ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 6,Nov. 2023,pp. 1458－1468
DOI:10. 16337/j. 1004-9037. 2023. 06. 019

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

多尺度富有表现力的汉语语音合成

高　洁[1]，肖大军[2]，徐遐龄[2]，刘绍翰[1]，杨　群[1]

(1. 南京航空航天大学计算机科学与技术学院,南京 211106;2. 国家电网公司华中分部,武汉 430070)

摘　要：常见的增强合成语音表现力方法通常是将参考音频编码为固定维度的韵律嵌入,与文本信息一起输入语音合成模型的解码器,从而向语音合成模型中引入变化的韵律信息,但这种方法仅提取了音频整体级别的韵律信息,忽略了字或音素级别的细粒度韵律信息,导致合成语音依然存在部分字词发音不自然、音调语速平缓的现象。针对这些问题,本文提出一种基于Tacotron2语音合成模型的多尺度富有表现力的汉语语音合成方法。该方法利用基于变分自编码器的多尺度韵律编码网络,提取参考音频整体级别的韵律信息和音素级别的音高信息,然后将其与文本信息一起输入语音合成模型的解码器。此外,在训练过程中通过最小化韵律嵌入与音高嵌入之间的互信息,消除不同特征表示之间的相互关联,分离不同特征表示。实验结果表明,该方法与单一尺度的增强表现力语音合成方法相比,听力主观平均意见得分提高了约2%,基频F_0帧错误率降低了约14%,该方法可以生成更加自然且富有表现力的语音。

关键词：语音合成;神经网络;变分自动编码器;注意力机制;韵律增强

中图分类号：TP391　　**文献标志码**：A

Multi-scale Expressive Chinese Speech Synthesis

GAO Jie[1], XIAO Dajun[2], XU Xialing[2], LIU Shaohan[1], YANG Qun[1]

(1. College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China; 2. Central China Branch of State Grid Corporation of China, Wuhan 430070, China)

Abstract: Common methods for enhancing the expressiveness of synthesized speech typically involve encoding the reference audio as a fixed-dimensional prosody embedding. This embedding is then fed into the decoder of the speech synthesis model along with the text embedding, thereby introducing prosody information into the speech synthesis process. However, this approach only captures prosody information at the global level of speech, neglecting fine-grained prosody details at the word or phoneme level. Consequently, the synthesized speech may still exhibit unnatural pronunciation and flat intonation in certain words. To tackle these issues, this paper introduces a multi-scale expressive Chinese speech synthesis method based on Tacontron2. Initially, two variational auto-encoders are employed to extract global-level prosody information and phoneme-level pitch information from the reference audio. This multi-scale variational information is then incorporated into the speech synthesis model. Additionally, during the training process, we minimize the mutual information between the rhyme embedding and the pitch embedding. This step aims to eliminate intercorrelation between different feature representations and to

收稿日期:2023-01-13;**修订日期**:2023-06-28

separate distinct feature representations. Experimental results demonstrate that our proposed method enhances the subjective mean opinion score by 2% and reduces the F_0 frame error rate by 14% compared to the single-scale expressive speech synthesis method. The findings suggest that our method generates speech that is more natural and expressive.

Key words: speech synthesis; neural networks; variational auto-encoder; attention mechanism; prosody enhancement

引　言

语音合成,又称文语转换(Text-to-speech,TTS)技术,是指通过计算机将文本转化为语音。基于神经网络的语音合成模型,例如:Tacotron[1]、MelNet[2]、Deep Voice 3[3]和TransformerTTS[4]已经能够根据输入文本合成较高质量的语音音频。这些模型经过训练,将输入文本映射到语音特征(例如梅尔频谱)。在现实生活中,文本与语音是一对多的映射关系。具有相同文本内容的真实人类语音并不是一模一样的,它会随着说话者的韵律特征,如时长、音高、音量等的变化而发生改变。仅从文本信息生成语音,缺少与语音韵律相关信息,在这种没有足够的输入信息的情况下训练模型,学习文本与语音一对多的映射,所生成语音的梅尔频谱往往过于平滑,模型倾向于学习数据集的平均韵律模式而不是学习每一个语句具体的韵律变化。这导致合成的语音音调平缓、缺乏节奏感和表现力、语音自然度较差,与真人语音有一定的差距。增强合成语音的表现力的一个关键是处理文本信息与语音特征之间一对多的映射关系。将变化的韵律信息作为输入引入语音合成模型,并对这些信息进行建模可以缓解以上问题,提高合成语音的表现力[5]。近年来,研究人员通过将参考音频作为语音合成模型的输入之一,把韵律信息引入语音合成模型中。在这些方法中,通常是通过参考编码器将参考音频编码为韵律嵌入,再将韵律嵌入与文本嵌入一起输入到解码器进行解码。例如Wang等[6]提出的全局样式标记(Global style token,GST),这种方法通过参考编码器和一个基于多头注意力的GST层提取参考音频的句子级别的韵律信息。当GST进行训练时,它会生成多个标记,这些标记的加权和作为音频的韵律嵌入。另一类增强合成语音表现力的方法通过变分自动编码器(Variational auto-encoder,VAE)[7]实现。VAE通过从潜在变量的分布中采样,从而生成具有特定特征的样本。潜在变量连续并且可以插值,类似于语音中的隐式特征。Zhang等[8]在语音合成模型中添加了一个VAE网络来学习代表语音韵律信息的潜在变量。

上述方法在学习参考音频韵律表示方面表现出良好的性能,并在一定程度上提高了合成语音的表现力。但是,它们都是将韵律信息编码为固定维度,仅关注了句子级别这一尺度的韵律信息。事实上,人类语音的韵律表达本质上是多尺度的,从粗粒度到细粒度都应有体现,而不仅在单尺度上。在句子的整体级别,可以对语句的韵律模式进行整体概括;而在语音音频的局部中,语句中每个音素的韵律特征都在发生变化。比如音调特征往往就在音素间发生变化。仅对句子级别的韵律信息进行建模,就会忽略更细粒度的如音素级别的变化信息。汉语作为一种音调语言系统,有着音调多变及音调载义的特点。汉语中有众多的同音字,通过不同的音调用来区分词义。音调可以帮助听者理解语音所表达的含义[9]。而汉语音调往往在音或字之间发生变化,仅对韵律信息在句子级别进行建模,会导致细粒度的音素级别的语调变化信息丢失,在合成汉语句子时,生成的语音仍然存在停顿不当、字词发音不自然甚至错误等问题。

针对上文对于仅在单一尺度学习音频韵律信息方法不足的分析,本文在Tacotron2模型[10]的基础上提出了一种多尺度的汉语语音合成方法,旨在学习多尺度的韵律信息。Tacotron2是Tacotron的改进

版本，是目前最先进的语音合成模型之一，本文所提方法对Tacotron2进行了拓展。考虑到汉语音调具体表现为音高随时间而变化的模式[11]，而语音信号的基频F_0特征直接关系到语音的音高，所以本文设计将基频引入语音合成模型，通过模型学习音频的音高信息。为了得到合适的信息表示以及受到VAE网络在语音合成领域应用的启发，本文提出了基于VAE网络的多尺度韵律编码网络，对参考音频的韵律信息进行多尺度建模。具体来说，在多尺度韵律编码网络中，通过一个VAE对参考音频的韵律信息在句子级别进行建模，将参考语音中的韵律信息编码为一个固定维度的全局韵律嵌入，该VAE网络称作韵律编码器；而多尺度韵律编码网络中的另一个VAE网络对参考音频音高特征在音素级别进行建模，将参考语音中的音高信息编码为一个的音素级别的韵律嵌入序列，它被称作音高编码器。通过两个VAE学习整体和局部的多尺度的韵律变化，帮助模型进行文本到语音特征的一对多映射的建模，增强合成语音的表现力。此外，考虑到VAE作为一种无监督学习的模型，它学习到的韵律表示为所有韵律特征的混合表示。各个韵律特征彼此纠缠在一起，对一个特征进行操作可能会影响其他维度，模型无法对韵律特征进行更明确精准的控制。并且训练时，因为输入的参考音频与输出的生成音频相同，所以两个编码器会出现信息冗余的情况，解码器会使用一个编码网络学习到的信息去重建语音，同时忽略另一个编码器学习的信息，在推理时影响合成语音的质量。所以，本文设计在训练过程中通过最小化韵律嵌入和音高嵌入之间的互信息，进一步对音高表示和韵律表示进行分离，增强不同特征之间的独立性，减少编码器之间出现信息冗余的现象，使模型对于合成语音的控制更加灵活。

1 Tacotron2模型结构

本文所提方法基于Tacotron2[10]语音合成模型。它是一种先进的端到端语音合成模型，由谷歌公司在2018年提出。Tacotron2是Tacotron模型的升级版，它优化了模型的编码器和解码器结构，并采用更高效的注意力机制，从而显著提升了合成效果。Tacotron2通过循环的序列到序列特征预测模型，将文本内容直接映射到梅尔频谱，实现直接从文本进行语音合成。这种端到端的思路避免了多个分阶段的处理，提高了合成效率。同时，Tacotron2的注意力机制能够有效地关注文本与音频之间的对应关系，使得合成语音的音质更加自然流畅。

如图1所示，Tacotron2由两个部分组成：(1)带有注意力机制的循环序列到序列特征预测网络。它从输入的字符序列中预测梅尔谱频谱序列。(2)声码器。它通过预测的梅尔频谱，生成时域波形。具体来说，序列到序列特征预测模型是由编码器和带有注意力机制的解码器两部分组成。编码器将字符序列转换为相应的隐藏特征表示，而解码器通过编码器生成的特征表示来预测频谱图。编码器由3层卷积神经网络(Convolutional neural networks, CNN)和一层双向长短期记忆网络(Long short-term memory, LSTM)组成。当文本序列输入后，编码器会输出相应的编码序列作为文本嵌入。文本嵌入序列通过注意力网络生成固定长度的内容向量，用于后续解码器进行特征预测。Tacotron2中注意力网络使用带有位置敏感的注意力网络，它使解码器只能单向进行解码，减少了一些潜在的解码时会出现的问题。解码器部分是一个自回归循环神经网络，它能够利用上一帧信息来预测下一帧信息。解码器每次从输入的编码预测一帧梅尔谱图，并且逐帧地预测停止符。上一个时间步的预测结果通过一个由两个全连接层构成的预处理网络和

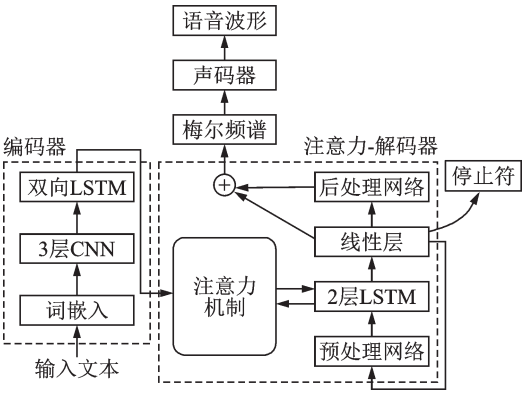

图1 Tacotron2架构

Fig.1 System architecture of Tacotron2

两个单向LSTM与内容向量一起预测目标梅尔频谱帧,以及是否为最后一帧。预测的结果通过由5层卷积网络构成的后处理网络对其进行改善。

Tacotron2的损失函数主要由两部分组成:基于梅尔频谱图的损失和基于停止标志的损失。基于谱图的损失是用来衡量模型生成的声音与目标声音之间的差异,而基于停止标志的损失则是用来衡量模型是否在正确的时间停止生成声音。

2　多尺度富有表现力的语音合成模型

本文针对合成语音的语调单一、韵律不够丰富、表现力有限的问题,在Tacotron2的基础上引入了基于VAE的多尺度韵律编码网络和互信息估计器,使其能多尺度地学习音频韵律相关信息,以改善语音合成模型合成语音语调平缓以及部分字词发音不自然的问题,提升合成语音表现力和自然度。

原始的Tacotron2模型结构主要由编码器和带有注意力机制的解码器两部分组成。相比于原始的Tacotron2模型,本文提出的模型添加了基于VAE的多尺度韵律编码网络和一个互信息估计器。基于VAE的多尺度韵律编码网络以参考音频的梅尔频谱和F_0作为输入,提取参考音频中句子级别的韵律信息和音素级别的音高信息,并输出韵律嵌入和音高嵌入,之后与文本嵌入一起输入解码器中,以合成音调丰富且富有韵律的合成语音。而互信息估计器会计算音高嵌入和韵律嵌入之间的互信息,并在训练过程中最小化它们之间的互信息,从而分离音高特征与其余韵律信息,使模型可以直接控制合成语音的音高特征。模型总体架构如图2所示。

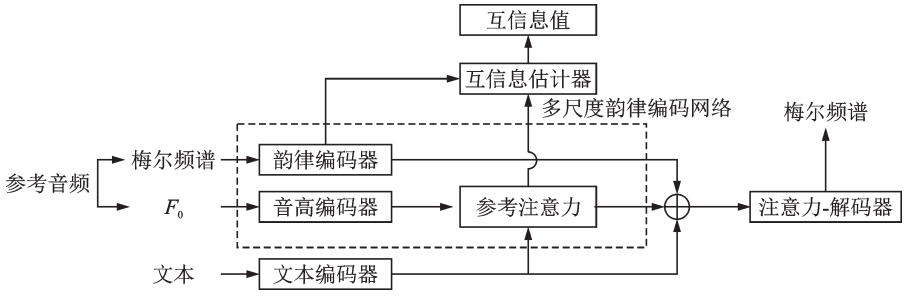

图2　本文所提模型架构

Fig.2　System architecture of the proposed model

2.1　基于VAE的多尺度韵律编码网络

如图2所示,基于VAE的多尺度韵律编码网络由韵律编码器、音高编码器和参考注意力3部分组成。韵律编码器和音高编码器都基于VAE模型。VAE模型早期在图像领域取得了不错的效果,其主要思想是对目标的潜在信息进行编码,然后利用潜在信息重建目标。本文VAE模型的输入是音频的声学特征$\boldsymbol{x}$,输出则是编码的潜在向量$\boldsymbol{z}$的分布。$\boldsymbol{z}$的分布通常为多维高斯分布,因此模型只需要预测多维高斯分布的均值和标准差,即

$$(\boldsymbol{\mu},\boldsymbol{\sigma})=\mathrm{VAE}(\boldsymbol{x}) \tag{1}$$

式中:$\boldsymbol{\mu}$和$\boldsymbol{\sigma}$分别为高斯分布的均值和标准差;$\boldsymbol{x}$为输入的声学特征。因为直接从高斯分布$N\left(\boldsymbol{\mu},\boldsymbol{\sigma}^2\right)$中采样潜在向量$\boldsymbol{z}$会导致无法计算网络的梯度,所以为了能够在不改变潜在向量分布的情况下使网络的梯度可以回传,模型采用重新参数化得到潜在向量$\boldsymbol{z}$,即

$$\boldsymbol{z}=\boldsymbol{\mu}+\boldsymbol{\sigma}\cdot\boldsymbol{\varepsilon} \tag{2}$$

式中$\boldsymbol{\varepsilon}$从标准高斯分布中采样获得,$\boldsymbol{\varepsilon}\in N(0,I)$。

如图2所示，本文采用参考音频的梅尔频谱和F_0分别作为韵律编码器和音高编码器的输入。其中，使用WORLD声码器[12]从音频波形文件中提取F_0序列。在每个音素的持续时间内对F_0值进行平均处理，从而得到音素级别的F_0，以便模型能够更好地学习每个音素的分布。每个音素的持续时间由其对应的帧数表示，这样确保F_0和持续时间在帧级别对齐。

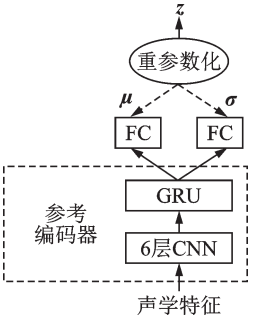

图3 VAE模型架构

Fig.3 System architecture of VAE model

基于VAE模型的韵律编码器和音高编码器具有相似的结构。如图3所示，它们都包含1个参考编码器和2个全连接层(Fully connected layers，FC)。参考编码器的结构与Skerry-Ryan等[13]提出的结构相同，具体来说，由6个CNN层和1个门控循环单元(Gate recurrent unit，GRU)层组成。对于韵律编码器，它学习音频句子级别韵律信息，将韵律信息编码为一个固定维度的向量，所以在GRU层，仅将GRU层的最后一个状态作为参考编码器的最终输出。而音高编码器学习参考音频音素级别的音高信息，它将音高信息编码为一个变长的向量序列，所以将音高编码器中的GRU层所有的状态都作为参考编码器的最终输出。参考编码器的输出通过2层FC，从而得到潜在向量$\boldsymbol{z}$分布的均值和标准差，最后通过重参数化操作得到音高嵌入序列$\boldsymbol{z}_{\mathrm{p}}$和韵律嵌入$\boldsymbol{z}_{\mathrm{s}}$。

为了便于将韵律嵌入、音高嵌入序列与文本嵌入序列连接，本文对韵律嵌入和音高嵌入做了如下处理。对于韵律嵌入向量在时间轴进行复制，将其拓展成与文本嵌入序列长度相同的序列。对于音高嵌入序列，将其通过参考注意力重组为与文本嵌入序列长度相同的序列。在本文的模型中，使用缩放点积注意力网络[14]作为参考注意力机制，学习音高嵌入序列和文本嵌入序列之间的对齐。首先将音高嵌入序列分为两个子序列，分别作为参考注意力机制的Key和Value。同时，文本嵌入序列作为Query输入。最后，参考注意力输出与文本嵌入序列长度相同的对齐的音高嵌入序列。处理后的音高嵌入，韵律嵌入与文本嵌入相连接，输入解码器进行梅尔频谱的预测。

在训练中，本文模型根据文本$\boldsymbol{y}_{\mathrm{t}}$以及从参考音频中学到的音高嵌入$\boldsymbol{z}_{\mathrm{p}}$和韵律嵌入$\boldsymbol{z}_{\mathrm{s}}$共同生成音频。并且向量$\boldsymbol{z}_{\mathrm{p}}$、$\boldsymbol{z}_{\mathrm{s}}$的先验分布定义为标准正态分布$N(0,I)$。因此，语音合成模型的条件生成分布为$p\left(\boldsymbol{X}\middle|\boldsymbol{y}_t,\boldsymbol{z}_{\mathrm{s}},\boldsymbol{z}_{\mathrm{p}}\right)$，其中$\boldsymbol{X}$表示语音。

按照VAE模型[7]的原理，它会引入两种变分分布$q\left(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{X}\right)$、$q\left(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{X}\right)$来近似变量$\boldsymbol{z}_{\mathrm{p}}$、$\boldsymbol{z}_{\mathrm{s}}$难以计算的真实后验分布，并通过最小化引入的变分分布与真实后验分布间的KL散度使变分分布接近真实后验分布。因此语音合成模型训练时的损失函数为

$$L_{\mathrm{TTS}}(\boldsymbol{y}_{\mathrm{t}},\boldsymbol{z}_{\mathrm{p}},\boldsymbol{z}_{\mathrm{s}})=E_{q(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{X})q(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{X})}\left[\ln p\left(\boldsymbol{X}|\boldsymbol{y}_{\mathrm{t}},\boldsymbol{z}_{\mathrm{p}},\boldsymbol{z}_{\mathrm{s}}\right)\right]-D_{\mathrm{KL}}\left(q\left(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{X}\right)\|N(0,I)\right)-D_{\mathrm{KL}}\left(q(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{X})\|N(0,I)\right) \tag{3}$$

式中：期望项$E_{q(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{X})q(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{X})}\left[\ln p\left(\boldsymbol{X}|\boldsymbol{y}_{\mathrm{t}},\boldsymbol{z}_{\mathrm{p}},\boldsymbol{z}_{\mathrm{s}}\right)\right]$为重建损失；$L_{\mathrm{TTS}}$表示模型的训练目标；$\boldsymbol{y}_{\mathrm{t}}$表示输入文本；$\boldsymbol{z}_{\mathrm{p}}$表示音高嵌入；$\boldsymbol{z}_{\mathrm{s}}$表示韵律嵌入；$\boldsymbol{X}$表示对应语音；$D_{\mathrm{KL}}$表示计算分布之间的KL(Kullback-Leikler)散度。

2.2 基于最小化互信息的特征分离

尽管基于VAE模型的方法已被证明VAE模型对不同特征有一定的分离能力[15]，但因为它是一种无监督的方法，所生成的潜在向量往往难以解释。并且由于各个韵律特征彼此纠缠在一起，因此对一个特征进行操作可能会影响其他维度。在语音合成中，得到可解释的独立语音特征对于实现对语音的精确控制至关重要。如2.1节所述，本文通过采用两个独立的VAE编码器，分别显式学习音高特征，隐式学习整体韵律特征，初步分离音高特征和其余韵律特征。

虽然通过对不同的特征分别独立建模,增强了特征嵌入间的独立性。然而,梅尔频谱作为韵律编码器的输入,它包含了大量信息,其中也包含关于音高的信息,所以韵律编码器学习的潜在向量空间中仍然对音高信息进行了编码。编码器所输出的音高嵌入并不完全独立于韵律嵌入,无法实现对于音高特征的单独控制。此外,考虑到在训练时参考音频与合成语音相同,多个编码器会造成信息重复,解码器可能会从韵律嵌入中获取音高信息,从而忽略音高嵌入提供的信息。在推理时,如果进行音高和韵律的非并行合成(即F_0和梅尔频谱不是来自同一个参考音频),解码器就会忽略目标音高信息,导致合成语音质量下降。针对以上分析,本文提出通过最小化不同特征嵌入之间的互信息,进一步区分音高特征和韵律特征,从而使模型可以直接控制音高特征,保证合成语音的质量。

互信息是一种基于香农熵的随机变量之间依赖关系的度量,它度量了两个变量之间相互依赖的程度,可以看成是一个变量中包含的关于另一个变量的信息量。韵律嵌入z_{s}和音高嵌入z_{p}的互信息$I\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)$等价于它们的联合分布$P_{z_{\mathrm{s}},z_{\mathrm{p}}}$与它们边缘分布乘积$P_{z_{\mathrm{s}}}P_{z_{\mathrm{p}}}$之间的KL散度[16],即

$$I\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)=D_{\mathrm{KL}}\left(P_{z_{\mathrm{s}},z_{\mathrm{p}}}\|P_{z_{\mathrm{s}}}*P_{z_{\mathrm{p}}}\right) \tag{4}$$

式(4)表达了联合分布与边缘分布的差异越大,变量之间的依赖性越强。具体来说,互信息越小,来自不同分布的两个向量之间的关系越低;互信息越大,相关性越高。当分布$P_{z_{\mathrm{s}}}$和$P_{z_{\mathrm{p}}}$独立时,$I\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)$等于0。因此本文通过最小化韵律嵌入$z_{\mathrm{s}}$与音高嵌入$z_{\mathrm{p}}$之间的互信息,降低不同嵌入之间的依赖性,从而将它们分离。

本文使用Belghazi等[17]提出的互信息估计器(Mutual information estimator,MINE)计算特征嵌入之间的互信息。该方法基于KL散度的Donsker-Varadhan表示[18]构造互信息的下界,有

$$I\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)\geqslant I_{\Theta}\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)=\sup_{\theta\in\Theta}E_{P_{\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)}}[T]-\log_{a}\left(E_{P_{\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)}}\left[\mathrm{e}^{T}\right]\right) \tag{5}$$

式中:T为可以使上述方程中的两个期望收敛的任何函数,在MINE中,通过使用深度神经网络得到T,这个方法通过梯度下降最大化关于T的下界,估计变量z_{s}、z_{p}之间的互信息;Θ为深度神经网络的参数。

在训练过程中,本文将同时最小化语音合成模型的重建损失和韵律嵌入z_{s}与音高嵌入z_{p}之间的互信息。由于互信息值为非负,如果神经互信息估计器输出的互信息值为负,本文将互信息值取为0。模型训练时的整体目标函数为

$$\min_{y_{\mathrm{t}},z_{\mathrm{p}},z_{\mathrm{s}}}\max_{\Theta}L_{\mathrm{TTS}}(y_{\mathrm{t}},z_{\mathrm{p}},z_{\mathrm{s}})+\lambda\times\max\left(0,I_{\Theta}\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)\right) \tag{6}$$

式中λ为平衡两种损失的超参数。在本文的实验中,λ设置为0.1。与常见的生成对抗网络的训练类似,在训练中的每一步交替更新语音合成模型和互信息估计函数T。具体来说,在训练中,通过最小化语音合成模型的损失L_{TTS}和嵌入z_{s}、z_{p}之间的互信息训练语音合成模型;通过最大化互信息下界$I_{\Theta}\left(z_{\mathrm{s}},z_{\mathrm{p}}\right)$训练互信息估计器。由于音高嵌入$z_{\mathrm{p}}$是可变长度的向量序列,本文会从音高嵌入序列中随机采样一个向量来计算互信息。通过上述过程训练模型,可以既保证语音特征重建的质量,又能使音高编码器和韵律编码器提取的信息相互独立。

3　实验评估

为了评估本文方法在语音合成任务上的性能,本文在中文语音语料库上进行了广泛的实验。在本节中,将介绍用于语音合成模型训练的语音数据集。此外,本节还将介绍模型具体实现细节、模型比较方法和评估方法。

3.1 数据集

本文主要的研究目的是可以合成韵律丰富,富有表现力的语音。而目前开源的中文语音数据集多是基于语音识别任务准备的,其语音一般不带情绪,音量、语速等韵律特征保持一致,不利于所提模型的训练。所以本文所有实验都基于一个内部中文语音数据集。它是一个高品质的有声读物数据集,数据集中的语音音频都是富有情感和韵律的。相较于目前的开源中文单人语音数据集,该数据集语音的韵律更加丰富多变,更利于进行表达性语音合成模型的训练。该数据集由一位男性说话人的简短语音录音片段组成,所有的录音片段均来自于一本有声小说。数据集为每段音频提供了对应的文本和三音素转录。每段音频的长度从1 s到1 min不等,一共包含8 312条单声道的录音片段,总时长大约为16 h。本文在数据集中随机抽取了100个语音片段以及相对应的文本作为后续测试的测试集。除此之外的所有数据均作为固定训练集,用于所提模型以及对比模型的训练。

对于数据集,首先本文将所有的录音音频重新采样为16 000 Hz,并对全部音频做了归一化处理,避免录音声音忽大忽小。然后对于每个录音音频,本文将每个语音片段的开头、结尾及语句中间的长时间(时长大于0.2 s)的静音片段替换为时长0.2 s的静音片段。在所提语音合成模型的训练中,需要每段语音相对应的梅尔频谱和F_0序列。在本文中,梅尔频谱通过窗长为1 024,帧长为1 024,帧移为256的短时傅里叶变化(Short time Fourier transform,STFT)得到。语音的F_0序列通过使用WORLD声码器从音频的波形文件中提取得到。根据音频对应的三音素序列,在每个音素的持续时间内进行平均。其中,音素的持续时间使用帧数表示,这保证了F_0序列与音素序列在帧级别的对齐。

3.2 实验设置

为了更加合理与准确地衡量所提模型的表达性语音合成效果,本文后续实验将对以下3种模型进行对比并分析实验结果。

基线模型1(Baseline 1):基线1采用原始的Tacotron2语音合成的方法。Tacotron2模型结构如第1节中所描述的,由编码器和带有注意力机制的解码器两部分组成。基线模型1以音素序列作为文本输入,通过编码器生成512维的文本嵌入序列。解码器根据文本嵌入生成预测梅尔频谱。

基线模型2(Baseline 2):基线2为一种单一尺度增强合成语音表现力的语音合成方法[8]。这种方法是一种基于VAE的单一尺度的表达性语音合成方法。这种方法在Tacotron2的基础上,加入基于VAE的句子级别的韵律编码器,从而学习音频句子级别的韵律信息。具体来说,参考音频的梅尔频谱作为韵律编码器的输入,编码器结构如图3所示。参考编码器根据输入生成32维的潜在向量。然后,潜在向量通过全连接层得到256维的韵律嵌入。对韵律嵌入进行复制,将其扩展到和文本嵌入序列长度相同,之后与文本嵌入序列连接,输入解码器预测梅尔频谱。

本文模型(The proposed):采用本文提出的多尺度表达性语音合成方法,具体结构如上文所述。其中韵律编码器根据输入的梅尔频谱生成32维的潜在变量,之后通过全连接层得到256维的韵律嵌入向量;音高编码器根据输入的F_0序列生成32维的潜在向量,之后通过全连接层和参考注意力,最终得到256维的对齐的音高嵌入向量序列。韵律嵌入经过复制,扩展成与文本嵌入序列长度相同。对齐的音高嵌入序列、扩展后的韵律嵌入与文本嵌入序列连接,并输入解码器进而预测梅尔频谱。

对于上述的所有模型,本文都使用单个NVIDIA TESLA V100 GPU进行训练训练。在训练时,Batch大小为64,初始学习率为1e−3,模型均使用Adam优化器。所有模型都使用3.1节中介绍的数据集所划分出的固定训练集进行训练,并且所有模型都训练200个左右的Epoch。在实验中,本文使用一个经过训练HiFi-GAN声码器[19]将语音合成模型生成梅尔频谱转换为相应的波形文件。

3.3 评　　估

为了衡量所提系统性能,本文从3.1节中介绍的测试集中抽取20个具有不同长度的样本作为固定的评估集。这些样本的时长均在20 s以内。在保证文本内容一致并排除其他干扰因素的前提下,本文主要使用主观打分对语音合成模型进行评价,同时也使用一些客观指标对模型进行辅助分析。

3.3.1　主观分析

合成语音主观评测采用了语音质量评价中常用指标:平均意见得分(Mean opinion score,MOS),即依靠人的听觉印象来对听到的语音进行评价打分。参与打分的志愿者共18人,母语均为汉语。志愿者根据自己的主观听觉感受对每条语音给出1~5分并以0.5分为1个跨度的分数。MOS分数越高,代表该语音的听感越好、越自然。

本文对基线模型1(Baseline 1)、基线模型2(Baseline 2)、本文模型(The proposed)以及真实音频分别进行了MOS评分,然后对获得的数据进行了分析处理。对比模型的具体结构如3.2节所述。最终MOS打分的均值及95%置信度区间的结果如表1所示,其中GroundTruth为真实音频。如表中所示,本文所提的方法表现优于其他2种基线方法,更接近于真实音频。相比于基线1,本文模型的MOS分数提高了约2.43%;相比于基线2,本文模型的MOS分数上升了约2.18%。主观评测结果显示相比于没有韵律信息基线模型1,加入了韵律信息的基线模型2和所提模型的MOS分数更高,证明了加入韵律信息有利于语音合成模型对于语音的重建。而将本文模型与基线模型2作比较,本文模型的MOS分数更高,证明了所提出的多尺度韵律编码器的有效性。相比于学习参考音频单一尺度的韵律信息,学习参考音频多尺度的韵律信息能帮助语音合成模型更好地对于语音的重建。主观听力实验证明所提模型的合成的语音更加自然,在人类听感上明显好于两种基线模型合成的语音,并更接近于真实的人类的语音。

表1　语音主观评测结果

Table 1　Subjective evaluation results of speech

方法	MOS值
GroundTruth	4.43±0.07
Baseline 1	4.11±0.07
Baseline 2	4.12±0.07
The proposed	4.21±0.07

3.3.2　客观分析

本文使用F_0帧错误率(F_0 frame error,FFE)[20]作为客观指标,用于测量出现音高误差大于20%或发声决策错误的帧的百分比,表达式为

$$E_{\mathrm{FFE}}=\frac{N_{\mathrm{UV}}+N_{\mathrm{VU}}+N_{\mathrm{FOE}}}{N}\times 100\% \tag{7}$$

式中:N表示帧的总数;N_{UV}表示真实值为清音被预测为浊音时的帧数;N_{VU}表示真实值为浊音被预测为清音时的帧数;N_{FOE}表示满足$\left|\frac{F_{0\,\text{估计值}}}{F_{0\,\text{参考值}}}-1\right|>\delta$条件的帧数,$\delta$是一个阈值,通常设置为20%。FFE用于计算预测音高和真实的音高之间的差异比值,可以体现出F_0轨迹的重构误差。FFE值越低证明预测值与真实值之间的误差越小。

本文对基线模型1(Baseline 1)、基线模型2(Baseline 2)和本文模型(The proposed)合成音频分别计算了FFE,结果如表2所列。相比于没有使用韵律信息的原始Tacotron2结构的基线模型1和仅在单一尺度上学习韵律信息的基线模型2,本文方法因为对F_0特征进行了音素级别的建模,其合成语音的FFE更低,对于F_0的预测更为准确。相比于基线1,本文模型的

表2　语音客观评测结果

Table 2　Objective evaluation results of speech

方法	FFE/%
Baseline 1	53.93
Baseline 2	57.05
The proposed	48.67

FFE下降了约10%;相比于基线2,本文模型的FFE下降了约14%。客观评测结果表明,本文方法相较于基线模型对于音高特征的预测更为准确,预测值与真实值之间的误差最小。

3.4 特征预测情况

图4给出了句子“母亲帮着黑娃说话了”的合成音频与真实音频的F_0对比图。通过F_0对比图,可以直观地观察可以看到,两个基线模型合成的音频与真实音频的时长相比,分别过长或过短。而本文模型合成的音频时长更接近与真实音频,其中对于停顿和语调等细节特征的预测相比两个基线模型更为准确。相比于两个基线模型合成的音频,本文模型合成的音频的音高包络也更接近与真实音频的音高包络。

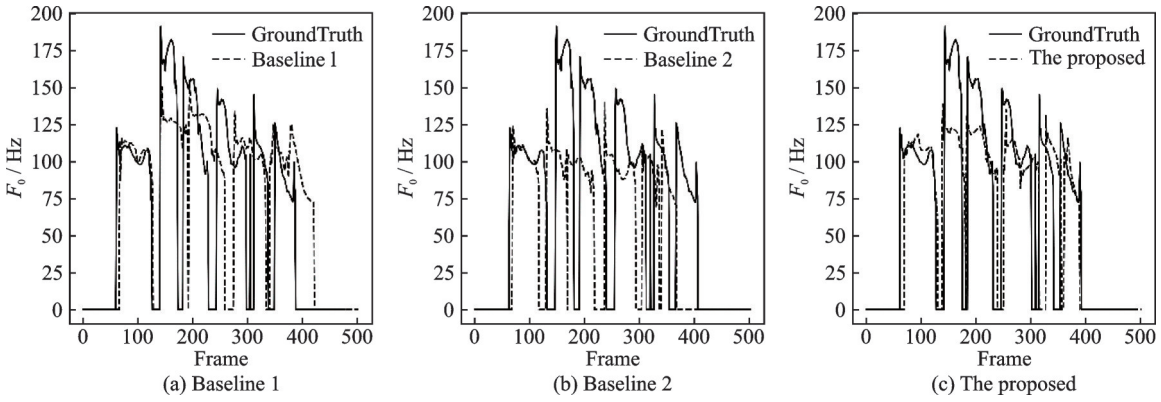

图4 合成语音F_0

Fig.4 Synthetic speech F_0

图5给出了句子“母亲帮着黑娃说话了”的真实的梅尔频谱图和所提模型预测的梅尔频谱图。通过对比观察特征图可以直接看到,合成语音频谱图中的频率形状和真实音频的大致一样,合成的梅尔频谱和原始音频的梅尔频谱十分接近,证明了本文模型可以高质量地重建音频。

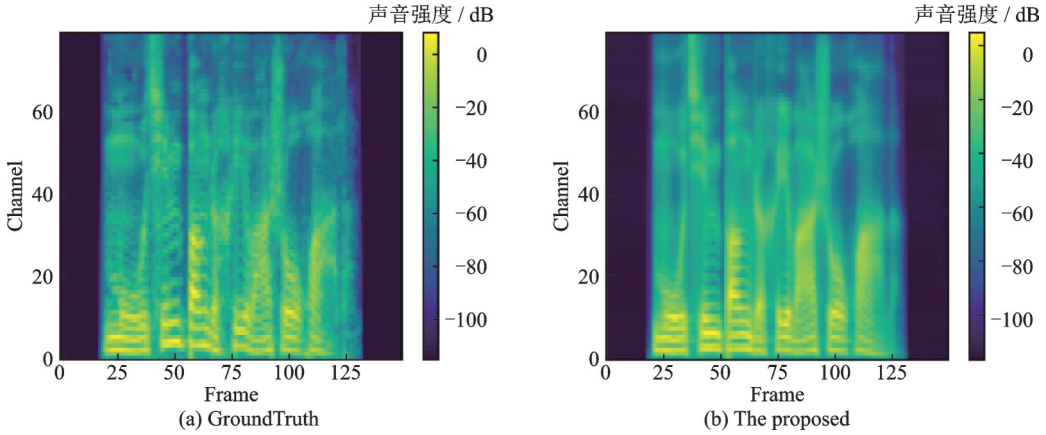

图5 合成语音梅尔频谱图

Fig.5 Synthetic speech Mel-spectrogram diagram

4 结束语

本文提出了一种多尺度富有表现力的汉语语音合成方法。该方法通过多尺度参考编码网络学习

音频中音素级别的音高信息和句子级别韵律信息,增强了合成语音的表现力,丰富了合成语音的韵律多样性。在训练过程中,通过最小化音高嵌入和韵律嵌入之间的互信息,对不同韵律特征进行分离。实验结果表明,本文提出的模型在客观评价和主观评价上性能都有所提升,证明了该方法在表达性汉语语音合成中的有效性,但是合成后的语音与真实目标语音仍存在一定的差距。在未来的工作中,将继续改进系统解决合成中出现的不正确停顿问题,以及合成中预测停止符不确定造成音频时长变短,语速过快的现象,进一步提高合成语音的质量。

参考文献:

[1] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[EB/OL]. (2017-04-06)[2023-01-13]. https://arxiv.org/abs/1703.10135.

[2] VASQUEZ S, LEWIS M. MelNet: A generative model for audio in the frequency domain[EB/OL]. (2019-06-04)[2023-01-13]. https://arxiv.org/abs/1906.01083.

[3] PING W, PEMG K, GIBIANSKY A, et al. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning[EB/OL]. (2018-02-22)[2023-01-13]. https://arxiv.org/abs/1710.07654.

[4] LI N, LIU S, LIU Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019: 6706-6713.

[5] TAN X, QIN T, SOONG F, et al. A survey on neural speech synthesis[[EB/OL].(2021-07-23)[2023-01-13]. https://arxiv.org/abs/2106.15561.

[6] WANG Y, STANTON D, ZHANG Y, et al. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 5180-5189.

[7] KINGMA D P, WELLING M. Auto-encoding variational Bayes[EB/OL]. (2022-11-10)[2023-01-13]. https://arxiv.org/abs/1312.6114.

[8] ZHANG Y J, PAN S, HE L, et al. Learning latent representations for style control and transfer in end-to-end speech synthesis [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 6945-6949.

[9] 许希明.英汉语声调的音系差异[J].宁波大学学报(人文科学版),2019,32(4): 71-77.
XU Ximing. Phonological difference of tone in English and Chinese[J]. Journal of Ningbo University (Liberal Arts Edition), 2019,32(4): 71-77.

[10] SHEN J, PANG R, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 4779-4783.

[11] 曹剑芬.汉语声调与语调的关系[J].中国语文,2002(3): 195-202,286.
CAO Jianfen. The relationship between tone and intonation in Mandrain Chinese[J]. Studies of the Chinese Language,2002(3): 195-202,286.

[12] 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, 99(7): 1877-1884.

[13] SKERRY-RYAN R J, BATTENBERG E, XIAO Y, et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 4693-4702.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY,USA: Curran Associates Inc., 2017: 6000-6010.

[15] HSU W N, ZHANG Y, WEISS R J, et al. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 5901-5905.

[16] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[17] BELGHAZI M I, BARATIN A, RAJESHWAR S, et al. Mutual information neural estimation[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 531-540.

[18] DONSKER M D, VARADHAN S R S. Asymptotic evaluation of certain Markov process expectations for large time[J]. Communications on Pure and Applied Mathematics, 1975, 28(1): 1-47.

[19] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 17022-17033.

[20] CHU W, ALWAN A. Reducing f_0 frame error of f_0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend[C]//Proceedings of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2009: 3969-3972.

作者简介:

高洁(1997-),女,硕士研究生,研究方向:语音合成,E-mail:gaoo0_0@163.com。

肖大军(1991-),工程师,研究方向:电网调度自动化、调控人工智能应用、源网荷储协同互动技术。

徐遐龄(1980-),女,博士,教授级高级工程师,研究方向:电网调度自动化,E-mail:xuxialing@foxmail.com。

刘绍翰(1974-),博士,讲师,研究方向:人工智能,E-mail: 267443267@qq.com。

杨群(1971-),**通信作者**,女,副教授,研究方向:语音处理、自然语言处理、文本挖掘、图像识别、目标检测,E-mail: qun. yang@nuaa. edu. cn。

(编辑:刘彦东)