转换方向是基于语音后验图（phonetic posteriorgrams，PPG）[37]的语音转换[38-40]，这种方法将PPG作为中间特征来表征源语音的内容信息，并借助自动语音识别器（automatic speech recognizer，ASR）来提取PPG。值得一提的是，初期基于语音后验图的语音转换只能完成多对一的语音转换，随着深度学习在语音及语音转换领域的发展，近年仍有部分语音转换工作在语音后验图的基础上进行[41-42]，有研究[43-44]基于语音后验图设计跨语言的语音合成或语音转换方法。此外，随着语音识别的发展，使用自动语音识别器的基于语音后验图的语音转换方法一般在词错误率的控制上有较好的表现。

2.2 域间迁移的语音转换

语音转换是一项对说话者身份进行改变的任务，这与图像领域的风格迁移任务（图片到图片的转换，image-to-image）[45]相似。生成对抗网络[46]（generative adversarial network，GAN）作为一种杰出的生成模型，应用于各个领域。生成对抗网络主要包含生成器（generator）和判别器（discriminator）两部分，它们相互对抗，通过竞争性的学习来不断提升模型的性能。其中生成器的目标是生成与训练数据相似的合成数据，通过学习逐渐生成能与训练数据难以分辨的数据，在不同的任务中这些生成的数据可以是图像、文本、语音等；而判别器由另一个神经网络构成，其目标是将生成器产生的合成数据与真实的训练数据分辨开，接收来自生成器的合成数据和真实数据作为输入，并尝试判断哪个是真实数据，生成器和判别器由此相

互对抗而相互促进。大量研究将生成对抗网络应用于图像风格迁移，随之在语音领域，研究者将这些方法成功应用于语音转换[47-50]。CycleGAN-VC[47]是一种使用生成对抗网络的经典语音转换方法，该模型将来自两个说话人域的语音分别进行多次风格迁移，并设计对抗损失来鉴别转换语音的自然程度，循环一致性损失对经过两次域间迁移的语音和原始语音进行比较，全等映射损失则衡量生成器对原本属于该生成器目标域的语音进行风格迁移的影响。其中对抗损失来自网络中的两个生成器和两个判别器，如**图2**所示，对抗损失衡量生成器的输出是否接近真实的人类语音；循环一致性损失用于防止GAN模型坍塌，即对任何输入都输出一个相同的结果，产生循环一致性损失的方法是将一段语音经转换到目标说话人的域后再次转换到源说话人的域，要求经过两次转换的语音尽可能和原始语音相似；全等映射损失补充循环一致性损失未考虑到的更极端的情况，计算该损失的方法为将源语音通过生成器转换到源说话人的域，要求转换前后的语音保持尽可能的一致，对目标域的生成器也进行类似的全等映射损失计算。

由此可以发现基于CycleGAN的语音转换模型训练过程中损失函数的计算均不需要用到平行语料，而仅需将各说话人的任意语句进行生成器的域间迁移训练。CycleGAN-VC成功作为后续基于GAN的语音转换模型的基础。StarGAN-VC[48]是另一种基于生成对抗网络的语音转换模型，具有多对多转换的能力。StarGAN-VC在构建生成器网络的同时学习多个域之间的映射，由此完成多对多的语音转换。此外，后续还有工作将基于CycleGAN等语音转换方法用于语音转换中的跨语言语音转换[51]、韵律转换[52]和情感语音转换[53]等任务中。生成对抗网络作为一种生成模型，不需要像自编码器模型那样考虑中间特征表示，将其用于语音转换时没有显式地对说话者身份进行提取，表现灵活。同样地，对于语音组成的韵律、音调、情感等构成属性，生成对抗网络模型将这些语音构成的改变转换为不同域之间的迁移转换，以完成特定的语音属性转换任务。

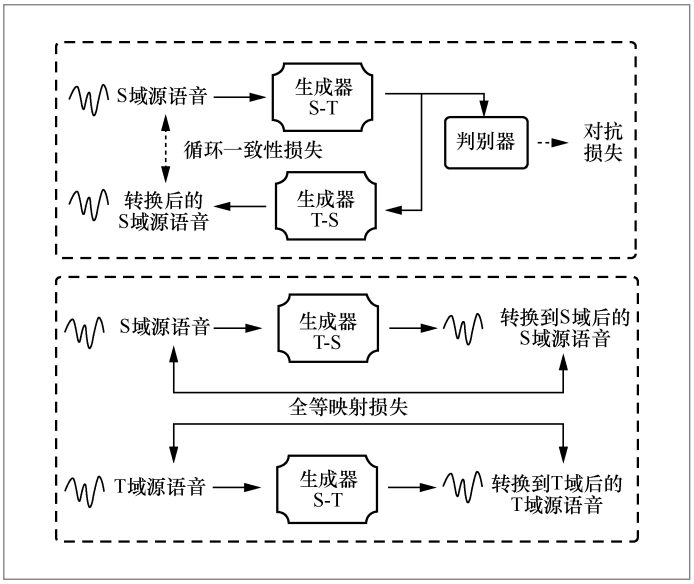

图 2 CycleGAN-VC 语音转换模型及训练过程

2.3 基于文本的语音转换

语音转换与语音合成存在许多联系，二者有不少相似之处，如语音合成和语音转换的目标都属于生成特定内容的语音，区别在于语音合成一般根据文本内容生成（文本转语音，TTS），而语音转换则从源语音中获取内容。由此容易联想到将较为成熟的TTS语音合成技术应用到语音转换任务中，使用文本等作为内容一致的中间保证，这种方法的优势在于使用大型数据集训练出的TTS模型能有效保证合成语音的质量。研究者将TTS用于语音转换中的语音重构模块，这类语音转换模型一般为编码器-解码器结构，如**图3**所示。序列到

序列语音合成模型Tacotron[54]被研究者引入到了语音转换中[55-56]，构建出编码器-解码器结构的语音转换模型。其中文献[8]提出一种语音合成和语音转换的联合训练框架，根据输入的类型（文本、语音、文本加语音）来执行语音合成或语音转换。文献[57]提出一种迁移学习技术，从TTS中注意力机制导出的语音上下文向量中学习，与TTS共享解码器。Cotatron[58]语音转换模型建立在多说话人Tacotron语音合成模型基础上，使用预训练的TTS模型来推导源语音中与说话人无关的特征，此过程需要同时输入语音的文本转录，然后解码器以目标说话人为条件，根据推导出的与说话人无关的特征生成目标语音，该语音转换模型能执行一对多的语音转换。

基于文本的语音转换另一系列工作则将自动语音识别系统用于语音转换模型中的语音分析模块，使用ASR提取的上下文后验概率序列来生成目标语言的特征序列[56]，许多研究使用ASR提取PPG作为中间特征进行内容表示[38-39,41]，再根据PPG重构语音，与引入TTS的语音转换方法一样，引入ASR的语音转换方法能有效保证合成语音内容的一致。这些方法将ASR作为编码器，用语音合成模型作为解码器。使用ASR能有效地解耦出语音信号中的内容信息。也有文献[55]设计瓶颈特征增强输入，使用ASR对序列到序列（sequence-to-sequence）的语音转换模型进行指导。在其他语音转换的任务中，文献[59]提出了一种PPG到波形的转换方法。文献[43-44]使用PPG进行跨语言的语音转换，文献[60]提出使用PPG进行情感语音转换。与后文将介绍的基于特征解耦的语音转换方法类似，这类语音转换模型使用ASR来提取与说话人无关的内容信息，将模型训练的重点放在从多说话人、非平行语料中学习映射模块，从而不再像域间迁移那样为每个说话人训练一个转换模型，有效实现多说话人对多说话人乃至任意说话人对任意说话人的语音转换。此外，有研究工作[58,61]使用文本作为内容提取的指导（text-guide），在语音转换训练过程中输入语音的文本转录，通过分别对文本和语音内容编码到同一空间，减小二者的差异以达到准确从语音中提取内容的目的。文献[58]和文献[61]分别在Tacotron[54]和Transformer[62]的基础上实现。这种文本指导的方法要求数据集含有语音对应的文本转录数据。这些引入TTS或ASR的语音转换方法都是面向非平行语料的，一般在训练过程中每次只需要输入一段语音进行重构，且引入的TTS和ASR模型大多是在大型数据集上进行预训练的。也有研究者考虑数据量受限的情形，提出低资源场景下的语音转换模型[63-64]，使用说话人自适应的语音合成模型减少对数据量的依赖。

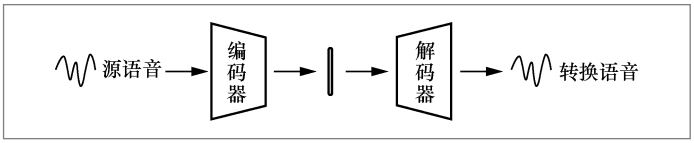

图 3 编码器 – 解码器结构的语音转换模型

2.4 基于特征解耦的语音转换

基于解耦的语音转换的基本设定是将语音信号分解为说话人不相关信息和说话人相关信息，使用不同的方法将这两部分信息分别从源语音和目标语音中解耦出后，使用源语音中解耦的说话人不相关信息和目标语音中解耦的说话人相关信息进行语音重构，输出含有目标说话人信息和源语音说话人不相关信息的转换语音。这类方法是当下语音转换领域的热点。使用自编码器结构[65]是这类方法的主流，**图4**展示了基于特征解耦的语音转换模型的一

般结构，由变分自编码器网络（variational autoencoder，VAE）实现。由两个编码器分别从语音中提取说话人表示特征（说话人嵌入）和内容表示特征（内容嵌入），再由解码器根据说话人表示特征和内容表示特征合成语音。在训练阶段执行语音的重构，即不区分源语音和目标语音，每次向说话人编码器和内容编码器中输入同一段语音，从这段语音中分别提取出说话人表示特征和内容表示特征后再由解码器重构这段语音，重构语音和原始语音间的差异（一般使用L1损失）作为重构损失。在推理阶段说话人编码器和内容编码器分别输入目标语音和源语音进行语音转换。因此基于特征解耦的语音转换模型不依赖于平行语料。

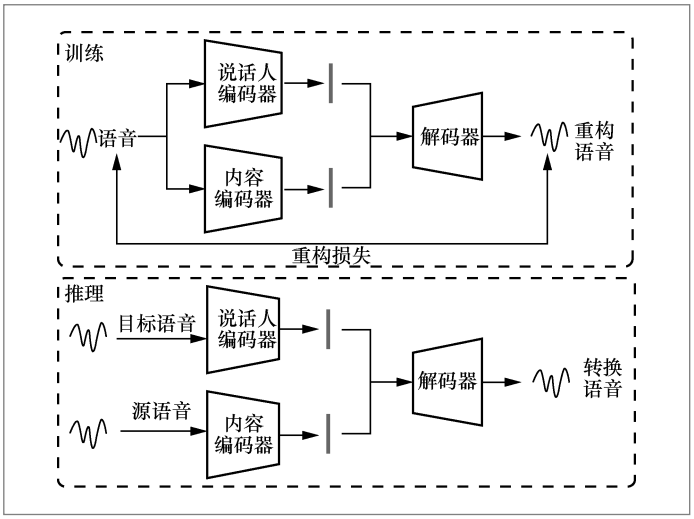

图 4 基于特征解耦的语音转换模型的一般结构及训练和推理过程

一些使用变分自编码器的语音转换方法[66-67]使用独热向量或i-vector或d-vector作为说话人表示，使用独热向量作为说话人表示的方法无法对训练数据外的说话人进行语音转换，即无法实现任意到任意或零样本语音转换。AutoVC[68]在内容编码器中设置瓶颈过滤掉与内容无关的说话人信息、噪声等，损失函数只用到了自编码器损失函数，SpeechSplit[69]使用瓶颈特征对内容表示、韵律表示和音高表示进行提取，相较于一般的内容表示加说话人表示考虑了语音的韵律和音高，提升转换后的语音在韵律上的自然度。但这类使用瓶颈特征进行解耦的方法依赖于对瓶颈的调节和把控，SpeechSplit 2.0[70]则加入音色扰动、音高平滑等信号处理手段以减小对瓶颈设置的依赖。另一系列文献[71-72]在内容编码器中加入实例归一化层来去除说话人信息，并且使用一种自适应归一化（adaptive instance normalization）的方法来将提取出的说话人信息迁移到内容表示中并合成出转换语音，Again-VC将说话人编码器和内容编码器进行合并，只使用一个编码器进行说话人嵌入和内容嵌入的提取，简化了网络的结构。基于自适应归一化的语音转换方法相较于使用瓶颈特征的方法更加灵活，网络结构一般更简洁。矢量量化（vector quantization，VQ）也是基于特征解耦的语音转换方法中的重要手段。大量零样本语音转换研究工作[73-75]使用VQ作为内容表示提取的方法，研究者发现对语音特征进行量化得到的表示在一定程度上与音素相关联，可以作为内容表示，而量化产生的损失则视为内容无关的信息，即说话者信息。VQVC+[74]在模型中使用了U-net的网络结构。VQMIVC[75]在分别使用内容编码器、说话人编码器和音高提取器提取出语音的内容表示、说话人表示和音高信息之后，再引入互信息减少这3种特征间的相互关联。

部分工作[76-77]更细粒度地对语音特征进行解耦，考虑韵律、音高等信息，从语音中提取出说话人信息和内容信息以外的特征，使语音转换模型不仅能进行说话人身份的转换，也能改变源语音的韵律或音高。文献[76]结合数据增强的思想

设计出一种自动解耦的框架。文献[77]在SpeechSplit 2.0[70]工作的基础上引入互信息强化模型的解耦能力，使用VAE作为网络框架进行语音特征解耦的语音转换模型普遍存在输出语音过度平滑的问题。有研究[78]将生成对抗网络引入这类方法中，要求生成的语音尽可能地不被判别器分辨为合成的语音，从而解决语音过度平滑的问题，提升转换语音的质量。去噪扩散概率模型（denoising diffusion probabilistic model，DDPM）[79]的出现让各类语音任务有了新的可能，已被用于声码器[80]和语音合成[81]中生成高质量的语音，也有研究[82]根据DDPM设计出解码器用于歌声语音转换。基于特征解耦的语音转换方法灵活易扩展，训练得到的模型具有普遍性，能应用于任意说话人到任意说话人的零样本语音转换。

本文对历年面向非平行语料的语音转换的经典工作进行了总结，见**表1**。

3 语音转换性能评价

本节将介绍一些语音转换工作中常用的性能评估指标，包括主观评估指标和客观评估指标两类。然后介绍一些现阶段语音转换研究常用的数据集，其中包含平行语料数据集和非平行语料数据集。

3.1 语音转换常用评估指标

3.1.1 主观评估指标

MOS（mean opinion score）常用于评估语音质量和自然度，是一种主观评估指标，广泛用于文本到语音的语音合成、语音转换等任务。在MOS值的测试实验中，参与测试的人员对语音的质量进行评分，分数从1分到5分，分越高则代表语音质量越好，一般能体现出语音在清晰度、自然度和易理解程度等方面的水平。

VSS（voice similarity score）用于评估语音转换结果与目标说话人语音的相似度。参与测试的人员对各组语音进行评分，每组语音一般包含两条语音数据，分别为目标说话人的真实语音和由语音转换模型合成的目标说话人的语音，VSS的值从1分到5分，分数越高则代表参与者认为两段语音来自同一个说话人的可能性越大或越有把握认为由同一个说话人说出。

表1 面向非平行语料的语音转换重要工作总结

类型	方法名称	年份	简述
域间迁移	CycleGAN-VC[47]	2018年	使用生成对抗网络，将不同说话人的语音视为来自不同的域进行风格迁移
	StarGAN-VC[48]	2018年	使用单一的生成网络实现多说话人到多说话人的语音转换
基于文本	VTN[61]	2019年	使用预训练的语音合成模型
	DBLSTM-VC[38]	2016年	使用ASR提取PPG作为语音转换的中间特征表示
	Cotatron[58]	2020年	建立在多说话人语音合成模型上并引入文本指导的思想
特征解耦	AutoVC[68]	2019年	使用自编码器和精调的瓶颈解耦内容信息
	AdaIN-VC[71]	2019年	使用实例归一化去除说话人信息并用自适应归一化的方法融入说话人信息
	VQVC[73]	2019年	使用矢量量化的思想提取内容信息
	SpeechSplit[69]	2020年	使用多个编码器并结合瓶颈解耦出内容、韵律、音高以实现更细粒度的语音转换

AB/ABX测试是另一种较为常见的语音转换主观评估指标。在AB测试中，参与测试的人员每次对两段语音样本进行某一方面的判断，如判断哪一段语音更清晰或拥有更高的自然度。在ABX测试中，相较于AB测试多了参考样本X（一般将真实的语音样本作为参考的基准样本），参与者需要在AB样本中选出与参考样本X在某方面（如韵律、音色、音调等）更贴近的一个。

3.1.2 客观评估指标

梅尔倒谱失真（Mel cepstral distortion，MCD）是语音合成和语音转换中常用的客观评估指标。MCD能衡量两段语音的梅尔倒谱系数之间的距离或相似度，值越小则表明合成的语音失真越少，在语音转换的MCD测量中，一般会用到平行语料，测量目标说话人真实话语和含有相同内容的源语音转换到目标说话人的转换语音之间的MCD，可通过Python库提供的工具进行对齐和测量。

其他客观评估指标还有皮尔逊相关系数（Pearson correlation coefficient，PCC），一般用于语音韵律转换等任务，如在韵律转换任务中，使用皮尔逊相关系数测量两段语音中提取的韵律特征之间的关联程度或依赖度，计算方法如下：

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

其中，cov表示协方差，σ表示标准差，X和Y在实际运算中分别代表两段语音的某种特征，如韵律表示。

客观评估指标无法总是与人们对语音做出的评价保持一致、难以准确衡量语音的真实质量，故一般将主观和客观指标结合用于对语音转换模型性能的评估。

3.2 数据集

优质的语音数据集是基于深度学习的语音转换技术的有力保障，语音转换常用的数据集见**表2**。

4 总结与展望

本文概述了语音转换的发展历程，从早期的统计方法的语音转换到现阶段的神经网络语音转换；重点对面向非平行语料的语音转换进行了介绍，将面向非平行语料的神经网络语音转换分为域间迁移语音转换、基于文本的语音转换和基于特征解耦的语音转换3类方法，分别介绍了各方

表2 语音转换常用数据集

数据集名称	时间	基本信息及特点
CMU-Arctic Database[83]	2004年	英语语音，含18名说话人，平行语料
LibriSpeech[84]	2015年	英语语音，含约2 000名说话人，语音数据约1 000小时
CSTR VCTK Corpus[85]	2016年	英语语音，含109名说话人，部分平行语料
VCC 2016[86]	2016年	平行语料，含8名说话人的163组语音
AISHELL-1[87]	2017年	汉语普通话数据集，含400名说话人，语音数据时长约178小时
VCC 2018[88]	2018年	平行语料和非平行语料，各81组
VCC 2020[89]	2020年	平行语料和非平行语料，多语言语音数据，共140组

法中经典的工作并分析了方法的优势与局限，文末介绍了评价语音转换模型的常用指标以及现有工作中广泛使用的数据集。

综合近年来语音转换领域的文献和相关工作，可将未来语音转换工作的方向分为以下几类。

● 提高转换质量。未来的语音转换工作将继续专注于提升模型生成的语音质量。这包括提升语音的清晰度，使生成的语音更易理解和辨识；提高语音的自然度，使生成的语音听起来更加真实和流畅；以及提升与目标说话人音色的相似度。

● 轻量化和实时性。未来的工作也将着手于降低语音转换模型的复杂度、减少计算量，以实现轻量化和实时性。这将使模型能够在资源有限的设备上运行，同时也适用于需要实时语音转换的场景，如通信、智能语音助手等。

● 多样性和灵活性。未来语音转换方面的研究将不仅仅局限于转换说话人的身份，还会关注其他语音属性的转换，包括对语音韵律的灵活转换，使生成的语音更贴近目标风格；对说话情感的转换，使模型能将源语音的说话情感根据指定情感或参考目标语音进行转换。这些多样性的语音转换将会被应用在不同的场景。

● 无监督学习和自监督学习。未来语音转换方面的研究还可能深入探索无监督和自监督学习的方法，以应对低资源场景下的语音转换模型训练。

● 隐私和伦理。随着语音合成和语音转换技术的快速发展，未来的研究将更加强调隐私和道德伦理，也会将语音转换技术更多地应用于安全领域，保护语音数据的隐私安全，同时应对可能的技术滥用和伦理挑战。

总之，未来的语音转换研究将在提升质量、实时性、多样性的同时，也关注安全方面的问题，以满足不断变化和增长的语音处理需求。这将推动语音转换技术走向更加成熟和全面的发展。

参考文献:

[1] YUAN R B, WU Y X, LI J, et al. DeIDVC: speaker de-identification via zero-shot pseudo voice conversion[C]//Proceedings of Interspeech 2022. [S.l.]: ISCA, 2022: 2593-2597.

[2] SRIVASTAVA B M L, VAUQUIER N, SAHIDULLAH M, et al. Evaluating voice conversion-based privacy protection against informed attackers[EB]. arXiv preprint, 2019, arXiv: 1911.03934.

[3] YE Z, MAO T R, DONG L, et al. Fake the real: backdoor attack on deep speech classification via voice conversion[C]// Proceedings of INTERSPEECH 2023. [S.l.]: ISCA, 2023: 4923-4927.

[4] WU Z Z, LI H Z. Voice conversion versus speaker verification: an overview[J]. APSIPA Transactions on Signal and Information Processing, 2014, 3(1): 1-16.

[5] HUANG C Y, LIN Y Y, LEE H Y, et al. Defending your voice: adversarial attack on voice conversion[C]//Proceedings of 2021 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2021: 552-559.

[6] TODA T, NAKAGIRI M, SHIKANO K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(9): 2505-2517.

[7] MA D, VIOLETA L P, KOBAYASHI K, et al. Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion[C]// Proceedings of 2022 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2023: 949-954.

[8] ZHANG M Y, WANG X, FANG F M, et al. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and WaveNet[C]// Proceedings of Interspeech 2019. [S.l.]: ISCA, 2019: 1298-1302.

[9] KAIN A, MACON M W. Spectral voice conversion for text-to-speech synthesis[C]//Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2002: 285-288.

[10] VEAUX C, YAMAGISHI J, KING S. Towards personalised synthesised voices for individuals with vocal disabilities: voice banking and reconstruction[C]// Proceedings of SLPAT 2013. [S.l.:s.n.], 2013: 107-111.

[11] WANG S J, BORTH D. Zero-shot voice conversion via self-supervised prosody representation learning[C]//Proceedings of 2022 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2022: 1-8.

[12] DU Z Y, SISMAN B, ZHOU K, et al. Expressive voice conversion: a joint framework for speaker identity and emotional style transfer[C]//Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE Press, 2022: 594-601.

[13] DU Z Y, SISMAN B, ZHOU K, et al. Disentanglement of emotional style and speaker identity for expressive voice conversion[C]//Proceedings of Interspeech 2022. [S.l.]: ISCA, 2022: 2603-2607.

[14] NAKAMURA K, TODA T, SARUWATARI H, et al. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech[J]. Speech Communication, 2012, 54(1): 134-146.

[15] CHIEN Y L, CHEN H H, YEN M C, et al. Audio-visual mandarin electrolaryngeal speech voice conversion[C]//Proceedings of INTERSPEECH 2023. [S.l.]: ISCA, 2023: 5023-5026.

[16] SISMAN B, YAMAGISHI J, KING S, et al. An overview of voice conversion and its challenges: from statistical modeling to deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 132-157.

[17] HELANDER E, SCHWARZ J, NURMINEN J, et al. On the impact of alignment on voice conversion performance[C]// Proceedings of Interspeech 2008. [S.l.]: ISCA, 2008.

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

[19] ZEN H G, NANKAKU Y, TOKUDA K. Probabilistic feature mapping based on trajectory HMMs[C]//Proceedings of Interspeech 2008. [S.l.]: ISCA, 2008.

[20] KOBAYASHI K, TAKAMICHI S, NAKAMURA S, et al. The NU-NAIST voice conversion system for the voice conversion challenge 2016[C]// Proceedings of Interspeech 2016. [S.l.]: ISCA, 2016: 1667-1671.

[21] 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.

[22] HELANDER E, SILEN H, VIRTANEN T, et al. Voice conversion using dynamic kernel partial least squares regression[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(3): 806-817.

[23] LUAN Y, SAITO D, KASHIWAGI Y, et al. Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition[C]//Proceedings of 2014

IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2014.

[24] NARENDRANATH M, MURTHY H A, RAJENDRAN S, et al. Transformation of formants for voice conversion using artificial neural networks[J]. Speech Communication, 1995, 16(2): 207–216.

[25] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[EB]. arXiv preprint, 2016, arXiv: 1609.03499.

[26] KALCHBRENNER N, ELSEN E, SIMONYAN K, et al. Efficient neural audio synthesis[EB]. arXiv preprint, 2018, arXiv: 1802.08435.

[27] KIM S, LEE S G, SONG J, et al. FloWaveNet: a generative flow for raw audio[EB]. arXiv preprint, 2018, arXiv: 1811.02155.

[28] KONG J, KIM J, BAE J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis[EB]. arXiv preprint, 2020, arXiv: 2010.05646.

[29] KUMAR K, KUMAR R, DE BOISSIERE T, et al. MelGAN: generative adversarial networks for conditional waveform synthesis[EB]. arXiv preprint, 2019, arXiv: 1910.06711.

[30] REN Y, HU C, QIN T, et al. FastSpeech 2: fast and high-quality end-to-end text-to-speech[EB]. arXiv preprint, 2020, arXiv: 2006.04558.

[31] DONAHUE J, DIELEMAN S, BIŃKOWSKI M, et al. End-to-end adversarial text-to-speech[EB]. arXiv preprint, 2020, arXiv: 2006.03575.

[32] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[EB]. arXiv preprint, 2021, arXiv: 2106.06103.

[33] LI J Y, TU W P, XIAO L. Freevc: towards high-quality text-free one-shot voice conversion[C]//Proceedings of ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1–5.

[34] CHOROWSKI J, WEISS R J, BENGIO S, et al. Unsupervised speech representation learning using WaveNet autoencoders[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(12): 2041–2053.

[35] ERRO D, MORENO A, BONAFONTE A. INCA algorithm for training voice conversion systems from nonparallel corpora[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(5): 944–953.

[36] TAO J H, ZHANG M, NURMINEN J, et al. Supervisory data alignment for text-independent voice conversion[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(5): 932–943.

[37] HAZEN T J, SHEN W D, WHITE C. Query-by-example spoken term detection using phonetic posteriorgram templates[C]//Proceedings of 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. Piscataway: IEEE Press, 2010: 421–426.

[38] SUN L F, LI K, WANG H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]//Proceedings of 2016 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2016: 1–6.

[39] SUNDERMANN D, NEY H, HOGE H. VTLN-based cross-language voice conversion[C]//Proceedings of 2003 IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway: IEEE Press, 2004: 676–681.

[40] QIAN Y, XU J, SOONG F K. A frame mapping based HMM approach to

cross-lingual voice transformation[C]// Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2011: 5120-5123.

[41] LIU S X, SUN L F, WU X X, et al. The HCCL-CUHK system for the voice conversion challenge 2018[C]// Proceedings of Speaker and Language Recognition Workshop (Odyssey 2018). [S.l.]: ISCA, 2018: 248-254.

[42] LIU S X, CAO Y W, WANG D S, et al. Any-to-many voice conversion with location-relative sequence-to-sequence modeling[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1717-1728.

[43] ZHOU Y, TIAN X H, YıLMAZ E, et al. A modularized neural network with language-specific output layers for cross-lingual voice conversion[C]// Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE Press, 2020: 160-167.

[44] ZHOU Y, TIAN X H, XU H H, et al. Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling[C]//Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6790-6794.

[45] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5967-5976.

[46] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2242-2251.

[47] KANEKO T, KAMEOKA H. CycleGAN-VC: non-parallel voice conversion using cycle-consistent adversarial networks[C]//Proceedings of 2018 26th European Signal Processing Conference. Piscataway: IEEE Press, 2018: 2100-2104.

[48] KAMEOKA H, KANEKO T, TANAKA K, et al. StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks[C]// Proceedings of 2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2019: 266-273.

[49] KANEKO T, KAMEOKA H, TANAKA K, et al. StarGAN-VC2: rethinking conditional methods for StarGAN-based voice conversion[C]//Proceedings of Interspeech 2019. [S.l.]: ISCA, 2019: 679-683.

[50] LI Y A, ZARE A, MESGARANI N. StarGANv2-VC: a diverse, unsupervised, non-parallel framework for natural-sounding voice conversion[C]// Proceedings of Interspeech 2021. [S.l.]: ISCA, 2021: 1349-1353.

[51] SISMAN B, ZHANG M Y, DONG M H, et al. On the study of generative adversarial networks for cross-lingual voice conversion[C]//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE Press, 2020: 144-151.

[52] YEH C C, HSU P C, CHOU J C, et al. Rhythm-flexible voice conversion without parallel data using cycle-GAN over phoneme posteriorgram sequences[C]// Proceedings of 2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2019: 274-281.

[53] ZHOU K, SISMAN B, LI H Z. Transforming spectrum and prosody for emotional voice conversion with non-

parallel training data[C]//Proceedings of Speaker and Language Recognition Workshop. [S.l.]: ISCA, 2020.

[54] WANG Y X, SKERRY-RYAN R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis[C]// Proceedings of Interspeech 2017. [S.l.]: ISCA, 2017: 4006-4010.

[55] ZHANG J X, LING Z H, LIU L J, et al. Sequence-to-sequence acoustic modeling for voice conversion[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 631-644.

[56] MIYOSHI H, SAITO Y, TAKAMICHI S, et al. Voice conversion using sequence-to-sequence learning of context posterior probabilities[C]//Proceedings of Interspeech 2017. [S.l.]: ISCA, 2017: 1268-1272.

[57] ZHANG M Y, ZHOU Y, ZHAO L, et al. Transfer learning from speech synthesis to voice conversion with non-parallel training data[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1290-1302.

[58] PARK S W, KIM D Y, JOE M C. Cotatron: transcription-guided speech encoder for any-to-many voice conversion without parallel data[EB]. arXiv preprint, 2020, arXiv: 2005.03295.

[59] TIAN X H, CHNG E S, LI H Z. A speaker-dependent WaveNet for voice conversion with non-parallel data[C]// Proceedings of Interspeech 2019. [S.l.]: ISCA, 2019: 15-19.

[60] LIU S, CAO Y, MENG H. Multi-target emotional voice conversion with neural vocoders[EB]. arXiv preprint, 2020, arXiv: 2004.03782.

[61] HUANG W C, HAYASHI T, WU Y C, et al. Voice transformer network: sequence-to-sequence voice conversion using transformer with text-to-speech pretraining[C]//Proceedings of Interspeech 2020. [S.l.]: ISCA, 2020: 4676-4680.

[62] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[EB]. arXiv preprint, 2017, arXiv: 1706.03762.

[63] LUONG H T, YAMAGISHI J. NAUTILUS: a versatile voice cloning system[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2967-2981.

[64] LUONG H T, YAMAGISHI J. Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech[C]//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE Press, 2020: 200-207.

[65] BOESEN A, LARSEN L, SONDERBY S K, et al. Autoencoder beyond pixels using a learned similarity metric[C]//Proceedings of International Conference on Machine Learning. [S.l.:s.n.], 2016.

[66] HSU C C, HWANG H T, WU Y C, et al. Voice conversion from non-parallel corpora using variational auto-encoder[C]//Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE Press, 2017: 1-6.

[67] HUANG W C, HWANG H T, PENG Y H, et al. Voice conversion based on cross-domain features using variational auto encoders[C]//Proceedings of 2018 11th International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE Press, 2019: 51-55.

[68] QIAN K, ZHANG Y, CHANG S, et al. Zero-shot voice style transfer with only autoencoder loss[EB]. arXiv preprint, 2019: arXiv: 1905.05879.

[69] QIAN K, ZHANG Y, CHANG S, et al. Unsupervised speech decomposition via triple information bottleneck[EB]. arXiv preprint, 2020, arXiv: 2004.11284.

[70] HO CHAN C, QIAN K Z, ZHANG Y, et al. SpeechSplit2.0: unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks[C]//Proceedings of ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2022: 6332–6336.

[71] CHOU J C, LEE H Y. One–shot voice conversion by separating speaker and content representations with instance normalization[C]//Proceedings of Interspeech 2019. [S.l.]: ISCA, 2019: 664–668.

[72] 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]//Proceedings of ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 5954–5958.

[73] WU D Y, LEE H Y. One–shot voice conversion by vector quantization[C]//Proceedings of ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2020: 7734–7738.

[74] WU D Y, CHEN Y H, LEE H Y. VQVC+: one–shot voice conversion by vector quantization and U–net architecture[C]//Proceedings of Interspeech 2020. [S.l.]: ISCA, 2020: 4691–4695.

[75] WANG D S, DENG L Q, YEUNG Y T, et al. VQMIVC: vector quantization and mutual information–based unsupervised speech representation disentanglement for one–shot voice conversion[C]//Proceedings of Interspeech 2021. [S.l.]: ISCA, 2021: 1344–1348.

[76] LIU Z H, WANG S J, CHEN N. Automatic speech disentanglement for voice conversion using rank module and speech augmentation[C]//Proceedings of Interspeech 2023. [S.l.]: ISCA, 2023.

[77] YANG S C, TANTRAWENITH M, ZHUANG H L, et al. Speech representation disentanglement with adversarial mutual information learning for one–shot voice conversion[C]//Proceedings of Interspeech 2022. [S.l.]: ISCA, 2022: 2553–2557.

[78] KANEKO T, KAMEOKA H, HIRAMATSU K, et al. Sequence–to–sequence voice conversion with similarity metric learned using generative adversarial networks[C]//Proceedings of Interspeech 2017. [S.l.]: ISCA, 2017: 1283–1287.

[79] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[EB]. arXiv preprint, 2020, arXiv: 2006.11239.

[80] KONG Z, PING W, HUANG J, et al. DiffWave: a versatile diffusion model for audio synthesis[EB]. arXiv preprint, 2020, arXiv: 2009.09761.

[81] HUANG R, LAM M W Y , WANG J, et al. FastDiff: a fast conditional diffusion model for high–quality speech synthesis[EB]. arXiv preprint, 2022, arXiv: 2204.09934.

[82] LIU S X, CAO Y W, SU D, et al. DiffSVC: a diffusion probabilistic model for singing voice conversion[C]//Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop. Piscataway: IEEE Press, 2022: 741–748.

[83] KOMINEK J, BLACK A. The CMU Arctic speech databases[EB]. ResearchGate, 2004: 228978129.

[84] PANAYOTOV V, CHEN G G, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 5206–5210.

[85] VEAUX C, YAMAGISHI J, MACDONALD K. CSTR VCTK Corpus: English multi–speaker Corpus for CSTR voice cloning

toolkit[Z]. 2016.
[86] TODA T, CHEN L H, SAITO D, et al. The voice conversion challenge 2016[C]// Proceedings of Interspeech 2016. [S.l.:s.n.], 2016: 1632-1636.
[87] BU H, DU J, NA X, et al. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline[EB]. arXiv preprint, 2017, arXiv: 1709.05522.
[88] LORENZO-TRUEBA J, YAMAGISHI J, TODA T, et al. The voice conversion challenge 2018: promoting development of parallel and nonparallel methods[EB]. arXiv preprint, 2018, arXiv: 1804.04262.
[89] ZHAO Y, HUANG W C, TIAN X, et al. Voice Conversion Challenge 2020: intra-lingual semi-parallel and cross-lingual voice conversion[EB]. arXiv preprint, 2020, arXiv: 2008.12527.

作者简介

李鹏程（1999- ），男，中国科学技术大学硕士生，平安科技（深圳）有限公司算法工程师，主要研究方向为语音合成、语音转换和语音安全等。

张旭龙（1988- ），男，博士，平安科技（深圳）有限公司高级算法研究员，复旦大学计算机理学博士，主要研究方向为语音合成、语音转换、音乐信息检索以及机器学习和深度学习方法在人工智能领域应用。担任清华大学深圳研究院以及中国科学技术大学先进技术研究院校外导师，目前是IEEE、中国自动化学会以及中国计算机学会会员，担任联邦数据与联邦智能专委会委员，2023年入选上海市东方英才计划青年项目。

王健宗（1983- ），男，博士，平安科技（深圳）有限公司副总工程师，资深人工智能总监，联邦学习技术部总经理，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会资深会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为大模型、联邦学习和深度学习等。

程宁（1981- ），男，博士，平安科技（深圳）有限公司高级人工智能专家、中国科学院自动化所博士。专注于人工智能算法研究以及其在语音处理和自然语言处理领域的应用。目前在大数据、机器学习、人工智能国际顶会或期刊上发表学术论文50余篇，发明专利申请100余项。

肖京（1972– ），男，博士，美国卡耐基梅隆大学博士，国家特聘专家。国家新一代普惠金融人工智能开放创新平台技术负责人、深圳市政协委员、深圳市决策咨询委员会委员，兼中国计算机学会深圳分部副主席、广东省人工智能与机器人学会副理事长、深圳市人工智能行业协会会长、深圳市人工智能学会副理事长，清华大学、上海交通大学、同济大学等客座教授。长期从事人工智能与大数据分析挖掘相关领域的研究，先后在爱普生美国研究院及美国微软公司担任高级研发管理职务，现任平安集团首席科学家，负责人工智能技术研发及在金融、医疗、智慧城市等领域的应用，带领团队树立了多项传统行业智能化经营的标杆。发表学术论文249篇，美国授权专利101项，中国发明专利155项，参与及承担国家级项目8项。凭借在技术创新及应用方面的杰出贡献，先后获得2018年中国专利奖、2019年吴文俊人工智能杰出贡献奖、2020年吴文俊人工智能科技进步奖一等奖、2020年上海市科技进步奖一等奖、2020年中国人工智能十大风云人物、2021年深圳市五一劳动奖章、2022年深圳市最美科技工作者等荣誉。

收稿日期： 2023-09-26

通信作者： 王健宗，jzwang@188.com

基金项目： 广东省重点领域研发计划“新一代人工智能”重大专项（No.2021B0101400003）

Foundation Item: The Key Research and Development Program of Guangdong Province (No.2021B0101400003)