Computer Science and Application 计算机科学与应用, 2023, 13(1), 126-135
Published Online January 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.131013

Hans汉斯

基于全局风格嵌入的多说话人印尼语语音合成

杨益灵，杨　鉴*，王发亮

云南大学，信息学院，云南　昆明

收稿日期：2022年12月26日；录用日期：2023年1月23日；发布日期：2023年1月31日

摘　要

由于印尼语高质量语料数据库的稀缺，该语种多说话人语音合成系统性能仍有待提升。因此以缓解低资源对多说话人语音合成性能的影响为目的，研究并实现了基于GST-Tacotron2模型框架的印尼语端到端语音合成系统。选用8.5小时的单说话人印尼语数据训练的合成系统，合成语音的MOS评分达4.11。在此基础上，设计多说话人印尼语语音合成系统，着重探索了在仅利用其他印尼语说话人少量语音数据进行混合训练时，采用说话人编码方法对多说话人合成自然度的影响。实验结果表明，利用合计14.5小时多说话人语音数据训练的合成模型，主位说话人合成语音的MOS评分到达了4.12，梅尔倒谱失真比单说话人最优模型降低了7.2%。其他说话人合成语音的MOS评分均大于3.60，验证了所提方法的有效性。

关键词

语音合成，多说话人，风格迁移，低资源，印尼语

Multi-Speaker Indonesian Speech Synthesis Based on Global Style Embedding

Yiling Yang, Jian Yang*, Faliang Wang

School of Information Science and Engineering, Yunnan University, Kunming Yunnan

Received: Dec. 26th, 2022; accepted: Jan. 23rd, 2023; published: Jan. 31st, 2023

Abstract

Due to the scarcity of high-quality Indonesian corpus databases, the performance of Indonesian multi-speaker speech synthesis systems still needs to be improved. Therefore, in order to alleviate the impact of low-resources on the performance of multi-speaker speech synthesis, an end-to-end

*通讯作者。

文章引用: 杨益灵, 杨鉴, 王发亮. 基于全局风格嵌入的多说话人印尼语语音合成[J]. 计算机科学与应用, 2023, 13(1): 126-135. DOI: 10.12677/csa.2023.131013

speech synthesis system in Indonesian based on the GST-Tacotron2 model framework is studied and implemented. A synthesis system trained on 8.5 hours of single-speaker Indonesian data achieves a MOS (Mean Opinion Score) score of 4.11 for synthesized speech. On this basis, a multi-speaker Indonesian speech synthesis system is designed, and the influence of the speaker coding method on the naturalness of multi-speaker synthesis is emphatically explored when only a small amount of speech data of other Indonesian speakers is used for hybrid training. The experimental results show that the MOS score of the synthesized speech of the main speaker reaches 4.12 using the synthesis model trained with a total of 14.5 hours of multi-speaker speech data. The MCD is 7.2% lower than the single-speaker optimal model. The MOS scores of the synthesized speech of other speakers are all greater than 3.60, which verifies the effectiveness of the proposed method.

Keywords

Speech Synthesis, End-to-End, Style Transfer, Low-Resource, Indonesian

1. 引言

语音合成是将输入文本转换为相对应的语音的过程，也称为 Text-to-speech (TTS)，语音合成的目标是在给定要合成的文本的情况下，能够生成自然发音的语音信号。近年来，基于深度神经网络的端到端 TTS 在产生自然可懂和带有情感的语音方面表现出相当强大的能力[1]-[7]。与统计参数语音合成相比，端到端 TTS 更易于设计，而且还允许进行组件嵌入，提高了模型的可操作性。

WaveNet [8]实现了直接从语言特征生成波形，可以说是早期真正意义上的现代神经 TTS 模型。WaveNet 是一个自回归的生成模型，它在语音特征提取，语音合成和音乐生成上取得了明显的效果。在语音合成领域，Tacotron [1]是近年来先进的端到端语音合成模型，可以直接从字素或音素生成语音，而 Tacotron2 [2]是 Tacotron 的改进版本。

Tacotron2 可以看作是一个基于注意力机制的自回归端到端模型，在该模型中，通过注意机制从文本中总结上下文信息，然后在端到端结构中进行训练。这种模型架构摆脱了原有的系统框架，直接将文本转换为语音，简化了文本处理的步骤，需要较少的专业语言知识和人力。然而，每种语言的语言特征有很大差异，Tacotron2 语音合成系统需要根据不同语言的语言特征进行优化，以便从输入文本中获得有用的隐式声学特征表示。

印度尼西亚语(Bahasa Indonesia)，简称印尼语，是印度尼西亚共和国的官方语言。在语言学分类中，印尼语属于马来－波利尼西亚语族西印度尼西亚语支。通用印尼语采用拉丁字母拼写系统，共 26 个字母，音素是其最小语音单位，共 34 个音素。相较于以英语为代表的通用语种，作为非通用语种的印尼语在语音合成领域仍存在一些问题，例如对于印尼语应用端到端语音合成方法的研究还相对较少，印尼语可用平行语料的缺乏，优质训练数据缺乏等。

作为一个先进的端到端语音合成系统，尽管 Tacotron2 在英语语音合成可懂度与自然度方面优于传统的统计参数语音合成系统，取得了显著的效果。但如果要将 Tacotron2 应用于低资源非通用语种的语音合成的研究工作，则还需要对其进行优化和改进[9]。本文设计并实现了以嵌入全局风格令牌(Global Style Tokens, GST) [10]的 Tacotron2 作为系统框架的印尼语语音合成系统，在低资源的情况下合成了具有风格

表现力的印尼语。针对 Tacotron2 在训练阶段过于依赖真实语音产生过拟合从而导致的暴露偏差(Exposure Bias) [11]问题，本文使用了渐变式交替训练的方法。针对无法在有限训练步数稳定地训练出注意力对齐图的问题，本文采用了迁移学习策略。在此基础上，还提到了多说话人混合训练，语言信息能够被共享，风格信息能被更精细地提取，并设计了多组低资源多说话人混合训练实验对 GST 以及嵌入说话人编码的 GST 的风格迁移性能进行了对比验证。

2. 语音合成模型及训练方法

在端到端印尼语语音合成系统中，可以直接将印尼语文本序列作为编码器的输入来进行训练。但是对于以 Tacotron2 为代表的基于深度神经网络的端到端语音合成系统，由于模型训练的信息只由输入的文本数据提供，当文本数量很少时，提取的信息通常不够丰富，导致系统难以取得很好的效果。因此，本文嵌入了风格特征提取模块 GST，以 GST-Tacotron2 模型作为基础架构，针对低资源语音合成提出了以下的方法对其进行优化。

2.1. 模型架构

本文设计的端到端印尼语语音合成系统基于 GST-Tacotron2 模型实现，其结构如图 1 所示。

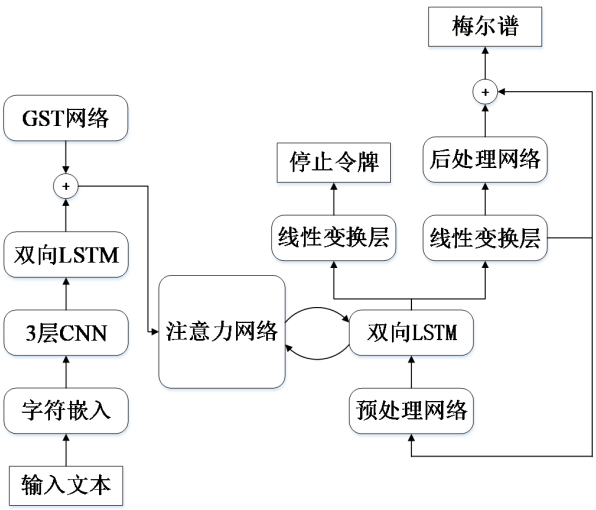

Figure 1. Diagram of GST-Tacotron2 model structure
图 1. GST-Tacotron2 模型结构框图

基于 GST-Tacotron2 模型构建的印尼语语音合成系统输入采用印尼语音素序列，Tacotron2 词嵌入维度为 512，GST 词嵌入维度为 256。Tacotron2 编码器为 3 层卷积核尺寸为 5 的 CNN 和一个 512 单元的 BiLSTM；解码器预处理网络为 2 层全连接层，隐含层节点个数为 256，解码器核心结构为两层 1024 单元的 LSTM 和一个位置敏感注意力网络；停止令牌通过一个激活函数为 sigmoid 的全连接层完成预测；后处理网络为5层卷积核为5的CNN，采用残差连接。本文所有实验都使用 Waveglow [12]来代替 WaveNet 作为声码器将声学参数转换为最终的语音波形。

GST 网络以两种方式完成推断过程。方式一：直接向 GST 网络输入任意的一个韵律音频，利用训练好 GST 网络建模输入音频的韵律特征，再将提取的韵律特征加入语音合成系统，合成出带有输入音频韵律的语音；方式二：取出训练完成的可学习向量组中的某一个向量，同编码器提取的文本信息，一起送入解码器，合成出音频。

2.2. 渐变式交替训练方法

由于训练和测试的解码模式不一致，造成模型过于依赖真实音频数据，当 Tacotron2 试图合成一个长句时，它就容易受到自回归过程中的误差积累的影响。在训练过程中，有真实数据的帮助，模型会取得较好的效果，但是在测试阶段因为不能得到真实数据的支持，在面对集外数据时模型就会变得无比脆弱，表现很差。这就是 Tacotron2 所面临的暴露偏差(Exposure Bias)问题。

为了缓解暴露偏差问题所带来的不利影响，本文提出了渐变式交替训练方法(Progressive Alternate Training Method)，训练阶段可供选择的教师强迫(Teacher Forcing, TF)和教师强制自由运行(Free Running, FR)两种模式以一个概率值交替进行，如公式(1)、(2)所示。

$$P_{TF} = 1 - \log_{10}\left[(\gamma + \alpha)/\alpha\right] \tag{1}$$

$$P_{FR} = \log_{10}\left[(\gamma + \alpha)/\alpha\right] \tag{2}$$

式中，P_{TF} 为解码器采用 Teacher Forcing 模式进行训练的概率；P_{FR} 为解码器采用 Free Running 模式进行训练的概率；γ 为当前训练轮次，γ 初始值为 0，每完成一个训练轮次 γ 增加 1，直至完成最后一个训练轮次 γ 达到最大值；α 为交替训练超参数，根据训练轮次选择合适的 α 使最后一个训练周期 P_{TF} 保持在 40%至 50%之间为宜。应用渐变式交替训练的解码流程图见图 2，两种模式由公式(1)、(2)决定的渐变概率值随机地交替进行。

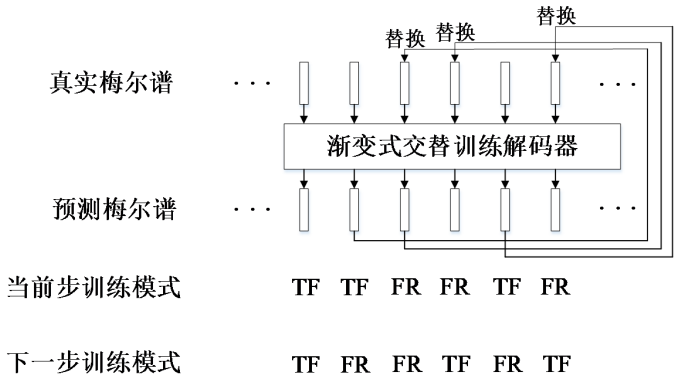

Figure 2. Flowchart of gradual alternate training decoding
图 2. 渐变式交替训练解码流程图

2.3. 预训练任务

针对端到端印尼语语音合成系统训练过程中模型鲁棒性差难以收敛的问题提出了添加预训练任务的策略。在单说话人语音合成实验中，将英语作为预训练任务并以此迁移到目标语言印尼语；在多说话人语音合成实验中，将单说话人实验中的印尼语作为第一说话人的预训练任务并以此迁移到目标多说话人印尼语语音合成系统中。

2.4. 说话人风格迁移

GST-Tacotron2 可以实现风格迁移，在低资源多说话人的情况下，提供额外的说话人信息会有助于说话人风格的学习。说话人编码由一组说话人嵌入向量(Language Embedding)组成，说话人嵌入向量为可学习向量，向量的数目为说话人数目，向量的词嵌入维度为 128 维。说话人编码作为文本所提供的额外信息，需要事先在文本中进行标记。同一说话人的所有文本都用同一说话人标记，一个说话人标记与一个

说话人嵌入向量一一对应。

3. 实验与结果分析

3.1. 实验数据与参数设置

实验中所使用的第一批印尼语数据集为实验室自建，由专业的播音员录音，播音员口音为印度尼西亚语标准口音，平行语料共有 4751 句，音频总时长为 8 小时 40 分钟，划分为训练集 4681 句(约 8 小时 30 分钟)，验证集 50 句，测试集 20 句。第二批印尼语数据集为开源多说话人数据集，音质稍差，根据低资源语音合成实验的需要，选择适当规模的语料进行实验。每个说话人分别选用 500 句(约 1 小时)或 250 句(约 0.5 小时)作为训练集，25 句作为验证集，20 句作为测试集。

实验中所使用音频的采样率均为 22,050 Hz，16 位 PCM 编码。整个实验基于 PyTorch 深度学习框架搭建模型，使用一块英伟达 RTX3090 显卡来训练模型，模型的训练批次设置为 32，训练轮次设置为 500。实验中使用了 Waveglow 声码器将声学参数转换为语音波形。GST 模块多头注意力网络，注意力头的数目设置为 8，可学习向量组数目设置为 10，每个向量的维度为 256，合成时的参考音频根据文本内容选用大致符合语境的。说话人风格迁移实验中，说话人编码的说话人嵌入向量维度为 128 维。其他参数每组实验均与 GST-Tacotron2 基线系统相同。

3.2. 多说话人实验设计

在多说话人混合训练实验中，采用 GST-Tacotron2 架构设计了以下 8 组以第一批印尼语数据集为主体的多说话人语音合成实验和 2 组单说话人语音合成对照实验。其中 2 组单说话人实验，即实验 1 和实验 2，分别为渐变式交替训练 GST-Tacotron2 实验和采用英语预训练的渐变式交替训练 GST-Tacotron2 实验，引入作为平均主观意见评分的对照。实验 1 不采用预训练，实验 2 使用从 LJSpeech [13]获取的 24.6 小时的英语数据集训练英语语音合成系统以此作为预训练任务。为了能稳定地训练出可用的语音合成系统，其余 8 个实验均采用渐变式交替训练以及预训练，并且每个实验的训练集均固定使用实验 2 中训练出的印尼语语音合成系统作为印尼语预训练任务。本节的实验设计如表 1 所示。

Table 1. Speaker style transfer experimental design
表 1. 说话人风格迁移实验设计

实验	说话人数	说话人	训练集规模(平行语料/单位：对)	是否采用说话人编码
实验 1	1	Sp0	4681	否
实验 2	1	Sp0	4681	否
实验 3	4	Sp0 + Sp1~Sp3	4681 + 3 × 250	否
实验 4	4	Sp0 + Sp1~Sp3	4681 + 3 × 250	是
实验 5	4	Sp0 + Sp1~Sp3	4681 + 3 × 500	否
实验 6	4	Sp0 + Sp1~Sp3	4681 + 3 × 500	是
实验 7	7	Sp0 + Sp1~Sp6	4681 + 6 × 250	否
实验 8	7	Sp0 + Sp1~Sp6	4681 + 6 × 250	是
实验 9	7	Sp0 + Sp1~Sp6	4681 + 6 × 500	否
实验 10	7	Sp0 + Sp1~Sp6	4681 + 6 × 500	是

8 组说话人风格迁移实验是为了比较原始的 GST-Tacotron2 与采用说话人编码的 GST-Tacotron2 在多种低资源情况下的风格提取能力与语音合成性能而设计的。Sp0 代表优质的第一批印尼语数据集的单说

话人(称为主位说话人)，在所有实验中固定使用 4681 对平行语料，不对其缩小规模。“Sp1~Sp6”，即“Sp1，Sp2，……，Sp6”代表音质稍差的第二批印尼语数据集中的“说话人 1，说话人 2，……，说话人 6”；“Sp1~Sp3”，即“Sp1，Sp2，Sp3”。其他有关数据集的详情已经在 3.1 节中说明。

注意力对齐的结果直接影响合成音频的质量。图 3 为可视化注意力对齐结果，注意力对齐结果呈对角线状，表示在生成音频序列时，解码器集中在正确的音素上，保证了每个字符的发音正确。

分析图 3 可视化对齐结果可发现，实验 10 的注意力对齐效果最佳，实验 9 也获得了较好的对齐图，但是其中存在一些断点和对齐模糊的部分。在其他采用说话人编码的实验组中，实验 6 的注意力对齐效果与实验 10 相当，实验 8 对齐图出现少许断点和模糊片段，实验 4 的对齐质量较实验 8 稍差，还没能正确预测出停止令牌。而在未采用说话人编码的实验组中，除了实验 9 以外，其余三组对齐图均存在明显的断点和噪点，实验 3 尤为突出。在风格数据极低资源的情况下，对比实验 3 和实验 4 可发现，采用了说话人编码的实验 4 在合成阶段的表现良好，对齐图断点较少；而未采用说话人编码的实验 3 的对齐图出现明显断点和模糊部分，总体合成效果很不理想。

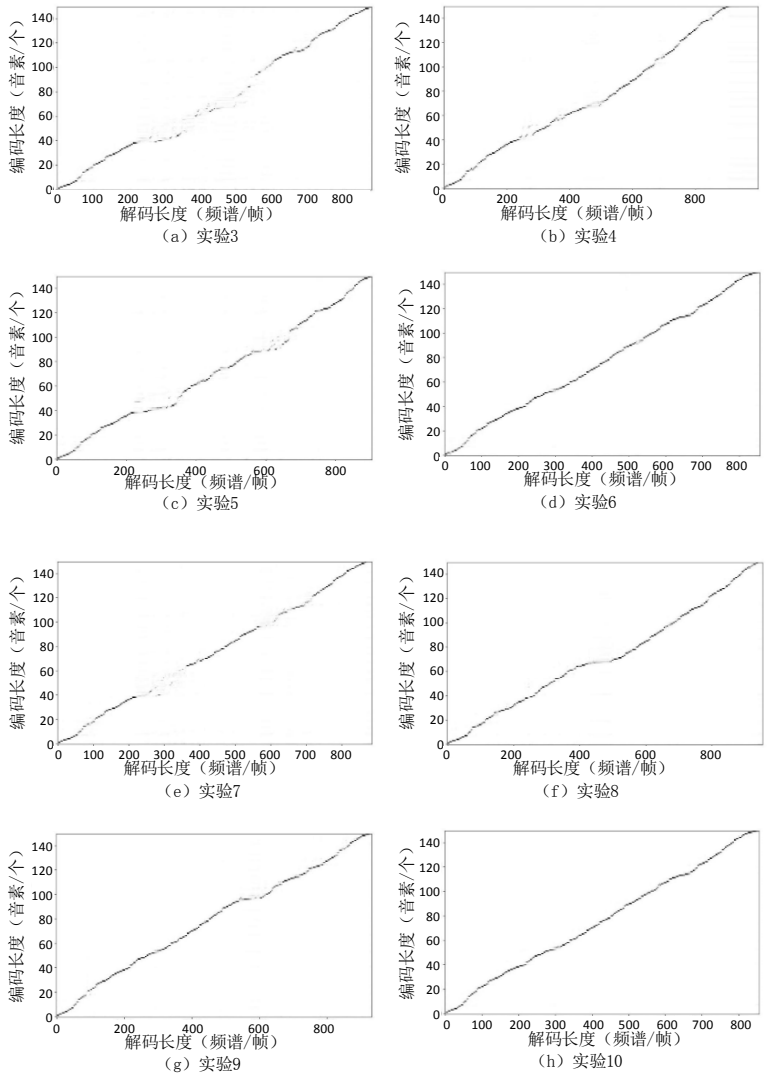

Figure 3. Visual attention alignment results of multi-speaker experiment
图 3. 多说话人实验可视化注意力对齐结果图

观察图 4 梅尔频谱图，我们注意到实验 10 和实验 6 的谱线较为自然且细节丰富，合成质量最好。实验 9 和实验 8 有不自然停顿但总体效果较好，实验 4 除了最后的部分因为没有正确预测出停止令牌而导致频谱图末端出现空白以外效果良好。实验 7 的频谱图良好但是其谱线缺乏细节，风格单一，韵律特征没有体现出来。实验 3 和实验 5 的频谱图断点过多，多为不符合句子韵律的不自然停顿，效果相对较差。总体而言，采用了说话人编码的 GST-Tacotron2 对于风格的提取更精细，对于停顿和语调的预测更准确，而不采用说话人编码的 GST-Tacotron2 在风格提取任务上的表现相对较差，尤其是在所需风格数据极低资源的情况下。

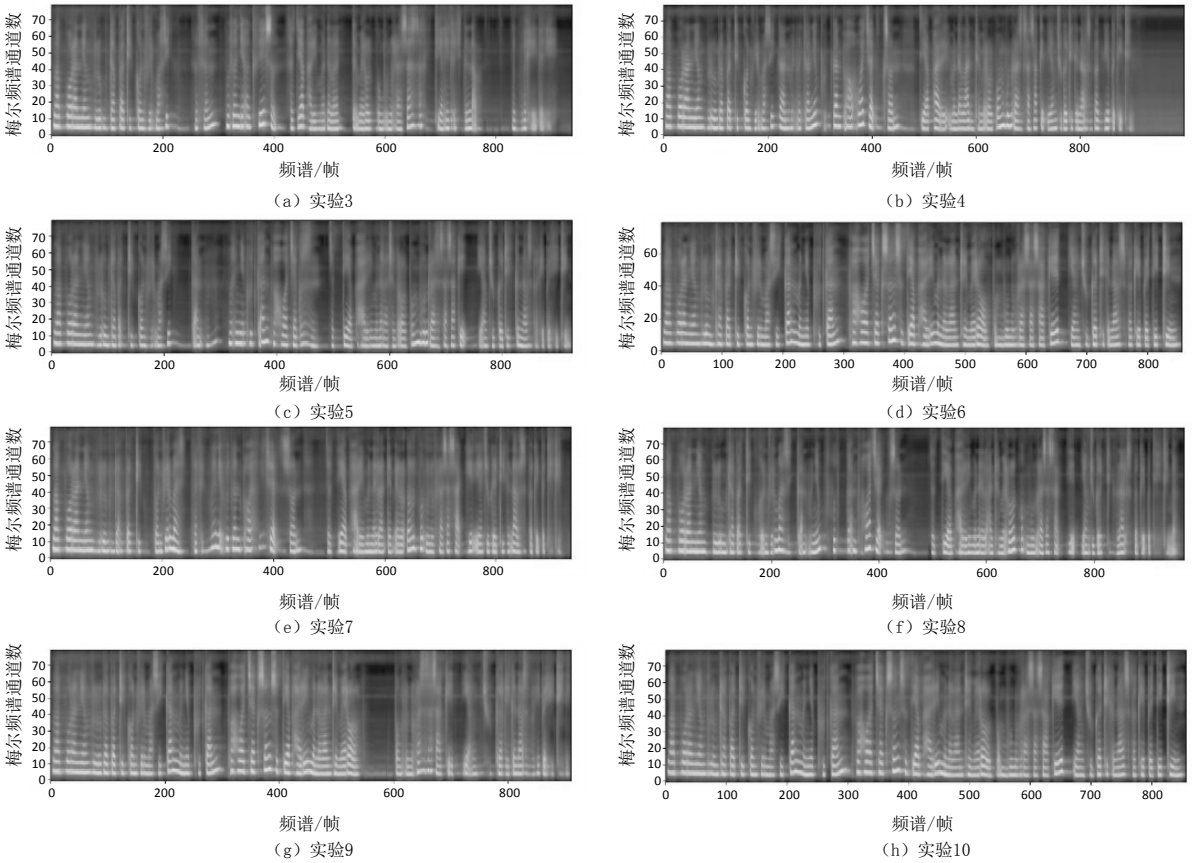

Figure 4. Mel-spectrogram of multi-speaker experiment
图 4. 多说话人实验梅尔频谱图

为了评价合成系统的性能，本文使用梅尔倒谱失真(Mel-Cepstrum Distortion, MCD)作为衡量音频质量的客观评测指标。梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)考虑了人耳对频率的非线性感知特性，MCD 把 MFCC 作为语音信号的特征描述，并用于表示合成音频与参考音频的客观失真距离。

$$D_{MCD}=\frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{k=1}^{K}\left(c_{t,k}-c_{t,k}^{*}\right)^{2}} \tag{3}$$

其中，参数 K 为 MFCC 的维数，T 为所有音频的总帧数，$c_{t,k}$ ($c_{t,k}^{*}$)为参考音频(合成音频)第 t 帧第 k 维梅尔频谱倒谱系数。D_{MCD} 为 MCD 值，取值大于等于零，其数值越小表示合成音频与参考音频的客观失真距离越小，模型效果越佳。

为了让合成音频和真实的参考音频能更好地进行比较，先对合成音频和参考音频进行了对齐处理，

即动态时间归整(Dynamic Time Warping, DTW)。评测内容为每组实验的每一个说话人风格测试集 20 句印尼语合成音频，结果取均值。本实验 MCD 评分见表 2。

Table 2. Evaluation results of MCD
表 2. MCD 评测结果

实验结果	MCD						
	Sp0	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
真实语音	0	0	0	0	0	0	0
实验 1	5.54						
实验 2	5.28						
实验 3	5.10	6.97	5.57	13.23			
实验 4	5.01	6.88	5.32	13.10			
实验 5	5.20	7.63	5.53	15.19			
实验 6	5.32	7.06	5.23	15.46			
实验 7	5.15	7.12	5.62	15.10	8.91	9.35	7.97
实验 8	5.56	7.28	5.85	14.45	8.76	6.77	7.89
实验 9	4.96	7.09	5.37	17.84	9.51	6.88	8.49
实验 10	4.90	6.56	5.56	15.85	8.10	5.37	6.56

分析 MCD 客观评测结果，可以看到，实验 10 的 Sp0 说话人风格的 MCD 值 4.90 为实验组中最优结果，相较实验 2 的 5.28 降低了约 7.2%。Sp1，Sp4，Sp5 和 Sp6 说话人风格的 MCD 值也为实验组中最优结果，其他说话人风格的结果也整体较好。通过采用可获得的音质稍差的同一语种不同说话人的语料来扩充训练集并嵌入 GST 与说话人编码来分离信息并建模韵律的方法，的确能够使系统的合成效果得到提升。在 7 个说话人的实验组中，采用说话人编码的实验组 MCD 值通常更低。对于 Sp5 说话人风格的 MCD 值，实验 10 相较实验 9 降低了约 21.9%，实验 8 相较实验 7 降低了约 27.6%，其他几组实验也大体呈现这个趋势。但是对于训练用第一批语料 Sp0 和第二批语料中音质相对较好的 Sp2，采用说话人编码的方法并不能使效果得到很大的提升，反而还会小幅下降。因此，采用说话人编码的系统对于所提供的训练语料低资源且音质稍差的情况，适应能力要强于基线系统，但是对于训练语料充分且语音质量高的情况，该方法也不会使系统性能得到很大的提升。而对于发音语料有明显外部环境噪音的 Sp3，与未使用说话人编码的实验组比较，说话人编码实验组虽然整体较优，但是 MCD 值还是较高。该方法能够使系统的抗噪性能得到小幅提高。

Table 3. Grading of mean opinion score
表 3. 平均主观意见评分标准

平均主观意见评分	等级	语音满意度
5.00	优	非常完美
4.00	良	比较自然
3.00	中	可以接受
2.00	差	不自然
1.00	劣	不能接受

合成音频的主观评测采用平均主观意见评分(Mean Opinion Score, MOS)，评分标准见表 3。我们邀请了 10 位印尼语专业的评测人对合成的音频进行了评测，评测内容与 MCD 客观评测相同，每一位评测人对每组实验的每一个说话人风格测试集 20 句合成音频评测分数取均值作为该项的个人 MOS 评分结果，再将 10 位评测人的个人 MOS 评分结果取均值得到该项的最终 MOS 评分结果。本实验 MOS 评分见表 4。

Table 4. Evaluation results of MOS
表 4. MOS 评测结果

实验结果	MOS						
	Sp0	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
真实语音	4.55	4.41	4.51	4.36	4.49	4.38	4.42
实验 1	4.02						
实验 2	4.11						
实验 3	4.08	3.51	3.71	3.38			
实验 4	4.10	3.61	3.67	3.64			
实验 5	4.01	3.57	3.76	3.49			
实验 6	4.00	3.68	3.74	3.66			
实验 7	4.01	3.73	3.82	3.70	3.70	3.44	3.67
实验 8	4.10	3.91	3.99	3.74	3.83	3.50	3.79
实验 9	4.07	3.81	4.01	3.87	3.76	3.51	3.71
实验 10	4.12	3.94	4.08	3.86	3.91	3.66	3.81

根据 MOS 评测结果，本文设计的单一说话人印尼语语音合成系统，实验 1 和实验 2，分别取得了 4.02 和 4.11 的 MOS 评分，已经很接近真实语音。多说话人语音合成系统也在低资源风格迁移的实验中取得了令人满意的表现，实验 10 的 Sp0 说话人风格的 MOS 评分为 4.12，实验组最优，其他说话人风格的 MOS 评分在多说话人语音合成系统中总体最高。另外几组嵌入说话人编码的实验均取得了不错的 MOS 评分，整体优于未采用说话人编码的合成系统，验证了嵌入说话人编码方法的有效性。

4. 结束语

本文围绕印尼语语音合成系统，对 Tacotron2 原始的训练模式进行了改进，提出了渐变式交替训练方法，缓解了暴露偏差问题所带来的不利影响。在可获取的优质语料有限的情况下，该系统也能够合成出高质量语音，取得了 4.11 的 MOS 评分。而且渐变式交替训练方法也可应用于其他低资源语言的语音合成系统。本文还设计了说话人编码并应用于 GST-Tacotron2 以此实现了多说话人语音合成系统，同时也采用渐变式交替训练方法并添加预训练任务，在仅有少量说话人风格数据的前提下实现了数据增强以及说话人语音风格特征的迁移，Sp0 的 MOS 评分达到 4.12。通过对合成音频的客观评测与主观评测结果分析进一步证明了嵌入说话人编码能够有效地提高系统的风格提取能力，增强系统风格迁移的可迁移性。但该方法对于训练用的风格语料有一定程度上的依赖，如果所提供的语料存在明显的环境音或是该说话人并没有用同样的语调风格讲话以及存在其他影响音质的音素时(例如第二批数据集 Sp3)，就会对分离文本信息与风格信息造成影响，进而影响风格的提取。下一步将针对上述问题继续改进系统，增强系统的抗噪声性能，解决合成中预测停止符不正确的问题，进一步提高合成语音的质量。

基金项目

科技创新 2030“新一代人工智能”项目(2020AAA0107901)。

参考文献

[1] Wang, Y., Skerry-Ryan, R.J., Stanton, D., *et al*. (2017) Tacotron: Towards End-to-End Speech Synthesis. *Proceedings of INTERSPEECH*, **2017**, 4006-4010. https://doi.org/10.21437/Interspeech.2017-1452

[2] Shen, J., Pang, R., Weiss, R.J., *et al*. (2018) Natural TTS Synthesis by Conditioning WaveNet on Mel-Spectrogram Predictions. 2018 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Calgary, 15-20 April 2018, 4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

[3] Ren, Y., Ruan, Y., Tan, X., *et al*. (2019) Fastspeech: Fast, Robust and Controllable Text to Speech. *Advances in Neural Information Processing Systems* 32 (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 1171-1179.

[4] Ren, Y., Hu, C., Tan, X., *et al*. (2020) FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *International Conference on Learning Representations*, Addis Ababa, 26-30 April 2020, 1-15. https://openreview.net/forum?id=piLPYqxtWuA

[5] Arık, S.Ö., Chrzanowski, M., Coates, A., *et al*. (2017) Deep Voice: Real-Time Neural Text-to-Speech. *International Conference on Machine Learning*, Sydney, 6-11 August 2017, 195-204.

[6] Gibiansky, A., Arik, S., Diamos, G., *et al*. (2017) Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 2966-2974.

[7] Ping, W., Peng, K., Gibiansky, A., *et al*. (2018) Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. *International Conference on Learning Representations*, Vancouver, 30 April-3 May 2018, 214-217.

[8] van den Oord, A., Dieleman, S., Zen, H., *et al*. (2016) WaveNet: A Generative Model for Raw Audio. 9*th ISCA Speech Synthesis Workshop*, Sunnyvale, 13-15 September 2016, 125.

[9] Debnath, A., Patil, S.S., Nadiger, G., *et al*. (2020) Low-Resource End-to-End Sanskrit TTS Using Tacotron2, WaveGlow and Transfer Learning. 2020 *IEEE* 17*th India Council International Conference* (*INDICON*), New Delhi, 10-13 December 2020, 1-5. https://doi.org/10.1109/INDICON49873.2020.9342071

[10] Wang, Y., Stanton, D., Zhang, Y., *et al*. (2018) Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 5180-5189.

[11] Guo, H., Soong, F.K., He, L., *et al*. (2019) A New GAN-Based End-to-End TTS Training Algorithm. *Proceedings of Interspeech*, **2019**, 1288-1292. https://doi.org/10.21437/Interspeech.2019-2176

[12] Prenger, R., Valle, R. and Catanzaro, B. (2019) WaveGlow: A Flow-Based Generative Network for Speech Synthesis. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 3617-3621. https://doi.org/10.1109/ICASSP.2019.8683143

[13] Ito, K. and Johnson, L. (2017) The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset