CycleGAN-VC3：检查和改进用于梅尔频谱图转换的CycleGAN-VC

Abstract

非并行语音转换（VC）是一种无需使用并行语料库即可学习源语音和目标语音之间的映射的技术。 最近，周期一致的对抗网络（CycleGAN）-VC和CycleGAN-VC2在此问题上显示出了令人鼓舞的结果，并已被广泛用作基准方法。 但是，由于CycleGAN-VC / VC2对mel谱图转换的有效性不明确，即使比较方法采用mel谱图作为转换目标，它们也通常用于mel-cepstrum转换。 为了解决这个问题，我们研究了CycleGAN-VC / VC2对mel谱图转换的适用性。

通过初步实验，我们发现它们的直接应用损害了在转换过程中应保留的时频结构。 为了解决这个问题，我们提出了CycleGAN-VC3，它是CycleGANVC2的改进，它结合了时频自适应归一化（TFAN）。 使用TFAN，我们可以调整转换特征的比例和偏差，同时反映源Mel频谱图的时频结构。 我们在性别间和性别内非平行VC上评估了CycleGAN-VC3。 对自然性和相似性的主观评估表明，对于每个VC对，CycleGAN-VC3的性能均优于或优于两种类型的CycleGAN-VC2，其中一种适用于mel倒谱，另一种适用于mel谱图。【1】

Index Terms: voice conversion (VC), non-parallel VC, generative adversarial networks (GANs), CycleGAN-VC, melspectrogram conversion

1. Introduction

语音转换（VC）是一种用于在保留语言信息的同时转换语音中非/副语言信息的技术。 由于VC在语音助手[1、2]，语音增强[3、4]和口音转换[5]等方面的潜在多样化应用，因此已经进行了积极的研究。 机器学习方法被广泛使用，包括基于高斯混合模型[6，7]和神经网络（NN）的统计方法，包括前馈NN [8]，递归NN [9]，卷积NN（CNN）[5]， 注意网络[10，11]和生成对抗网络（GANs）[5]。

为了便于学习，许多VC方法（包括上述方法）使用并行语料库学习从源语音到目标语音的映射。 但是，收集此类数据并不一定容易或实用。 此外，即使收集了此类数据，大多数VC方法所需的时间对齐过程仍然是一个挑战。

作为替代，不需要并行语料库进行训练的非并行VC最近受到关注。就数据收集成本而言，非并行VC具有吸引力。 然而，由于缺乏明确的监督，其学习具有挑战性。 为了解决这个问题，一些研究已经利用语言信息[12、13、14、15]。 尽管这种额外的监督可以提高性能，但是需要辅助数据或模块来提取语言信息。

为了避免这种要求并仅使用声学数据来实现非并行VC，已经提出了基于变分自动编码器的方法[16、17、18]和基于GAN的方法[17、19]。 其中，CycleGAN-VC [19，20]及其变体（CycleGAN-VC2 [21]和StarGAN-VCs [22，23，24]）已经引起了人们的关注，并且在一些研究中已被广泛用作基准方法（ 例如[15、25、26]）。

但是，由于它们对于蜜谱图转换的有效性含糊不清，即使比较方法采用蜜谱图作为转换目标时，它们也通常用于半谱峰转换（例如[15，26]）。

这些事实促使我们研究CycleGAN-VC [19]和CycleGAN-VC2 [21]在质谱图转换中的适用性。 通过最初的实验，我们发现将CycleGAN-VC / VC2直接应用于质谱图时，会损害转换期间应保留的时频结构，如图1所示。

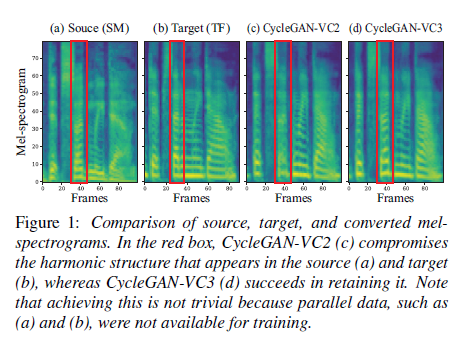


Figure 1: Comparison of source, target, and converted melspectrograms. In the red box, CycleGAN-VC2 (c) compromises the harmonic structure that appears in the source (a) and target (b), whereas CycleGAN-VC3 (d) succeeds in retaining it. Note that achieving this is not trivial because parallel data, such as (a) and (b), were not available for training.

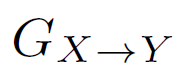
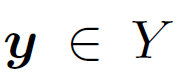
为了解决这个问题，我们提出了CycleGAN-VC3，它是对CycleGAN-VC2的改进，它结合了时频自适应归一化（TFAN）。 TFAN受到空间自适应（反）归一化（SPADE）[27]的启发，该方法最初提出用于语义图像合成。 我们修改了SPADE，以将其应用于1D和2D时频特征。 使用TFAN，我们可以调整转换特征的比例和偏差，同时反映源Mel频谱图的时频结构。

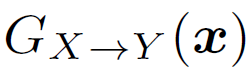
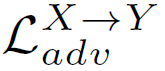
我们使用语音转换挑战2018（VCC 2018）数据集[28]，研究了CycleGAN-VC3在性别间和性别内非并行VC上的有效性。 对自然性和相似性的主观评估表明，对于arXiv：2010.11672v1 [cs.SD] 2020年10月22日，每个VC对，CycleGAN-VC3的性能均优于两种类型的CycleGANVC2（其中一种用于mel） -倒谱，另一到蜜谱图。

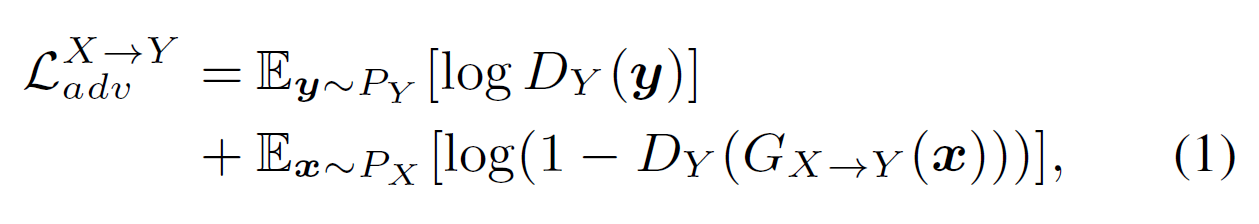
本文的其余部分安排如下。 在第2节中，我们简要回顾了传统的CycleGAN-VC / VC2。 在第3节中，我们介绍了拟议的CycleGAN-VC3。 在第4节中，我们描述了实验结果。 第5节简要概述和描述了未来的工作。

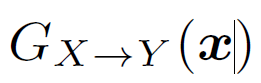
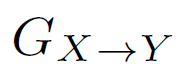
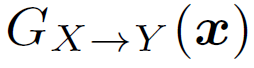
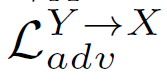
2. Conventional CycleGAN-VC/VC2

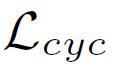
2.1. Training objectives

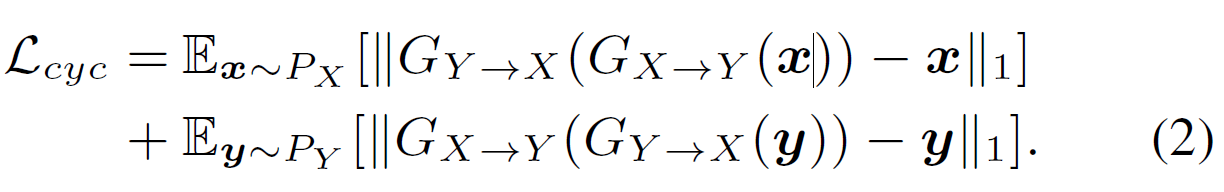
CycleGAN-VC / VC2旨在学习映射，该映射将源声学特征转换为目标声学特征，而无需使用并行语料库。 受最初提出用于不成对图像到图像转换的CycleGAN [29]的启发，CycleGAN-VC / VC2使用对抗性损失[30]，循环一致性损失[31]和身份映射损失[32]来学习映射。 此外，CycleGAN-VC2使用第二次对抗性损失来改善重构特征的细节。

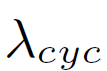
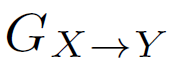
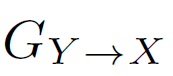
对抗损失 Adversarial loss。 为了确保转换后的特征在目标Y中，按如下方式使用对抗损失：



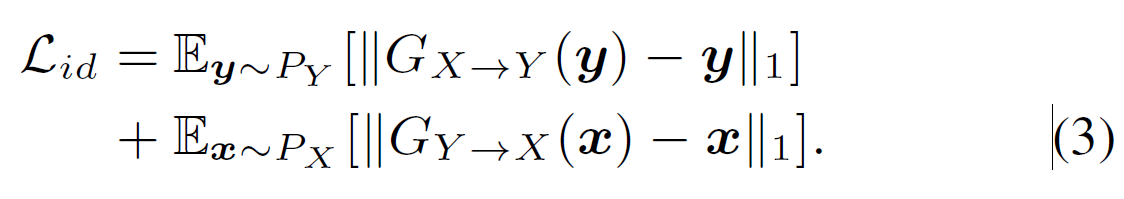
其中判别器试图通过使损耗最大来将合成的与实数y区分开，而试图通过最小化损耗来合成可以欺骗的。 类似地，逆映射和鉴别器使用进行对抗训练。

Cycle-consistency loss 周期一致性损失。 为了保持转换中的成分，循环一致性损失的用法如下：

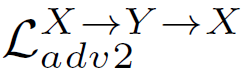


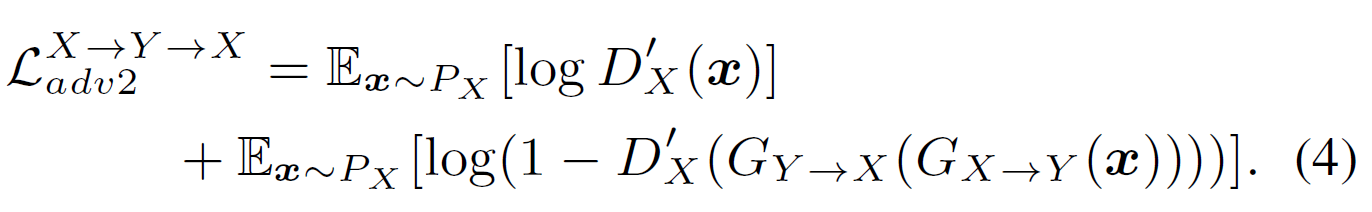
此损失与控制其相对重要性的超参数一起使用。 损失有助于和识别周期一致性约束内的伪对。

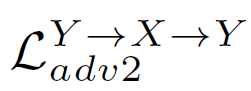
Identity-mapping loss 身份映射丢失。 为了便于保存输入，使用了一个标识映射丢失，如下所示：



此损失与控制其相对重要性的超参数一起使用。

Second adversarial loss 第二次对抗损失。 在CycleGAN-VC2中，为了减轻由L1损失引起的统计平均（公式2），引入了一个附加的鉴别符，并对循环转换后的特征施加了第二个对抗损失，如下所示：



类似地，引入鉴别符，并将施加于逆向映射。

2.2. Generator architectures

CycleGAN-VC使用一维CNN生成器[5]来捕获整体关系以及特征方向，同时保留时间结构。 特别地，网络由下采样，残差[33]和上采样块组成，以有效地捕获宽范围的时间关系，并且门控线性单元（GLU）[34]被用作激活来自适应地学习顺序和层次结构。

但是，对CycleGAN-VC2的研究[21]表明，降采样和升采样块中的一维CNN会影响应保留在转换中的结构。 为了缓解这种情况，CycleGAN-VC2引入了2-1-2D CNN，该模型在上采样和下采样块中使用2D CNN，而在剩余块中使用1D CNN。 前者用于在保留原始结构的同时提取时频结构。 后者用于执行动态更改。

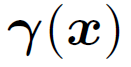
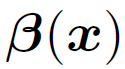
2.3. Discriminator architectures

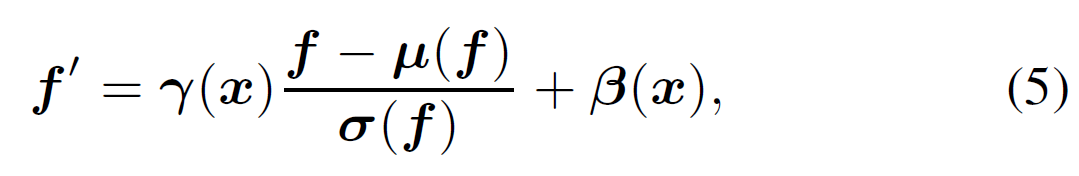
CycleGAN-VC使用2D CNN鉴别器[35]来基于2D频谱纹理鉴别数据。 特别是，它使用具有完全连接层作为最后一层的FullGAN，以根据整体输入结构来区分数据。 但是，在FullGAN中，需要学习许多参数会导致学习困难。 为了减轻这种情况，CycleGAN-VC2引入了在最后一层使用卷积的PatchGAN [36]。 这减少了参数并稳定了GAN训练。

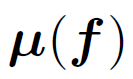
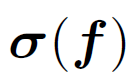
3. CycleGAN-VC3

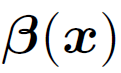
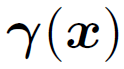
3.1. TFAN: Time-frequency adaptive normalization

CycleGAN-VC和CycleGAN-VC2最初是为梅尔-倒谱转换而设计的，而它们在梅谱图转换中的有效性还没有得到充分检验。 我们通过经验检查了它们的有效性，发现它们损害了应保留在转换中的时频结构，如图1所示。

出于这一发现，我们设计了TFAN来扩展实例归一化（IN）[37]，以调整转换特征的比例和偏差，同时以时间和频率方式反映源信息（即x）。 尤其是，我们针对1D和2D时频特征设计了TFAN，以便在2-1-2D CNN中使用（第2.2节）。 图2说明了TFAN的体系结构。 给定特征f，TFAN以类似于IN的通道方式对其进行归一化，然后使用标度和偏差以元素方式对归一化的特征进行调制，这是使用CNN从x计算得出的：



其中是输出特征，和分别是f的通道平均和标准偏差。

在IN中，以通道方式应用与x无关的比例和偏差，而在TFAN中，以元素方式应用根据x（即和）计算的比例和偏差。这些差异允许TFAN调整f的比例和偏差，同时以时间和频率方式反映x。

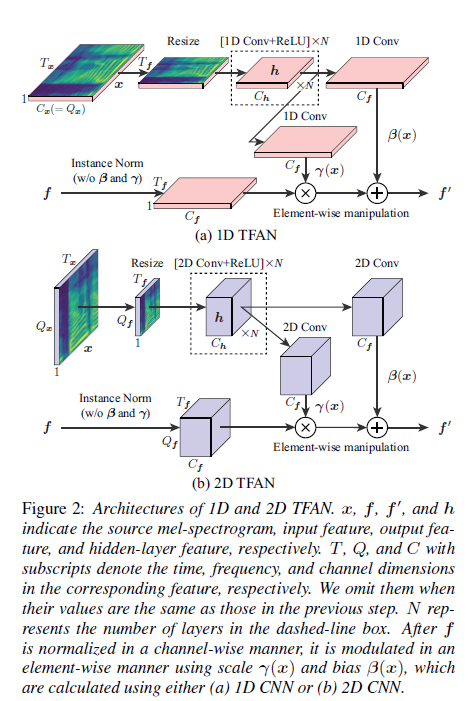
请注意，TFAN受SPADE [27]的启发，该技术最初是为语义图像合成而提出的。

主要区别在于

（1）SPADE是为2D图像特征而设计的，而TFAN是为1D和2D时频特征而设计的；

（2）SPADE在图2中的虚线框所示的组件中使用了一层CNN，因为语义图像合成中不需要进行大的更改，而TFAN使用多层CNN来确保动态更改，并且

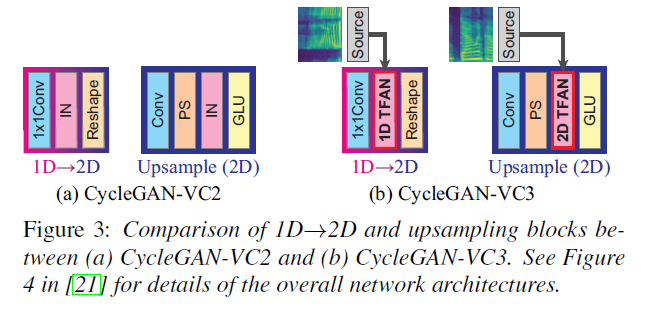
（3）SPADE基于批次归一化[38]，而TFAN基于IN。我们检查（1）和（2）的影响，并在4.2节中介绍我们的发现。



3.2. Implementation

在CycleGAN-VC3中，我们将TFAN合并到CycleGAN-VC2发生器中（即2-1-2D CNN（第2.2节））。

特别是，将块中的IN和上采样块中的IN分别替换为1D TFAN和2D TFAN，如图3所示。在TFAN中，我们设置了通道数（即）和内核大小。 h（图2）分别为128和5。 我们通过改变N并调整TFAN的插入位置来检查性能，并在4.2节中介绍我们的发现。 鉴别器与CycleGAN-VC2中使用的鉴别器相同（即PatchGAN（第2.3节））。



4. Experiments

数据集。 我们在VCC 2018 [28]的Spoke（即非并行VC）任务上评估了CycleGAN-VC，其中包含美国英语专业人士的录音。 我们选择了考虑所有性别间和性别内部VC的发言人子集：VCC2SF3（SF），VCC2SM3（SM），VCC2TF1（TF）和VCC2TM1（TM），其中S，T，F和M代表来源 ，目标，女性和男性。 使用2个来源和2个目标的组合进行评估。 对于每个说话者，分别使用81 utterances（大约5分钟，对于VC来说相对较低）和35 utterances进行训练和评估。 在训练集中，源话语和目标话语之间没有重叠； 因此，必须在完全不平行的情况下解决此问题。 记录被下采样到22.05 kHz。 在研究MelGAN [39]（我们在实验中将其用作声码器）之后，我们提取了一个80维对数梅尔声谱图，其窗口长度为1024，跃点长度为256个样本。

转换过程。 这项研究的目的是研究使用CycleGAN-VC进行蜜谱图转换的可行性。

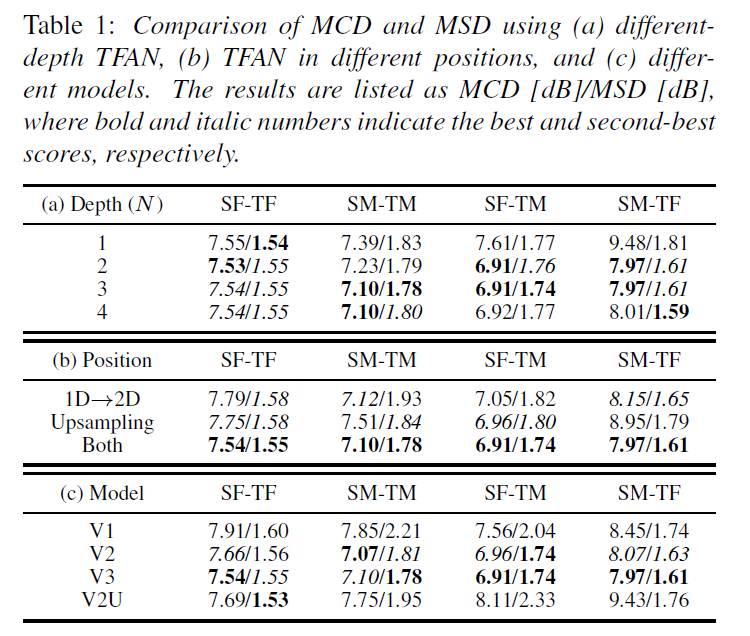
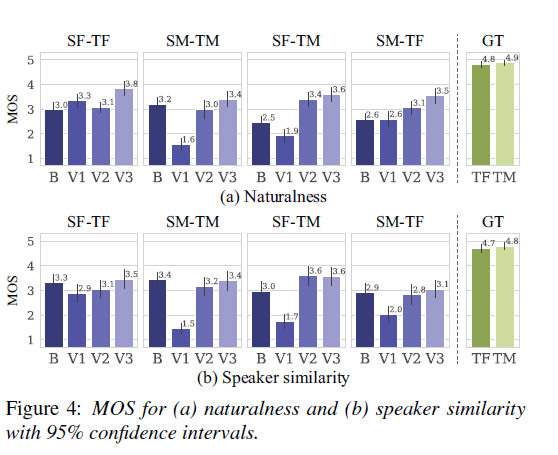
因此，我们将CycleGAN-VC用于Mel谱图转换，并使用预训练的MelGAN声码器合成波形[39]。2我们没有更改声码器的参数，因此我们可以专注于Melspectrogram转换的评估。 但是，对每个扬声器进行微调是一种可能的改进方法。

网络体系结构。 随着声学特征从mel倒频谱变为mel频谱图，特征维从35增大到80。然而，CycleGAN-VC的生成器是完全卷积的；它的生成是完全卷积的。 因此，可以在不修改网络体系结构的情况下使用它们。 关于鉴别器，我们使用与mel-cepstrum转换相同的网络架构，不同的是在CycleGANVC2 / VC3中，倒数第二个卷积层的内核大小在频率方向上增加了一倍（请参见[21]中的图4）。 原始网络架构的详细信息）。

训练设置。 训练设置类似于在CycleGAN-VC / VC2中用于mel-cepstrum转换的设置[19，21]。 对于预处理，我们使用训练数据的均值和方差对梅尔频谱图进行归一化。 我们使用最小二乘GAN [40]作为GAN目标。 我们使用Adam优化器[41]对网络进行了500k迭代训练，批处理大小为1。训练样本由随机裁剪的64帧（约0.75 s）组成。 生成器的学习率设置为0.0002，鉴别器的学习率设置为0.0001，动量项和分别为0.5和0.999。和分别设置为10和5，并且仅用于前10k迭代。 请注意，类似于原始的CycleGAN-VC / VC2，我们没有使用额外的数据，模块或时间对齐过程进行培训。

4.2. Objective evaluation

我们进行了客观评估，以研究TFAN参数选择的效果以及CycleGAN-VC之间的性能差异。 直接测量目标光谱图和转换的光谱图之间的差异非常困难，因为它们的对齐方式并不容易。 作为替代方案，我们使用了以前的VC [21、23、25]中常用的两个评估指标：梅尔谱倒谱失真（MCD），它根据目标和转换后的梅尔谱倒谱测量全局结构差异，以及 调制谱距离（MSD），它基于mel-cepstra的目标和转换后的调制谱评估局部结构差异。 对于这两个指标，值越小，性能越好。 使用WORLD [42]从目标或转换后的波形中提取了35维梅尔倒谱参数。

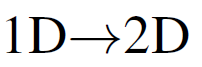
TFAN参数选择的效果。 我们最初检查了TFAN参数选择的效果。 特别是，我们研究了在改变TFAN的深度（图2中的N）和插入TFAN的位置（和/或上采样模块（见图3））下的性能。 表1（a）和（b）列出了各自的结果。 我们的主要发现如下。

（1）不同深度TFAN的比较（表1（a））。

我们发现

（i）1深度TFAN表现最差，但SF-TF除外，后者的表现与其他表现相当，并且

（ii）分数在大约N = 3时达到峰值。这表明使用多层CNN执行动态更改的重要性。 这不同于应用SPADE [27]的语义图像合成。

（2）TFAN位置的比较（表1（b））。 我们发现，在和升采样模块中联合使用1D TFAN和2D TFAN最为有效。 因此，在以下实验中，我们将N设置为3，并在两个位置都使用了TFAN。

CycleGAN-VC之间的比较。 我们分析了CycleGAN-VC之间的性能差异。 除了CycleGAN-VC（V1），-VC2（V2）和-VC3（V3）外，我们还通过U-net [43]（V2U）检查了CycleGAN-VC2，这可能是TFAN传播的替代选择 将源信息转换为要素。 表1（c）总结了结果。 V2和V3通常得分最高或第二。 在这两个指标中，V3在大多数情况下都表现出更好的性能。 此外，我们分别在图1和第4.3节中显示了定性和主观比较。

4.3. Subjective evaluation

我们进行了听力测试，以评估CycleGAN-VC在蜜谱图转换中的适用性。 我们比较了四种模型：具有mel-倒谱转换的CycleGAN-VC2（当前最佳和基准模型；用B表示）和具有mel-频谱图转换的CycleGAN-VC，-VC2和-VC3（由V1，V2和V3表示） ， 分别）。为了衡量自然性和说话人的相似性，我们进行了平均意见得分（MOS）测试3，其中包括目标地面真实语音（GT）作为锚点样本。 在两个测试中，所有评估数据（35个发音）均用于评估，每个发音均评估了一次。 在说话人相似性测试中，我们以随机顺序将转换后的语音和目标语音（话语内容不同）配对。 分别有9位和11位听众参加了自然性和说话人相似性测试。 音频样本可在线获得。1

图4显示了结果。 我们的主要发现如下。 （1）CycleGAN-VC2用于半倒谱（B）和mel频谱图（V2）的比较。 就自然性而言，V2对于跨性别的VC（SF-TM和SM-TF）特别有效，对于说话人的相似性，结果取决于情况。 这表明直接应用于蜜谱图（即V2）不一定是合理的。

（2）CycleGAN-VC（V1，V2和V3）之间的比较。 在大多数情况下，随着版本的增加，两种指标的性能都会提高。 与以前的版本相比，这确认了V2和V3的实用性。 （3）CycleGAN-VC3（V3）的有效性。 与其他型号相比，V3具有更好或更具竞争力的性能。这表明有可能将V3用作代替B的新基准方法。

5. Conclusions

尽管CycleGAN-VC被广泛用作基准方法，但仍未充分检验其用于mel谱图转换的可行性。 因此，在检查之后，我们提出了CycleGAN-VC3，这是对包含TFAN的CycleGANVC2的改进。 使用此模块，我们可以转换mel频谱图，同时自适应地反映源mel频谱图。 实验结果表明，使用CycleGAN-VC3作为替代BenchGAN-VC2的新基准方法的潜力。 将TFAN应用于高级任务，例如多域VC [22，23，24]和应用程序VC [1,2,3,4,5]，仍然是一个有趣的未来方向。

6. Acknowledgements

这项工作得到了日本JSPS KAKENHI 17H01763和JST CREST授权号JPMJCR19A3的支持。