个性化语音合成中说话人特征不同嵌入方式的研究*

汪涛[1,2]，易江燕[1]，傅睿博[1,2]，温正棋[1]，陶建华[1,2,3]

(1. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190；

2. 中国科学院大学 人工智能学院，北京 100190；

3. 中国科学院自动化研究所 中国科学院脑科学与智能技术研究中心，北京 100190)

文 摘： 个性化语音合成是语音合成中一个重要的研究领域。当前语音合成技术以基于端到端的框架为主，本文基于端到端语音合成，结合说话人识别技术提取出特定说话人的特征表示，研究了说话人嵌入至端到端语音合成系统中的不同方式对个性化语音合成的影响。本文结合之前的说话人嵌入方案又提出了两种说话人嵌入方式，通过在 VCTK 数据集中训练个性化语音合成系统并比较了三种说话人特征嵌入方式的不同，分析了不同方案合成语音的自然度和相似度，模型参数量的变化以及韵律上表现的效果，最终得出说话人特征嵌入较好的方式。

关键词：个性化语音合成；说话人识别；端到端语音合成

中图分类号：TP181

个性化语音合成的目的是建立一个能够合成任意说话人声音的语音合成系统，而不是仅仅能合成语料库中的语音，特别的在给定说话人语料很少的情况下，能够准确还原出说话人的音色。这样的技术可以有助于将语音合成推广到更大的应用范围。语音合成技术经历了拼接合成，基于参数统计的合成方法。目前以 Tacotron[1]和 WaveNet[2]结构为代表的端到端语音合成系统,合成出的声音主观评测评分能够达到 4.5 分以上。

但个性化语音合成目前还没有达到理想的阶段。个性化语音合成技术目前可以分为基于自适应的个性化语音合成和基于说话人特征嵌入的个性化语音合成。基于自适应的语音合成技术[11]希望通过大量语料学习一个通用的中性语音合成系统，然后再通过较少量的待合成说话人语音对网络进行调整，这样的方法合成出来的效果较好，但是一般需要几十分钟高质量的待合成说话人语料。然而想要获得一个人如此长时间的高质量语音往往是不切实际的。基于说话人特征嵌入的个性化语音合成技术[9][10]则不需要待合成说话人大量的语料，只需要通过给定的说话人语音提取出说话人特征嵌入到语音合成系统中即可，同时为了训练说话人特征提取模型，可以采用一些对音质要求不高仅含说话人信息的数据集，这可以在语音识别中获取大量关于说话人识别的数据集，如 LibriSpeech[8]数据集。2018 年 Jia Y[9]提出这种说话人嵌入技术，将提取的说话人特征 d-vector[4]向量嵌入到 Tacotron[1][5]中，但是采用这种方案合成出来的语音相似度不高，同时语音的韵律信息过于平滑，没有表现力。本文猜想这样的原因可能与说话人特征嵌入到 Tacotron[1][5]的位置有关系，为此我们提出了两种新的说话人特征嵌入方式，并且与 Jia Y[9]提出的特征嵌入方案进行比较，以此改进合成语音的效果。

说话人识别是用来判断一段语音来源于哪个说话人的技术，通过该技术也可以间接提取出说话人特征。本文采用了基于 d-vector[4]的说话人识别方案，其主要思路是利用深度神经网络对语音进行建模，建立特定的损失函数之后训练模型，并将网络最后一层的输出作为该语音的说话人特征表示，即 d-vector 向量。

声码器作为端到端语音合成中将声学特征转化为时序信号的模块，是一个典型的生成模型。目前以神经网络为基础的声码器可以合成出能够与真人媲美的合成效果。比较典型的神经网络声码器有 WaveNet[2]和 WaveRNN[6]。尽管 WaveNet 能合成出高音质的语音，但是其训练和合成速度都很慢，很难满足合成系统的实时性要求。然而 WaveRNN 在损失很小音质的情况下，能有比 WaveNet 快几十倍的训练和合成速度。为了满足个性化语音合成系统的实时性要求，本文选择 WaveRNN 作为声码器，并采用多说话人数据集 LibriSpeech[8]对 WaveRNN 进行训练作为一个通用的声码器。

下文将先介绍基于说话人特征嵌入的语音合成系统框架，然后再提出两种新的说话人特征嵌入

*基金项目：国家重点研发计划(No.2017YFB1002801)，国家自然科学基金(No.61425017, No.61831022, No.61771472, No.61603390)，以及中国科学院战略性先导科技专项（XDC02050100）。

作者简介：汪涛，男，安徽，硕士研究生。

通讯联系人：陶建华，研究员，E-mail：jhtao@nlpr.ia.ac.cn

方式，通过实验，比较不同说话人特征嵌入对于语音合成效果的影响。

1 个性化语音合成系统

本文采用的个性化语音合成系统框架如图 1 所示，其中包含了以 Tacotron[1][5]为基础的通用语音合成模块，说话人特征提取模块以及基于神经网络的声码器。下面将分别介绍各模块的具体内容。

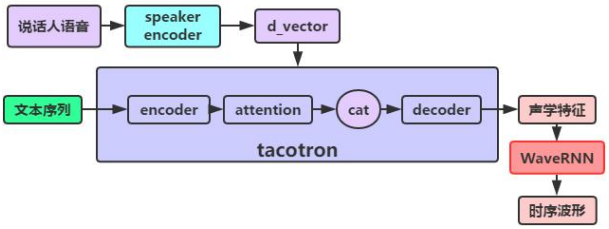

图 1：系统框架

1.1 通用语音合成模块

本文通过通用的语音合成模块学习语音合成的共性信息，以 Tacotron2[5]框架为基础。该框架简化了传统的管道式语音合成方案，能够直接将文本序列转化为可用于声码器的声学特征，同时也可以很方便的在前端插入韵律模型，从而使合成出来的声音更加自然。同时为了学习到自动停止功能，模块加入了输出截止的预测模块，通过结合梅尔谱损失，一块优化整个网络。在通用的语音合成模块中包含三个部分，编码器模块、注意力机制模块和解码器模块，分别对输入的文本序列进行编码，再利用注意力机制找出当前的注意信息，然后输入到解码器模块中进行解码生成声学特征。

1.2 说话人特征提取模块

说话人特征用于加入到通用的语音合成模块中，让网络能够根据给定的说话人特征合成该说话人的声音。为了能够合成更加相似的声音，对提取出的特征要求是能够尽可能的学习到说话人的音色信息而忽略其他信息，这对于合成出来的效果是否相似至关重要。

在此，我们采用的说话人特征提取的方案是以 d-vector[3]框架为基础的说话人识别系统。该网络通过将语音提取出来的梅尔谱特征映射到一个固定维度的隐含层中，从而实现说话人的特征提取任务。对于从隐含层中提取出的参数，被称为 d-vector 向量。该网络是通过优化一个端到端的说话人识别损失函数进行训练，思想是希望两个来自同一段说话人的 d-vector 向量距离尽可能的近而不同说话人的 d-vector 向量尽可能的远。训练该网络的数据不必来自于语音合成内的数据，也不需要文本信息，只需要标记出语音对应的说话人即可，同时可以使用较低音质的数据学习，而这样的数据可以在语音识别领域中大量获取。通过大量的数据训练，该网络可作为一个通用的说话人特征提取工具生成 d-vector 向量。这样也保证了对于集外的语音数据能够生成更好的说话人特征表示。

d-vector 向量提取流程为：对于 16k 的语音提取 40 维的梅尔谱参数，然后将这些提取出来的帧级序列输入到包含有三层具有 512 维的 LSTM 的网络中，每一层 LSTM 后包含一层全连接层，全连接层的输出维度为 64。最终的 d-vector 是通过最后一层的全连接层输出并做 L2 范式的标准化而得到。在推理阶段，对于一段长度不固定的语音，我们采用滑动窗的形式对语音进行切段，滑动窗长度为 800ms,步长 400ms。最终的 d-vector 是对所有的分段语音提取出的 d-vector 向量进行平均而获得的。

2 说话人特征不同的嵌入方式

Jia Y[9]采用的嵌入方案是在文本序列经过编码器模块编码之后与说话人特征相拼接而实现。具体的方式是将 d-vector 扩展为和编码器的输出序列长度一样的维度，然后进行拼接。通过这样的方式，相当于对每个文本序列加入了说话人特征，从而经过端到端语音合成训练之后能够合成不同的说话人。在此，我们将这样的嵌入方式称为在 attention 之前嵌入，作为实验的基准模型，该方式如图 2 所示。

采用在 attention 之前嵌入方式合成出来的声音存在相似度不高，韵律信息平滑等问题。本文猜想可能与说话人特征在合成系统中嵌入的位置有关。在文本经过编码器之后与说话人特征相拼接，此时对于说话人特征会有一个升采样操作以匹配经过编码之后的文本序列的维度，相当于对于每一个文本序列都引入了说话人特征，从而增加了网络的计算量。我们可以只需要一个说话人特征即可指导网络合成出个性化语音。在此，本文认为将上述的说话人特征嵌入到语音合成系统中还有其他不同的方式，除了在 attention 之前嵌入特征之外，还可以在 attention 之后嵌入特征以及在 attention 前后均嵌入特征。下文将详细介绍。

2.1 attention 之后嵌入

如图 3 所示，在 attention 之后的嵌入方式是在文本序列经过编码器编码和注意力机制之后，

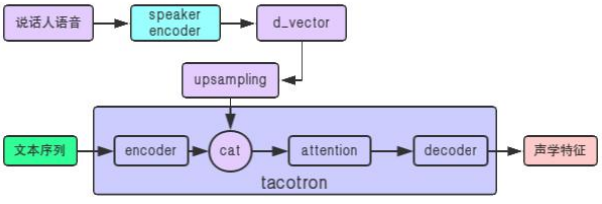

图 2：attention 之前嵌入

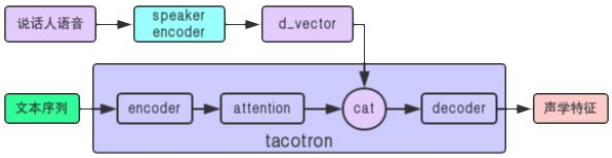

图 3：attention 之后嵌入

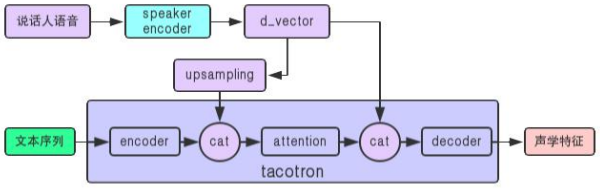

图 4：attention 前后均嵌入

此时已经通过注意力机制结合上下文信息计算出注意信息，所以不需要对 d-vector 进行升采样操作，而直接将其与计算出的注意信息拼接即可，然后指导解码器模块解码合成声学参数。通过这样的方式，我们可以理解为编码器和注意力模块学习出来的是一个通用的对于文本序列进行编码的模块，解码器是在给定说话人信息之后进行解码。我们将这种方式称为 attention 之后嵌入。

2.2 attention 前后均嵌入

如图 4 所示，attention 前后均嵌入的方式综合了上述两种方式，既在文本编码之后给每个文本序列赋予说话人信息，同时在 attention 之后也加入说话人信息，指导声学特征的预测。我们将这种方式称为 attention 前后均嵌入。

3 实验

3.1 实验数据

用于训练的语料集为 VCTK[7]。VCTK 包含有 44 小时的干净语音，其中包含有 109 个说话人，主要是英国式口音，语音采样率为 44k。我们将语音下采样为 22.05Khz。由于 VCTK 数据集中大部分语音前后静音段过长，对其计算帧能量再根据设定的阈值进行前后静音切除。并将语料分为三个子集：训练集，验证集（说话人包含在训练集中），测试集（说话人不包含在训练集中）。

说话人识别模块采用的训练语料为 LibriSpeech[8]。其包含有 436 小时的干净语音，共计说话人 1172 个，语音采样率为 16k。训练说话人识别模块时采用该语料集单独训练，训练结束之后，用于提取说话人特征，在个性化语音合成模型训练阶段不再调整参数。

3.2 实验设置

本文采用的通用语音合成模型为 Tacotron2 框架，用于训练的语料集为 VCTK[7]。Tacotron2 的实现根据 NVIDIA 的开源代码[1]。其中提取梅尔谱的参数设置为默认。

训练说话人识别模型使用了开源代码[2]提取 d-vector 向量。其中设置 d-vector 向量为 64 维。我们采用随机抽取该说话人对应的任意一个语音生成的 d-vector 向量作为输入的说话人特征。通过这样的乱序输入，可以丢弃 d-vector 中与文本无关的一面，从而使语音合成系统更加关注说话人本身的音色问题。同时乱序输入也相当于增加了数据量，因为通过随机挑选某一个说话人的 d-vector，理论上每次输入的文本和说话人配对是不一样的，从而相当于极大的扩充了数据，提高了模型的泛化性能，防止数据过拟合。

声码器部分采用的是 WaveRNN 算法，实现参考了开源代码[3]。该实现将语音进行 ulaw 编码之后，通过优化交叉熵损失更新网络参数。

3.3 模型参数量

表 1　参数量对比

方式	参数量	参数量提升百分比
通用模型	2.82e7	\
attention 之前嵌入	2.87e7	1.92%
attention 之后嵌入	2.85e7	0.96%
attention 前后均嵌入	2.90e7	2.87%

从表 1 中可以看出每种方式相对于通用模型所引入的参数量，比较得：attention 前后均嵌入>attention 之前嵌入>attention 之后嵌入。因为在 Tacotron2 框架下，具有两层 LSTM 模型，所以在 attention 之前嵌入时，需要将两层 LSTM 的维度均提高，相当于进入 attention 和解码器模块的数据维度是文本序列经过编码器之后的维度加上说话人特征的维度。所以将会加大后续模块总体的维度。

在 attention 之后嵌入说话人特征，不会影响 attention 模块的参数量，只会影响解码器阶段的 LSTM 模块以及后续预测声学特征和停止信息的全连接层的维度，所以在 attention 之后嵌入参数量要小一些。

3.4 主观评测和相似度

分别利用上述提出的三种嵌入方案实现个性化语音合成系统，并根据测试集中的说话人和文本作为系统的输入从而合成语音，并与原始语音

[1] https://github.com/NVIDIA/tacotron2
[2] https://github.com/HarryVolek/PyTorch_Speaker_Verification
[3] https://github.com/fatchord/WaveRNN

表 2　主观评测与相似度

方式	主观评测	相似度
attention 之前嵌入(基准)	4.20	2.51
attention 之后嵌入	4.30	2.60
attention 前后均嵌入	4.22	2.53

进行对比，分别对合成出来的语音的自然度和相似度打分，并计算了不同说话人嵌入方式相对于原来通用模型带来的参数量提升的百分比，数据结果如表 2 所示。

从表 2 可以看出，三者合成出来的声音主观评测差别不大并且效果都不错，可以达到很自然的程度。实验表明在 attention 之后嵌入说话人特征合成的语音更加自然一些，主观评测也有所提升。采用说话人嵌入的方式合成的相似度均不是很高，在 attention 之后嵌入说话人特征相对于其他两种方式效果要好一些。对比每种方式引入的参数量可以看出，采用在 attention 之后嵌入说话人特征，由于避免了对说话人特征采用升采样操作，从而显著的降低了引入的参数量，最终学习出的模型泛化性能更好，在测试的阶段模型表现也更加稳定。

3.5 梅尔谱

图 4-6 是同一段文字嵌入说话人特征后生成的梅尔谱对比，其中左边是生成的梅尔谱，右边是真实的梅尔谱。

可以看出，虽然三种方式都能够学习出比较清晰的谱信息，但是在生成的梅尔谱参数上表现得很平滑，而原始谱富含更多韵律上的信息。分析原因是模型的结构所造成的，通用的模型由于输入的数据包含有很多说话人，而不同的说话人具有不同的韵律特征，通用的语音合成模块只能够学习到众多人平均之后的韵律信息。同时由于我们嵌入到通用说话人模型中的是说话人特征，其中已经尽可能地去除了韵律特征对说话人的影响，所以总的模型无法学习出不同说话人各自的韵律特征，只能够学习到一个经过众多说话人平均后的韵律。

4　结论

本文研究了将说话人特征嵌入到端到端语音合成系统中的三种不同的方式。实验结果表明：三种嵌入的方式均能够合成出自然的语音，但在相似度上还有提升的空间。同时本文提出了在

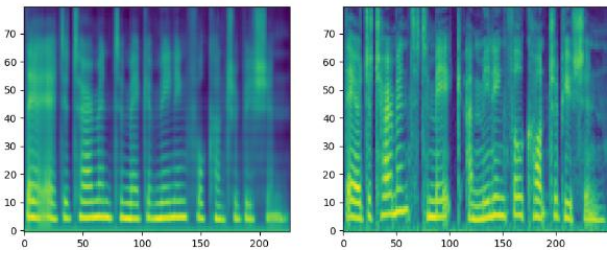

图 4：attention 之前嵌入

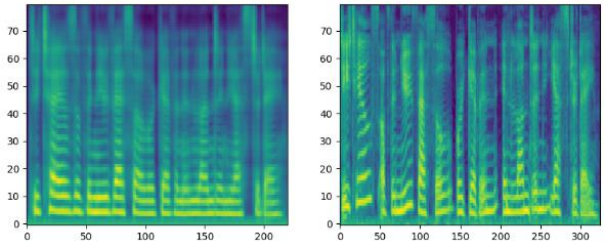

图 5：attention 之后嵌入

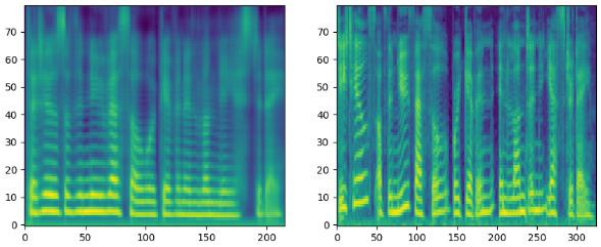

图 6：attention 前后均嵌入

attention 之后嵌入说话人特征的方法并且实验表明合成出的声音更加稳定和自然，并引入相对更少的参数量，易于训练。采用直接提取 d-vector 向量作为说话人特征嵌入到语音合成系统中合成出来的声音在韵律上表现过于平滑，不具有表现力，这也是未来需要改进的一点。未来工作将改进说话人特征的提取方式，使提取出的特征对于不同说话人更具有代表性。同时也考虑在个性化语音合成中再加入代表说话人韵律的信息，以此提高合成声音的表现力。

参 考 文 献

[1] Wang Y, Skerry-Ryan, RJ, Stanton D, et al. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model[J]. 2017.

[2] Oord A V D , Dieleman S , Zen H , et al. WaveNet: A Generative Model for Raw Audio[J]. 2016.

[3] Matjka P , Glembek O , Castaldo F , et al. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification[C]// IEEE International Conference on Acoustics. IEEE, 2011.

[4] Heigold G , Moreno I , Bengio S , et al. End-to-End Text-Dependent Speaker Verification[J]. Computer Science, 2015:5115-5119.

[5] Shen J , Pang R , Weiss R J , et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions[J]. 2017.

[6] Kalchbrenner N , Elsen E , Simonyan K , et al. Efficient Neural Audio

Synthesis[J]. 2018.

[7] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.

[8] Panayotov V , Chen G , Povey D , et al. Librispeech: An ASR corpus based on public domain audio books[C]// ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[9] Jia Y , Zhang Y , Weiss R J , et al. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis[J]. 2018.

[10] Skerry-Ryan R , Battenberg E , Xiao Y , et al. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron[J]. 2018.

[11] Taigman Y , Wolf L , Polyak A , et al. VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop[J]. 2017.

Research on Different Embedding Ways of Speaker Characteristics in Personalized Speech Synthesis

WANG Tao[1,2], YI Jiangyan[1], FU Ruibo[1,2], WEN Zhengqi[1], TAO Jianhua[1,2,3]

(1. National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, 100190;
2. School of Artificial Intelligence, University of Chinese Academy of Science, 10090;
3. CAS Center for Excellence in Brain Science and Intelligence Technology, CASIA, 100190)

Abstract: Personalized speech synthesis is an important research field in speech synthesis. Currently, speech synthesis technology is mainly based on end-to-end framework. This paper extracts the feature representation of specific speakers based on end-to-end speech synthesis and speaker recognition technology, and studies the influence of different ways of speakers embedded into end-to-end speech synthesis system on personalized speech synthesis effect. This paper compares three schemes for embedding speaker features into end-to-end speech synthesis systems. The naturalness and similarity of speech synthesized by different schemes, the variation of model parameters and the effect of prosody are analyzed.

Keywords: personalized speech synthesis; speaker recognition; end-to-end speech synthesis