用于自动配音的双向注意力机制联合多尺度跨语言说话风格迁移

Jingbei Li、Sipan Li、Ping Chen、Luwen Zhang、Yi Meng、Zhiyong Wu，*IEEE* 会员，Helen Meng，*IEEE* 会士，Qiao Tian、Yuping Wang、Yuxuan Wang

arXiv:2305.05203v2 [cs.SD] 5 Jul 2024

摘要——**自动配音能够生成与输入语音对应的另一种语言版本，可广泛应用于视频和游戏本地化等诸多现实场景。除了合成翻译后的脚本，自动配音还会将原语言中的说话风格迁移到配音语音中，使观众感受到角色仿佛在使用母语交谈。然而，当前最先进的自动配音系统仅对时长和语速的迁移进行建模，忽略了说话风格的其他方面，例如情绪、语调和重音；这些方面对于充分理解角色及其话语同样至关重要。本文提出一种联合多尺度跨语言说话风格迁移框架，同时对两种语言之间的双向说话风格迁移进行建模，涵盖全局尺度（即话语层面）和局部尺度（即词级别）。该框架提取每种语言中的全局和局部说话风格，并通过每个方向各自的编码器—解码器框架以及两个方向共享的双向注意力机制，利用这些风格预测另一种语言中的全局和局部说话风格。随后，采用多尺度说话风格增强的 FastSpeech 2，依据预测得到的各语言全局和局部说话风格合成目标语音。实验结果表明，所提出的框架效果显著，在客观和主观评估中均优于仅进行时长迁移的基线。**

索引词—**自动配音、跨语言说话风格迁移、多尺度说话风格迁移、双向注意力机制、文本到语音合成。**

稿件于2023年1月27日收到；于2023年8月19日和10月12日修订；于2023年11月3日接受。此项工作部分受中国国家自然科学基金（项目编号：62076144）、深圳市科技创新委员会（项目编号：WDZC20220816140515001）、AMiner、深圳市 SciBrain 基金以及深圳市下一代交互媒体创新技术重点实验室（项目编号：ZDSYS20210623092001004）资助。负责协调本稿件评审并批准其发表的副编辑为 Joseph Keshet 教授。（*Jingbei Li* 和 *Sipan Li* 对本项工作贡献相同。）（通讯作者：*Zhiyong Wu*。）

Jingbei Li、Sipan Li、Ping Chen、Luwen Zhang、Yi Meng 和 Zhiyong Wu 任职于清华大学深圳国际研究生院清华—香港中文大学媒体科学、技术与系统联合研究中心，中国深圳 518055（电子邮箱：lijb19@mails.tsinghua.edu.cn；lisp20@mails.tsinghua.edu.cn；p-chen21@mails.tsinghua.edu.cn；zlw20@mails.tsinghua.edu.cn；my20@mails.tsinghua.edu.cn；zywu@sz.tsinghua.edu. cn）。

Helen Meng 任职于中国香港特别行政区香港中文大学系统工程与工程管理系（电子邮箱：hmmeng@se.cuhk.edu.hk）。

Qiao Tian、Yuping Wang 和 Yuxuan Wang 就职于字节跳动，中国上海。（电子邮箱：tianqiao.wave@bytedance.com；wangyuping@bytedance.com；wangyuxuan.11@bytedance.com）。

数字对象唯一标识符 10.1109/TASLP.2023.3331813

I. 引言

TEXT-TO-SPEECH（TTS）合成旨在根据给定文本 [1], 生成自然语音，已在众多现实应用中得到广泛研究和应用。随着语音处理和神经网络技术的发展，先进的 TTS系统 [2]–[6] 通过将传统 TTS 系统中的不同组件整合到一个可训练框架中，实现了端到端语音合成；该框架能够合成高质量语音，同时减少对特征工程的需求。这类基于端到端神经网络的 TTS 模型已成功应用于多种语言[7]，并被广泛用作许多 TTS 相关任务的基础，例如情感语音合成 [8]，[9]、说话风格迁移 [10]–[12]、跨语言文本到语音系统[13]–[15] 以及自动配音 [16]。

在这些任务中，自动配音 [16], 将电影、电视或游戏中的语音转换为另一种语言的对应版本，是一个具有潜力的研究领域，可用于电影和游戏本地化，从而简化繁重的人工配音流程。与标准的 TTS 合成 [2]–[6] 直接根据给定文本合成语音不同，自动配音还需要进一步考虑语音中的说话风格，并在合成过程中将其迁移到配音语音中。在电影或游戏制作中，演员会运用各种说话风格来展现不同角色的个性、情感和意图，包括全局尺度（即话语层面）和局部尺度（即词级别）的说话风格。全局说话风格 [10] 用于控制每段话语的全局情感和意图，而局部说话风格 [17] 则用于控制每个词的重音和语调，以突出话语中的细节。这两种说话风格对于语音理解都至关重要，因此自动配音应将它们迁移过去，让观众觉得角色是在用自己的母语说话 [18]。

最先进的自动配音系统 [19] 已成功实现韵律短语级别的说话风格迁移。这些研究采用基于注意力机制[16] 或隐马尔可夫模型（HMMs） [20], [21]的算法，对齐两种语言中的韵律短语。随后，提出了多种韵律短语对齐方法 [16] ，用于调整

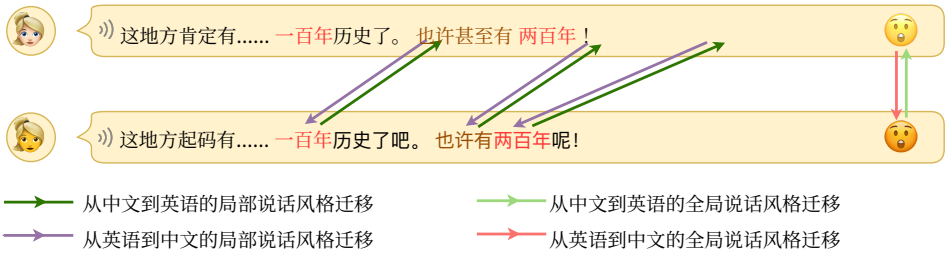


图 1. 两种语言之间在多个尺度上的跨语言说话风格迁移。

每个韵律短语 [23]的时长 [19]，、语速或说话风格 [22]，以匹配原语言语音中相应韵律短语的时长。

然而，当前的自动配音系统通常只适用于有限范围的简单配音场景，对多尺度说话风格迁移的建模也不充分。这些方法通常要求配音语言中的韵律短语数量与源语言中的韵律短语数量一致。对于复杂场景，例如中文与英语之间的配音，韵律短语数量往往不同，这些方法可能会合成不恰当的停顿和词语时长。此外，这些方法只能根据另一种语言中相同位置的韵律短语来调整每个韵律短语的说话风格，而不考虑它们的语义是否相同。当原文与配音语言中的韵律短语顺序差异显著时，这种设计可能导致不恰当的词语时长和语速。此外，现有方法也很少考虑全局尺度上的说话风格迁移建模。对于意译的话语，全局说话风格的迁移至关重要，因为两种语言中每个词的含义完全不同。对局部和全局说话风格迁移建模不足，会使合成语音缺少特定的情感和意图，并显著降低配音语音的表现力。

本文针对这些问题，旨在改进自动配音中的说话风格迁移，提出了一种多尺度跨语言说话风格迁移框架。该框架通过多任务学习，联合建模并优化两种语言之间两个配音方向的说话风格迁移。我们提取并利用每种语言语音的全局和局部说话风格，以预测另一种语言中相应的全局和局部说话风格。随后，通过适用于每种语言的相应多尺度说话风格增强型 FastSpeech 2（以下简称 MST-FastSpeech 2） [5], [17] ，将预测得到的另一种语言的全局和局部说话风格合成为语音。我们的工作源代码也已发布于 GitHub[1]。

我们使用预测的梅尔频谱图与真实梅尔频谱图之间的均方误差（MSE），以及自然度和说话风格迁移的平均选项得分（MOS）来评估我们方法的有效性。我们采用一种同样联合建模两个配音方向、但仅迁移词时长而非全局和局部说话风格的方法作为基线。实验结果表明，所提出的方法在客观评估和主观评估中均优于基线方法。

本文其余部分的组织结构如下：第二节介绍相关工作。随后，我们在第三节进行数据观察。第四节介绍所提出方法的详细内容。第五节给出实验结果和消融研究。第六节讨论我们的工作。最后，第七节总结全文。

II. 相关工作

A. 文本到语音合成

尽管数十年来一直在进行研究，文本到语音合成仍然是一项颇具挑战性的任务 [1]。传统TTS系统通常采用复杂的流水线 [24]，，包括语言学特征生成、时长预测、声学特征预测和声码器 [25], [26]。这些组件分别完成传统TTS流水线中的各项子任务，但它们是独立开发和训练的，因此容易在各个步骤中产生误差累积 [2], [27]。

随着深度学习的发展，研究人员提出了端到端TTS技术，将传统TTS系统 [2], [3] 中的各个组件整合到一个可训练的框架中。Tacotron [2], [3]和FastSpeech [5],等端到端TTS系统得益于联合训练，能够减轻误差累积，并以较少的特征工程提供更稳健的合成语音。这些端到端TTS系统采用编码器-解码器结构 [28]和注意力机制 [29]，：编码器用于生成语言学编码，解码器则直接生成原始频谱图。解码器中采用了 [30] 中的位置敏感注意力等注意力机制，通过学习文本与频谱图之间的对齐关系，提升性能和可解释性。随后，神经声码器根据频谱图生成波形，进一步提升合成语音的整体自然度和质量 [3]。

端到端TTS系统的改进为许多应用带来了越来越多的机遇，例如有声书朗读、新闻播报员和对话助手；这也促使端到端TTS系统提升其可控性，以便为不同应用场景选择合适的说话风格。全局风格令牌 [10]被提出，

[1]https://github.com/thuhcsi/StyleDub

用于将贯穿整个话语、与说话风格相关的属性（例如情感和意图）表示为单个向量，并已广泛应用于稳健的情感迁移和全局说话风格迁移 [10]。此外，还提出了一种对抗分支[31] ，用于从全局风格令牌中解耦说话人信息，以实现跨说话人的全局说话风格迁移 [10]。在词级别，提出了局部风格令牌 [17]，通过基于神经网络的强制对齐器 [32]，表示与每个词的说话风格相关的属性，例如重音、节奏和语调。

B. 自动配音

配音是一个复杂的过程，旨在将电影或游戏中的语音翻译成另一种语言，同时保留原有的说话风格，可视为一种特殊的语音到语音翻译 [18], [33]–[36]。传统上，配音由专业团队和配音演员完成，配音演员会在话语和词级别模仿原始语音的基频、能量及其他与说话风格相关的特征 [18]。然而，尽管投入了大量人力，一些配音作品仍因译文中的说话风格表现不尽如人意而受到观众争议。

随着 TTS 合成技术的发展，神经网络自动配音有望基于端到端 TTS 合成技术实现，从而简化并规范影视及游戏制作中的配音流程。目前已有若干自动配音系统被提出，用于将时长和语速迁移至配音语音。已有研究提出了一种基于韵律对齐的自动配音系统 [16]，用于迁移各个韵律短语的时长。该方法采用韵律对齐算法，根据预训练机器翻译模型提取的注意力权重，获取两种语言中韵律短语之间的映射关系。随后，算法调整每个韵律短语的时长，使其与原语言中对应映射韵律短语的时长相同。然而，当两种语言中相互映射的韵律短语之间的音素数量相差较大时，这种设计可能会生成过长或过短的时长。

此外，我们提出了一种基于 HMM [37] 的韵律对齐算法 [38] ，通过最大化配音语音中停顿位置的概率，将配音语音切分为与原语音停顿相匹配的韵律短语。我们还提出了对这种基于 HMM 的韵律对齐算法的改进 [23] ，使其能够同时迁移语速。我们进一步提出了一种额外的音素级时长预测器 [19]，以进一步改善韵律对齐和后续 TTS 系统中的时长与语速预测。然而，这些方法在切分韵律短语时忽略了每个词的语义。因此，原文与对应切分出的韵律短语很可能在语义上完全不同。在这种情况下，原语言中的时长和语速会被强行迁移到目标语言中语义无关的韵律短语这可能会大幅降低合成语音的自然度。

C. 贡献

与以往工作相比，我们的贡献如下：(1) 据我们所知，这是首项将多尺度说话风格迁移引入自动配音的工作。我们主张在全局尺度和局部尺度上对说话风格迁移进行建模，以显著提升自动配音的表现力；(2) 这也是首项采用多任务学习、联合优化两种语言之间双向跨语言说话风格迁移建模的工作；(3) 我们提出了一种有效的方法，用于推断自动配音在全局尺度和局部尺度上的说话风格。

III. 数据观察

我们对真实专业配音中的说话风格迁移进行主观和客观观察分析。我们随机收集了游戏《style id='1'>无主之地3 英语和中文配音中的25对平行话语用于观察。该游戏广为人知，其多种语言版本的配音均十分出色且广受好评。

A. 主观观察

在主观观察中，我们邀请了 25 位听众聆听每对所收集话语的英语和中文语音，并回答以下问题：(1) 每个话语中是否存在一种不同于陈述句正常说话风格的特殊 全局说话风格？(2) 这对话语在两种语言中的全局说话风格是相同、相似、无关还是相反？(3) 每个话语中的词语是否存在一种不同于陈述句中词语正常说话风格的特殊 局部说话风格？(4) 如果一个词在另一种语言中有一个含义和局部说话风格相同或相似的对应词，我们就说这个词可以与另一种语言中的相应词匹配。那么，对于每种语言，大约有多少个词可以与另一种语言中的词匹配?

前两个问题表明，在自动配音中对全局说话风格迁移进行建模既有必要性，也有可能性。汇总听众的回答（取平均值）后发现，英语和中文配音话语中分别有 76.1% 和 78.1% 的话语，其全局说话风格不同于常见陈述句的说话风格。这一发现表明，配音中普遍采用各种全局说话风格，因此应当予以考虑，并迁移到其他语言中。此外，汇总后的回答还表明，在这两种语言的平行话语对中，有 96.8% 的话语对，其全局说话风格相同或相似。这一发现证明了在自动配音中对全局说话风格迁移进行建模是可行的。

相比之下，接下来的两个问题旨在说明在局部尺度上建模说话风格迁移的必要性和可行性

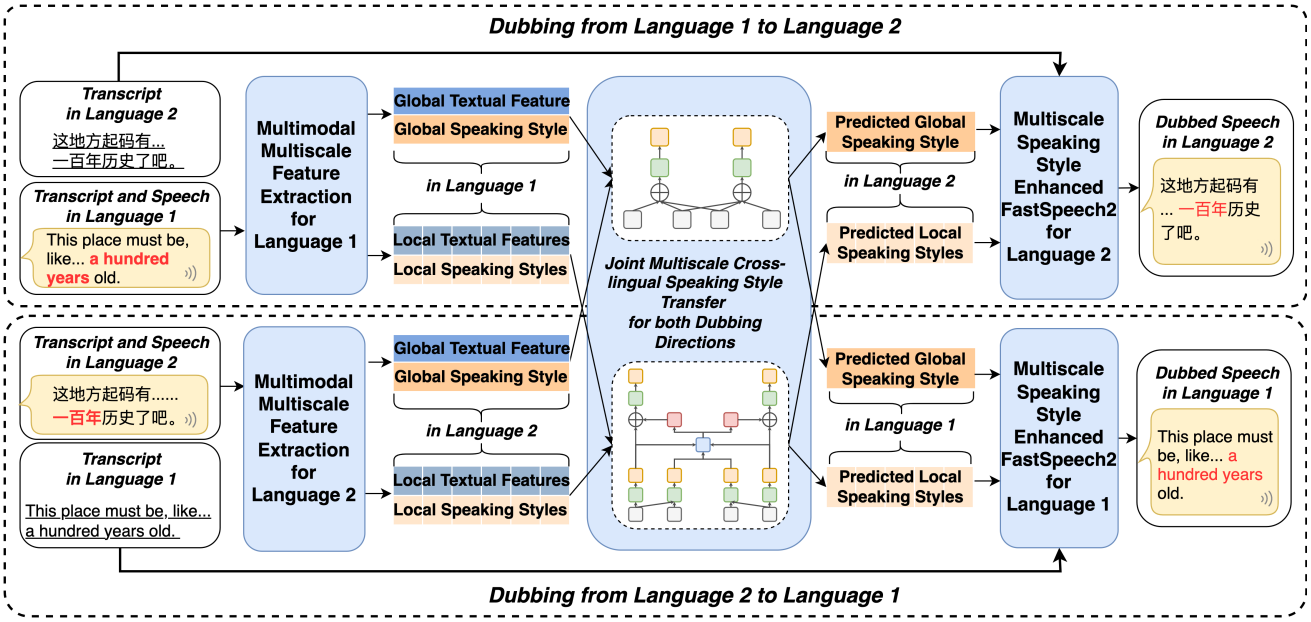


图 2. 所提出的联合多尺度跨语言说话风格迁移框架的架构。

表 I 主观观察中的汇总响应。

问题	英语	中文
是否具有特殊的全局说话风格?	76.1%	78.1%
全局说话风格是否相似?	96.8%	
是否具有任何特殊的 局部说话风格?	83.1%	83.7%
匹配 词语的百分比?	86。7%	87.1%

。对第三个问题的汇总回答显示，英语和中文配音中的话语分别有 83.1% 和 83.7% 包含具有不同于普通陈述句说话风格的局部说话风格的词语，这反映了为自动配音建模局部说话风格迁移的重要性。在最后一个问题中，汇总回答显示，英语配音中有 86.7% 的词语可以匹配 到中文中的词语，这意味着它们具有相似的语义信息和局部说话风格。反过来，中文配音中有 87.1% 的词语也可以匹配到英语中的词语。这表明，不同语言的平行话语之间，语义的相似性与局部说话风格的相似性相关，并展示了跨语言建模局部说话风格迁移的可行性。

我们在表 I 中总结了汇总后的回答。上述观察表明，在自动配音中考虑全局和局部说话风格的必要性（问题 1 和 3），以及在全局尺度（问题 2）和局部尺度（问题 4）上对话语之间的说话风格迁移进行建模的可能性。

B. 客观观察

我们参考了对人工配音过程进行大规模分析的实验 [18]，以进行客观

表 II 两种语言中平行话语及匹配词的各项音频属性的皮尔逊相关系数。

属性	话语层面	词级别
语速	0.782	0.482
基频均值	0.850	0.623
基频标准差	0.837	0.245
能量均值	0.621	0.414
能量标准差	0.691	0.261

观察。结果总结于表 II 中。

首先，我们计算了这些话语对之间语速、基频和能量的均值与标准差的皮尔逊相关系数。结果表明，这些平行配音在话语层面的这些属性上存在相关性。其次，我们计算了两种语言中明显匹配的话语之间相同的皮尔逊相关系数。每个词的时长由蒙特利尔强制对齐器 [39] 得出。结果显示，其皮尔逊相关系数低于话语层面的相关系数，因为这些词在英语和中文中的发音完全不同。不过，这些匹配 的词在这些属性上仍呈正相关。

在话语层面和词级别发现的模式与大规模分析 [18] 中的模式相似。这些模式表明，人工配音员会在全局尺度和局部尺度上模仿源音频的属性，也说明了自动配音中进行多尺度说话风格迁移的必要性。

IV. 方法论

基于数据观察，我们提出了一种用于自动配音的多尺度跨语言说话风格迁移框架，如图 2 所示，该框架联合学习了

两种语言之间在全局尺度和局部尺度上的说话风格迁移。

A. 多模态多尺度特征提取

为支持多尺度跨语言说话风格迁移的建模，我们在全局尺度和局部尺度上，分别提取文本特征，以及融合两种语言的文本信息和说话风格信息的多模态特征。

1) 文本特征提取：针对每种语言，采用一个预训练的基于变换器的双向编码器表示（BERT）[40] 模型，提取句子级和词级BERT嵌入。句子级BERT嵌入直接作为每种语言在全局尺度上的文本特征。词级BERT嵌入则通过每种语言的局部文本编码器转换为局部尺度上的文本特征。局部文本编码器由Tacotron [2]中的预网络和CBHG网络组成。上述过程可表示为：

$$g_1^t = SBERT_1 \tag{1}$$
$$g_2^t = SBERT_2 \tag{2}$$
$$s_1^t = E_1^t\left(BERT_1\right) \tag{3}$$
$$s_2^t = E_2^t\left(BERT_2\right) \tag{4}$$

其中，$SBERT_1$、$SBERT_2$、g_1^t 和 $g_2^t \in \mathcal{R}^{d_b}$ 分别表示每种语言的句子级BERT嵌入和全局文本特征，$BERT_1 \in \mathcal{R}^{l_1 \times d_b}$ 和 $BERT_2 \in \mathcal{R}^{l_2 \times d_b}$分别表示每种语言的词级BERT嵌入，$d_b$ 表示句子级和词级BERT嵌入的维度，l_1 和 l_2 分别表示两种话语中的词数，E_1^t 和 E_2^t 分别表示局部文本编码器，$s_2^t \in \mathcal{R}^{l_1 \times d_e}$ 和$s_1^t \in \mathcal{R}^{l_2 \times d_e}$ 分别表示每种语言的局部文本特征，d_e 表示局部文本编码器的输出维度。

2) 说话风格提取：随后，采用全局和局部说话风格编码器 [17]，提取每种语言在全局和局部尺度上的说话风格。全局说话风格编码器由 6 个步幅卷积神经网络（CNN）组成，每个网络包含 3×3 个卷积核，步幅为 2×2，滤波器数量分别为 32、32、64、64、128 和 128；此外，还包括一个 256 维 GRU层和一个 128 维风格注意力层。输入语音的梅尔频谱图由 CNN 和 GRU 处理。随后，将 GRU 的最终状态传入风格注意力层，得到全局说话风格向量 GST，作为 10 个自动学习的全局说话风格基嵌入的权重。该过程可表示为：

$$q = final(GRU(CNN(mel))) \tag{5}$$
$$GST = softmax\left(q^T GST^{table}\right) \tag{6}$$

其中，$final$ 返回 GRU 的最终状态，q 是风格注意力层的查询向量，GST^{table} 包含 10 个自动学习的全局说话风格基嵌入。

局部说话风格编码器的架构与全局说话风格编码器相同，但步幅现在为 1×2，且GRU层现在会返回每个输入帧的输出。随后，通过相乘对话语中的每个词对应的GRU输出进行汇总。

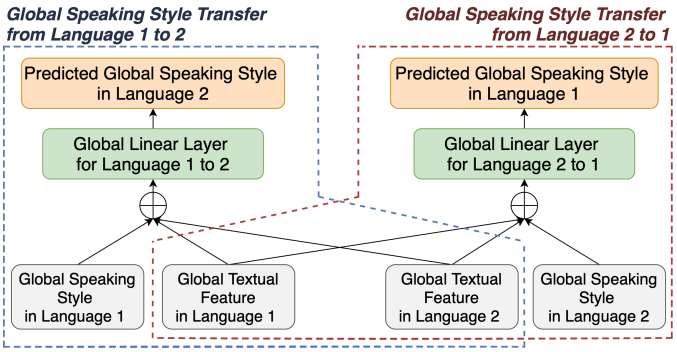


图 3. 全局尺度上的联合跨语言说话风格迁移。

将其与从预训练神经网络强制对齐器（NeuFA）中提取的语音到文本注意力权重[32]相结合。NeuFA 采用双向注意力机制[32]，以学习文本与语音之间的双向信息映射。沿语音到文本方向学习到的注意力权重可用于汇总话语中每个词的帧级信息，从而得到局部说话风格序列 LST。局部参考编码器的处理过程可表示为：

$$q' = W_{ASR}^T GRU(CNN(mel)) \tag{7}$$
$$LST = softmax\left(q'^T LST^{table}\right) \tag{8}$$

其中，W_{ASR} 是 NeuFA 在语音到文本方向上获得的注意力权重，q' 是局部风格注意力层的查询向量，LST^{table} 包含 10 个自动学习得到的局部说话风格基础嵌入。

3) 多模态特征融合：为了融合提取的文本特征和语音特征，随后将句子级 BERT 嵌入与每种语言的全局说话风格拼接，作为全局尺度上的多模态特征。将词级BERT嵌入与局部说话风格拼接，并输入每种语言对应的局部编码器，以生成局部尺度上的多模态特征序列。这里，局部编码器与局部文本编码器的架构相同。该过程可表示为：

$$g_1^m = [SBERT_1; GST_1] \tag{9}$$
$$g_2^m = [SBERT_2; GST_2] \tag{10}$$
$$s_1^m = E_1\left([BERT_1; LST_1]\right) \tag{11}$$
$$s_2^m = E_2\left([BERT_2; LST_2]\right) \tag{12}$$

其中，GST_1、$GST_2 \in \mathcal{R}^{d_{st}}$ 是每种语言的全局说话风格，$LST_1 \in \mathcal{R}^{l_1 \times d_{st}}$ 和 $LST_2 \in \mathcal{R}^{l_2 \times d_{st}}$是每种语言的局部说话风格，$E_1$ 和 E_2、$g_1^m \in \mathcal{R}^{d_b+d_{st}}$ 和 $g_2^m \in \mathcal{R}^{d_b+d_{st}}$、$s_1^m \in \mathcal{R}^{l_1 \times d_e}$ 和$s_2^m \in \mathcal{R}^{l_2 \times d_e}$ 分别是每种语言对应的局部编码器、全局多模态特征和局部多模态特征。

B. 联合多尺度跨语言说话风格迁移

1) 全局尺度上的说话风格迁移：如图 3 所示，将每种语言的全局多模态特征与另一种语言的全局文本特征拼接起来

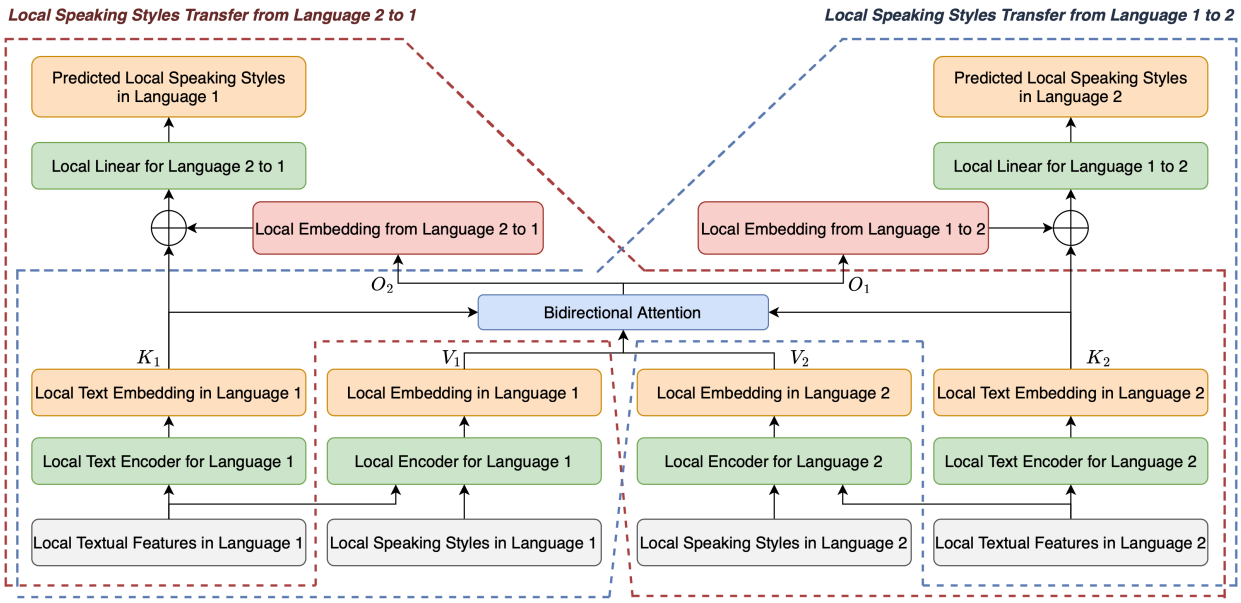


图 4. 局部尺度上的联合跨语言说话风格迁移。

并输入全局线性投影层，以预测另一种语言的全局说话风格：

$$GST_1' = f_{2to1}^g([g_2^m; g_1^t]) \tag{13}$$
$$GST_2' = f_{1to2}^g([g_1^m; g_2^t]) \tag{14}$$

其中，f_{2to1}^g和f_{1to2}^g、GST_1'和GST_2'分别表示各配音方向的全局线性投影层，以及各语言预测得到的全局说话风格。

2）局部尺度上的说话风格迁移：如图 4 所示，本文采用由两个编码器—解码器框架和一个共享的双向注意力机制组成的架构，学习两种语言之间局部尺度上的说话风格迁移。

双向注意力机制由传统注意力机制 [41]扩展而来，用于建立两组键值对之间的双向关系，如图 5 [32]所示。双向注意力机制以两组键值对作为输入，其中每组的键都作为另一组的查询。

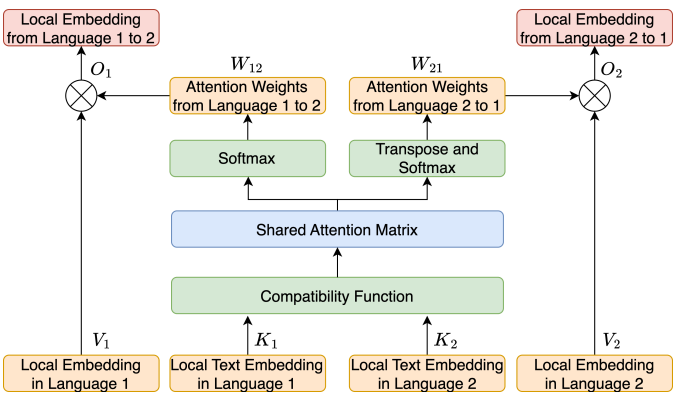


图 5. B局部说话风格迁移中的双向注意力机制 移。

键值对。随后，为两组键建立一个共享的兼容性函数，以计算两个方向各自的注意力权重。然后，利用注意力权重对两组值进行汇总，分别得到两个方向的输出。双向注意力机制的公式如下：

$$A = f(K_1, K_2) \tag{15}$$
$$W_{12} = softmax(A) \tag{16}$$
$$W_{21} = softmax(A^T) \tag{17}$$
$$O_1 = W_{12}^T V_1 \tag{18}$$
$$O_2 = W_{21}^T V_2 \tag{19}$$

其中，$K_1 \in R^{n_1 \times d_{k1}}$、$V_1 \in R^{n_1 \times d_{v1}}$ 和 $K_2 \in R^{n_2 \times d_{k2}}$、$V_2 \in R^{n_2 \times d_{v2}}$ 是两组键值对；n_1 和 n_2是键值对的数量；d_{k1}、d_{v1}、d_{k2} 和 d_{v2}是特征维度；f 是兼容性函数；$A \in R^{n_1 \times n_2}$ 是共享注意力矩阵；$W_{12} \in R^{n_1 \times n_2}$和 $W_{21} \in R^{n_2 \times n_1}$ 是两个方向的注意力权重；$O_1 \in R^{n_2 \times d_{v1}}$ 是 K_2中每个键对应的 V_1 加权和；$O_2 \in R^{n_1 \times d_{v2}}$ 是 K_1中每个键对应的 V_2 加权和。

我们采用双向注意力机制的乘法形式，在局部尺度上联合建模说话风格迁移，并汇总两个配音方向上的多模态特征。具体而言，将两种语言的局部文本特征和局部多模态特征分别作为两组键和值，注意力权重和输出的计算公式如下：

注意力权重和输出的公式如下：

$$A_{1to2} = f_1(s_1^t) \times f_2(s_2^t)^T \tag{20}$$
$$A_{2to1} = f_2(s_2^t) \times f_1(s_1^t)^T = A_{1to2}^T \tag{21}$$
$$W_{1to2} = softmax(A_{1to2}) \tag{22}$$
$$W_{2to1} = softmax(A_{2to1}) \tag{23}$$
$$O_{1to2} = W_{1to2}^T s_1^m \tag{24}$$
$$O_{2to1} = W_{2to1}^T s_2^m \tag{25}$$

其中， f_1 和 f_2 是两个线性投影， A_{1to2} 和 A_{2to1}是共享的注意力矩阵， W_{1to2} 和 W_{2to1} 是两个配音方向对应的学习得到的注意力权重，O_{1to2} 和 O_{2to1} 是每种语言的汇总多模态特征。

随后，将每种语言的汇总多模态特征与另一种语言的局部文本特征拼接，并输入局部线性投影层，以预测另一种语言的局部说话风格：

$$LST_1' = f_{2to1}^l([O_{2to1}; s_1^t]) \tag{26}$$
$$LST_2' = f_{1to2}^l([O_{1to2}; s_2^t]) \tag{27}$$

其中， f_{2to1}^l和 f_{1to2}^l、 LST_1'和 LST_2'分别是各个配音方向的局部线性投影，和是每种语言预测得到的局部说话风格。

C. 利用预测的多尺度说话风格进行语音合成

为每种语言训练一个MST-FastSpeech 2 [17]，以合成具有所提出框架推断出的全局和局部说话风格的语音。

每个话语的说话人嵌入以及预测得到的全局和局部说话风格均上采样至音素级别，并与 FastSpeech 2 的编码器输出进行拼接。随后，方差适配器根据拼接后的编码器输出推断每个音素的基频、时长和能量，并进一步由 FastSpeech 2 的解码器将其转换为梅尔频谱图。

预训练的 HiFi-GAN [26] 随后用作声码器，根据预测的梅尔频谱图合成语音，并迁移来自另一种语言的期望说话风格。

D. 训练策略

在训练所提出的联合跨语言多尺度说话风格迁移框架之前，需先训练每种语言的全局和局部说话风格编码器及其对应的 MST-FastSpeech 2，以确保提取的全局和局部说话风格与文本转语音主干网络兼容。为增强提取的说话风格及其跨语言迁移的鲁棒性，在此阶段还通过梯度反转层（GRL） [42]和域对抗学习 [31], [43]将文本和说话人信息与全局和局部说话风格解耦：

$$speaker'_{GST} = f_{speaker}^{GST}(GRL(GST)) \tag{28}$$
$$speaker'_{LST} = f_{speaker}^{LST}(GRL(LST)) \tag{29}$$
$$BERT' = f_{BERT}(GRL(LST)) \tag{30}$$

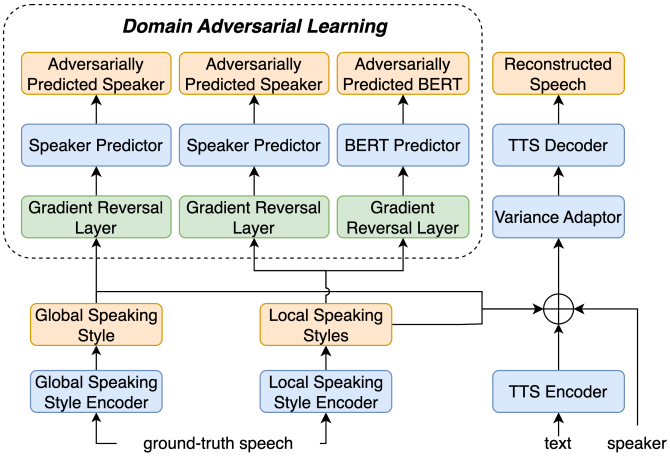


图 6. 针对每种语言 [17]预训练全局和局部说话风格编码器以及 FastSpeech 2。

其中， $f_{speaker}^{GST}$、 $f_{speaker}^{LST}$ 和 f_{BERT} 是用作对抗式说话人预测器和文本预测器的线性投影，GRL 对这些对抗式预测器的梯度进行反转，$speaker'_{GST}$ 和 $speaker'_{LST}$ 是通过对抗方式预测的说话人嵌入， $BERT'$ 是通过对抗方式预测的该话语的文本嵌入。预训练损失是重建损失与对抗损失之和，其中前者是预测的梅尔频谱图与真实梅尔频谱图之间的均方误差（MSE），后者是预测的说话人嵌入和 BERT 嵌入与真实说话人嵌入和 BERT 嵌入之间的均方误差。

随后，冻结每种语言的预训练全局和局部编码器以及 FastSpeech 2，并利用它们提取说话风格，以及在所提出的框架中使用预测的说话风格合成语音。然后，对所提出的联合多尺度跨语言说话风格迁移框架进行训练，并根据两种语言中预测的全局和局部说话风格与真实全局和局部说话风格之间的均方误差之和进行反向传播：

$$\begin{aligned} loss = & MSE(GST_1, GST_1') + MSE(LST_1, LST_1') \\ & + MSE(GST_2, GST_2') + MSE(LST_2, LST_2') \end{aligned} \tag{31}$$

V. 实验

A. 基线方法

为展示所提出的联合多尺度跨语言说话风格迁移框架的有效性，我们采用了两种使用不同说话风格迁移方法的方法作为基线，这些方法同样采用 FastSpeech 2 作为文本转语音主干网络。

1) 无说话风格迁移：第一种基线方法是一个不进行说话风格迁移的原版 FastSpeech 2 [5]，它也代表了最先进的传统文本转语音系统，以及使用理想翻译脚本的语音到语音翻译系统。

2) 时长迁移：基于所提出的方法，我们实现了另一种仅进行时长迁移的方法，作为第二种基线。该方法受到一种最先进的自动配音系统 [19]的启发，采用编码器—解码器框架和注意力机制时，其性能应相当或更优。在这种方法中，我们省略了全局说话风格迁移的建模，只将每个词的时长迁移到另一种语言的语音中。具体而言，我们用每个词的时长序列替代所提出方法中的局部说话风格。公式 (11)、(12)、(26) 和 (27) 改为：

$$s_1^m = E_1^l\left([BERT_1; d_1]\right) \tag{32}$$
$$s_2^m = E_2^l\left([BERT_2; d_2]\right) \tag{33}$$
$$d_1' = f_{2to1}^l([O_{2to1}; s_1^t]) \tag{34}$$
$$d_2' = f_{2to1}^l([O_{1to2}; s_2^t]) \tag{35}$$

其中，d_1、d_2、d_1' 和 d_2' 分别表示每种语言中每个词的时长序列和预测时长序列。然后，FastSpeech 2 利用每种语言的预测时长序列进行精确的时长控制，从而合成配音语音。

B. 训练设置

我们从游戏《无主之地3》的英语和中文配音中收集了更多平行话语，作为本研究的语料库。如第二节所述，游戏《无主之地3》是一款以太空西部科幻背景为设定的动作角色扮演、第一人称射击电子游戏，隶属《无主之地》游戏系列。该游戏提供了多种语言的出色配音，广受好评，且包含多种说话风格。《无主之地3》有4名可玩角色和数十名非玩家角色（NPC），其中约有20名主要非玩家角色，他们经常相互交谈，讲述背景故事或引导玩家完成任务。每个可玩角色和非玩家角色都有独特的说话风格，以展现不同的个性。此外，在35小时的主线任务和其他支线任务游戏时长中，角色还会根据不同场景改变说话风格。

我们从*Bor-derlands 3*中收集了 4,914 对平行话语，包括 5 小时英语语音、5 小时中文语音及其对应字幕。语料库中所有语音的采样率均为 48,000 Hz。语料库中英语语音的每条话语时长介于 0.21 s 至 12.63 s 之间，平均时长为 3.80 s，标准差为 2.35 s。每条话语的词数介于 1 至 46 之间，平均为 11.40，标准差为 7.72。语料库中中文语音的每条话语时长介于 0.31 s 至 12.63 s 之间，平均时长为 3.91 s，标准差为 2.44 s。每条话语的词数介于 1 至 70 之间，平均为 17.36，标准差为 11.92。

我们使用 25 ms 的窗口长度和 10 ms 的帧移，为语料库中的每个话语提取梅尔频谱图，这些梅尔频谱图将作为全局和局部说话风格编码器的输入，以及 MST-FastSpeech 的真实值

2. 我们采用一个预训练的 BERT [44]模型来提取句子级别和词级别的 BERT嵌入，其中每个词的 BERT嵌入是其对应子词的 BERT嵌入的平均值。

随后，我们使用整个语料库对 MST- FastSpeech 2 以及两种语言的全局和局部说话风格编码器进行预训练。我们遵循 Fast- Speech 2 [5] 的训练设置，以批大小为 32 对每种语言的模型训练 500,000 次迭代。

我们随机选取两种语言中的 4,414 对平行话语作为训练集，以训练所提出的多尺度说话风格迁移框架。该框架训练 10 个 epoch，批量大小为 32，学习率为 10^{-4}。

所有模型均使用 PyTorch [45] 实现，并在 NVIDIA TeslaV100 GPU 上训练。

C. 评估

我们将剩余的 500 对话语用作测试集进行评估。我们采用全频带预测梅尔频谱图与真实梅尔频谱图之间的均方误差（MSE）作为客观评估的指标。我们还计算梅尔频谱图最低 10 个维度与最高 10 个维度之间的均方误差，以研究说话风格如何在全局尺度和局部尺度上传递。我们使用最近邻插值调整预测梅尔频谱图的大小，使其长度与真实梅尔频谱图匹配。我们采用最近邻插值，而不是动态时间规整（DTW），因为动态时间规整会显著影响每个单词的时长，而时长也是说话风格中的一个重要属性。另一方面，最近邻插值不会对时长产生重大影响，因为每种方法中整个句子的时长始终较为接近。

在主观评估中，我们进一步随机选取了 20 个对话片段[2]，并由 25 位听众进行评估。听众需要按照 1 到 5 分的量表（间隔为 1 分），评价合成语音的自然度，以及从另一种语言的语音中迁移说话风格的效果，并据此计算主观平均意见分（MOS）。听众还需要从不同方法生成的配音语音中选出偏好的一种，并据此计算偏好率。

D. 实验结果

客观评估和主观评估的结果见表 III 和表 IV。在客观评估中，采用时长迁移的方法和所提出的多尺度说话风格迁移方法，均优于无说话风格迁移的基线。此外，所提出的方法在全频带、高频带和低频带的梅尔频谱图 MSE 上，进一步显著优于采用时长迁移的基线。在主观评估中，与无说话风格迁移的基线相比，采用时长迁移的方法在说话风格迁移方面的平均意见分在两个方向上分别提高了 0.762 和 0.634

[2]合成样本位于 https://thuhcsi.github.io/StyleD ub/.

表 III 全频带及最高和最低 10 个维度的梅尔频谱图均方误差。

说话风格迁移方法	方向	均方误差（梅尔）	均方误差（高维）	均方误差（低维）
无说话风格迁移 [5]	英语-中文	4.694	3.756	4.230
时长迁移 [19]		3.695	3.079	3.133
多尺度说话风格迁移（本文提出）		**1.392**	**1.090**	**1.293**
无说话风格迁移 [5]	中文-英语	5.385	4.305	4.627
时长迁移 [19]		3.301	2.697	2.593
多尺度说话风格迁移（所提出的方法）		**2.140**	**1.623**	**2.080**

表 IV 不同方法的主观评价。

说话风格迁移方法	方向	MOS（自然度）	MOS（迁移效果）	偏好率
无说话风格迁移 [5]	英语-中文	2.761 ± 0.097	3.161 ± 0.075	7.75%
时长迁移 [19]		3.562 ± 0.099	3.923 ± 0.073	26.10%
多尺度说话风格迁移（所提方法）		$\mathbf{3.788 \pm 0.088}$	$\mathbf{4.123 \pm 0.071}$	**66。15**%
无说话风格迁移 [5]	中文-英语	3.473 ± 0.090	3.358 ± 0.077	14.95%
时长迁移 [19]		3.615 ± 0.093	3.992 ± 0.069	40.72%
多尺度说话风格迁移（所提出的方法）		3.626 ± 0.097	4.005 ± 0.073	**44。33**%

表V 两种方向上的多尺度说话风格迁移与多任务学习消融研究。

方法	方向	均方误差（GST）	均方误差（LST）	均方误差（梅尔）	均方误差（高维）	均方误差（低维）
所提方法	英语-中文	4.220×10^{-3}	7.983×10^{-3}	**1.392**	**1.090**	1.293
-不包含局部尺度		3.274×10^{-3}	-	1.675	1.545	1.333
-不包含全局尺度		-	7.373×10^{-3}	1.660	1.291	1.536
-不包含中文-英语方向		4.228×10^{-3}	8.097×10^{-3}	1.402	1.135	**1.270**
所提方法	中文-英语	2.012×10^{-3}	1.075×10^{-2}	**2.140**	**1.623**	**2.080**
-不使用局部尺度		1.401×10^{-2}	-	2.699	2.678	2.184
-不使用全局尺度		-	1.092×10^{-2}	2.666	2.053	2.533
-不使用英语-中文方向		2.024×10^{-3}	1.080×10^{-2}	2.368	1.804	2.272

，相应的偏好率分别提高了 18.35% 和 25.77%。与采用时长迁移的基线相比，所提出的方法在说话风格迁移方面的平均意见分在两个方向上进一步提高了 0.2 和 0.013，相应的偏好率分别提高了 40.05% 和 3.61%。这些性能提升表明，对说话风格的各个方面进行迁移建模，而不仅仅是时长，既有必要也有效。

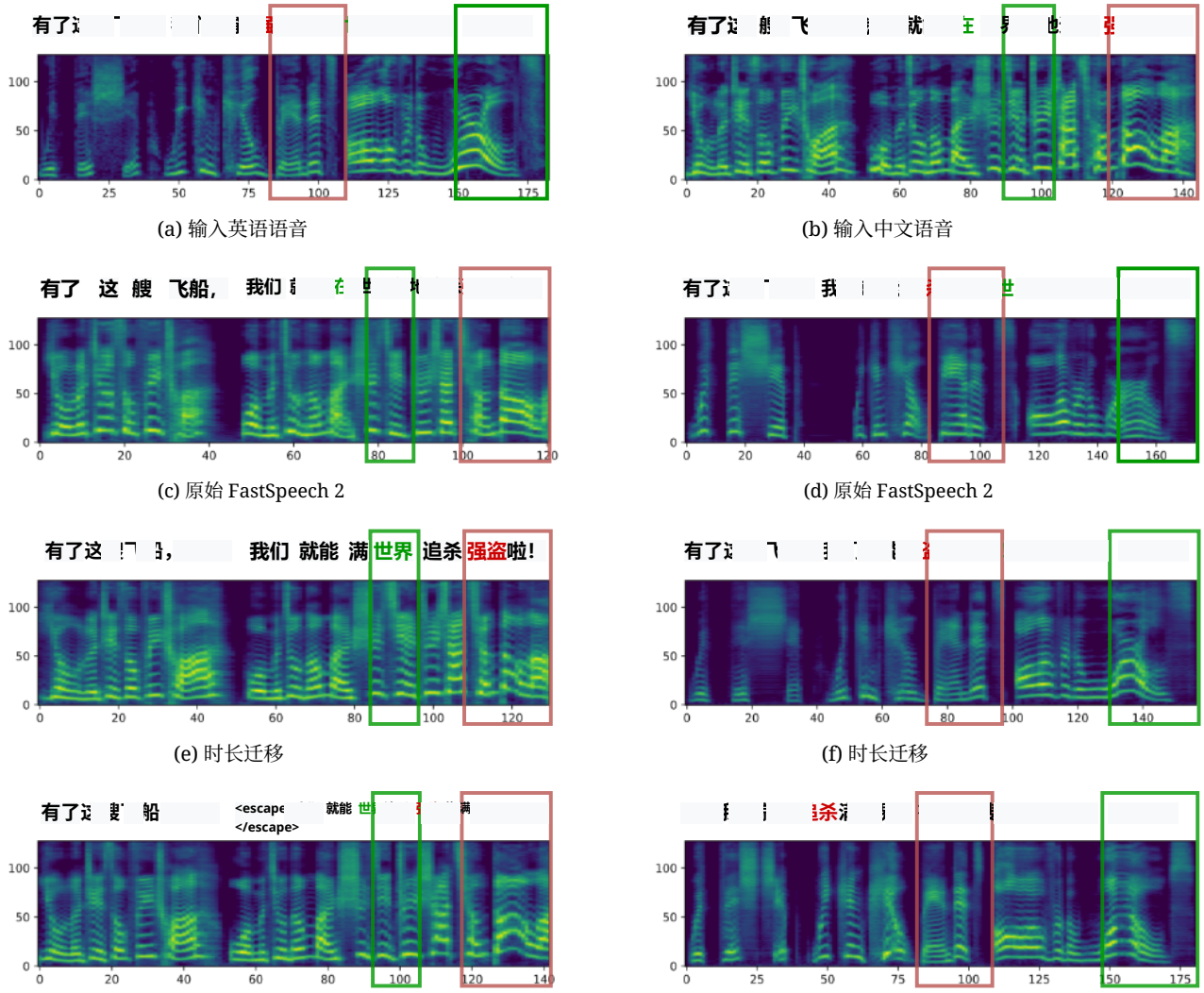
需要注意的是，尽管所提出的模型在英语到中文的配音方向上，效果显著优于采用时长迁移的基线方法，但从中文配音到英语配音时，提升幅度却微乎其微。这种差异可能是因为中文配音中的说话风格不如英语配音丰富。在我们的主观观察中，大多数听众表示，英语配音比中文配音更有表现力，说话风格也更加多样。由于英语配音是该游戏的原始配音，其他语言的配音过程中可能损失了部分说话风格。此外，配音演员在制作英语配音时，通常会为不同角色使用各种口音；而中文配音则始终采用标准普通话，不带任何口音。这种差异也可能导致难以将中文中的说话风格映射到英语中。

我们在图7中展示了各方法合成语音的梅尔频谱图，所用话语来自测试集：“With this ship, we can kill bandits all over the world”。我们用蓝色和红色边框标出这些话语中“bandits”和“world”两个词的对比。在图 7（a）和（b）中，原始英语配音和中文配音中的这两个词通过局部说话风格得到强调，以表现角色获得新飞船时的兴奋之情。如图7（g）和（h）所示，与原始 FastSpeech 2 和仅采用时长迁移的基线方法合成的语音相比，另一种语言中对应的词也得到了更鲜明的强调。

E. 消融研究

基于所提出的方法，我们实现了几种变体方法，用于消融实验。实验结果如表V所示。除了计算两种语言中预测的梅尔频谱图与真实梅尔频谱图之间的均方误差（MSE）外，我们还在消融实验中进一步考虑了预测的全局和局部说话风格与真实全局和局部说话风格之间的均方误差（MSE）。

前两个变体忽略了全局或局部尺度上的说话风格迁移。这些变体的文本转语音骨干网络也排除了对全局或局部说话风格的建模。



(a) 输入英语语音 (b) 输入中文语音

(c) 原始 FastSpeech 2 (d) 原始 FastSpeech 2

(e) 时长迁移 (f) 时长迁移

(g) 多尺度说话风格迁移（所提方法） (h) 多尺度说话风格迁移（所提方法）

图 7. 各方法在英语到中文（左列）和中文到英语（右列）配音方向下，合成语音“With this ship, we can kill bandits all over the world”中单词“bandits”和“worlds”的梅尔频谱图对比。

说话风格。这两种方法在说话风格迁移中都缺少某些尺度，因此其梅尔频谱图的 MSE 高于所提出的方法。具体而言，仅迁移局部说话风格的方法优于仅迁移全局说话风格的方法。这一发现证实了局部说话风格在自动配音中更为重要这一观点。然而，仅迁移局部说话风格无法处理意译以及在另一种语言中无法匹配的词语。所提出的方法通过联合建模全局和局部说话风格迁移，利用局部说话风格处理匹配的词语，并利用全局说话风格处理意译，从而获得了最低的梅尔频谱图 MSE。请注意，尽管这两个变体的全局或局部说话风格 MSE 低于所提出的方法，但由于它们是从不同的文本转语音骨干网络中提取的，因此无法进行比较。

最后两种变体在说话风格迁移过程中各自忽略了一个配音方向。所提出的方法略微优于这两种变体，表明对两个配音方向进行说话风格迁移联合建模是有效的。值得注意的是，与相反方向相比，中文到英语的配音方向似乎从联合建模中获益更多，因此梅尔频谱图的 MSE 降幅更大。这一发现也可能与中文和英语配音中说话风格密度的差异有关，并凸显了两个方向进行联合训练的潜在有效性，尤其是在源语音中的说话风格密度低于目标语音时。

VI. 讨论

自动配音是 TTS 领域一个具有潜力的研究方向，可用于简化现实游戏和影视制作中的配音流程。尽管已有一些研究致力于自动配音，并将每个韵律短语的时长迁移到配音语言中，但此前的工作尚未充分考虑并迁移自动配音中的说话风格

。本文首次提出联合多尺度跨语言说话风格迁移框架，以同时建模两种语言之间在全局尺度和局部尺度上的双向说话风格迁移。大量实验表明，所提出的框架能够联合考虑源语言中每个词的说话风格以及整个话语的说话风格，并在目标语言中合成具有恰当说话风格的语音。

收集意义和说话风格都相互对应的不同语言语音十分困难，这或许可以解释为何自动配音中尚无跨语言说话风格迁移建模方法。为应对这一挑战，我们收集了电子游戏配音中的平行语音。这些语音由经验丰富的配音演员演绎，并使用专业设备录制，因此非常适合用于学习不同语言间的说话风格迁移模式。然而，由于这些语音只能用于私人用途，复现我们的工作可能会受到限制。我们呼吁游戏公司公开发布不同语言版本的配音，以支持未来的自动配音研究。

为了证明自动配音中对多尺度跨语言说话风格迁移的需求，我们对所收集游戏语料库中的跨语言说话风格迁移进行了主观观察。观察结果显示，不同语言的平行话语之间，说话风格存在很强的相关性。具体而言，在话语层面，这些平行话语的全局说话风格始终相同或相似。在词级别，每个词的局部说话风格往往可以与另一种语言中的某个词匹配，这意味着这些词在语义信息和局部说话风格方面都相同或相似。

根据我们的观察，我们提出了一种新颖的框架，用于联合建模全局尺度和局部尺度上的跨语言说话风格迁移。所提出的框架利用源语言的全局多模态特征和配音语言的全局文本特征，预测配音语言的全局说话风格。在局部尺度上，我们采用双向注意力机制，根据两种语言中词语语义之间学习到的注意力权重，将多模态信息从源语言迁移到配音语言。此外，我们使用多任务学习统一两个配音方向上的跨语言说话风格迁移学习，并联合优化局部尺度上的注意力权重。尽管所提出的框架最初是针对我们的游戏语料库设计和优化的，但只要配音过程仍符合在游戏语料库中观察到的现象，该框架也可以普遍应用于其他配音场景，例如有声书配音。

然而，所提出框架的有效性可能会受到语料库规模的限制。我们研究所收集的语料库相当小，并且我们相信，使用更大的语料库，我们的设计能够取得更好的性能。当配音过程中存在额外要求时，所提出的框架可能无法良好运行，例如，以匹配相应视频中角色的面部动作。

VII. 结论

本文旨在自动配音中恰当地迁移全局尺度（即话语层面）和局部尺度（即词级别）的说话风格。为此，我们提出了一种联合多尺度跨语言说话风格迁移框架，通过多任务学习对两种语言之间两个尺度上的说话风格迁移进行建模。我们提取每种语言中的全局和局部说话风格，并利用它们预测另一种语言中相应的全局和局部说话风格，随后通过 MST-FastSpeech 2 将其合成为语音。实验和消融研究证明了所提框架的有效性。

参考文献

[1] P. Taylor，《文本到语音合成》。剑桥大学出版社，2009年。[2] Y. Wang、R. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio 等，“Tacotron：迈向端到端语音合成”，*arXiv*预印本 *arXiv:1703.10135*，2017年。[3] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，“通过以梅尔频谱图预测结果为条件的WaveNet实现自然文本到语音合成”，载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2018年，第4779–4783页。[4] N. Li、S. Liu、Y. Liu、S. Zhao和M. Liu，“基于Transformer网络的神经语音合成”，载于*AAAI*人工智能会议论文集，第33卷，第01期，2019年，第6706–6713页。[5] Y. Ren、C. Hu、X. Tan、T. Qin、S. Zhao、Z. Zhao和T.-Y. Liu，“FastSpeech 2：快速、高质量的端到端文本到语音”，载于国际学习表征会议，2020年9月。[在线]。网址：https://openreview.net/forum?id=piLPYqxtWuA [6] J. Kim、J. Kong和J. Son，“通过对抗学习实现端到端文本到语音的条件变分自编码器”，载于国际机器学习会议。PMLR，2021年，第5530–5540页。[7] J. Li、Z. Wu、R. Li、P. Zhi、S. Yang和H. Meng，“用于端到端普通话文本到语音合成的知识驱动语言编码”，载于*INTERSPEECH*，2019年，第4494–4498页。[8] X. Cai、D. Dai、Z. Wu、X. Li、J. Li和H. Meng，“借助跨领域语音情感识别，使用无情感标签数据集实现可控情感语音合成”，载于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第5734–5738页。[9] Y. Lei、S. Yang、X. Wang和L. Xie，“Msemotts：用于情感语音合成的多尺度情感迁移、预测与控制”，*IEEE/ACM*音频、语音与语言处理汇刊，第30卷，第853–864页，2022年。[10] Y. Wang、D. Stanton、Y. Zhang、R.-S. Ryan、E. Battenberg、J. Shor、Y. Xiao、Y. Jia、F. Ren和R. A. Saurous，“风格令牌：端到端语音合成中的无监督风格建模、控制与迁移”，载于国际机器学习会议。PMLR，2018年，第5180–5189页。[11] X. Li、C. Song、J. Li、Z. Wu、J. Jia和H. Meng，“迈向表现力语音合成的多尺度风格控制”，*arXiv*预印本 *arXiv:2104.03521*，2021年。[12] X. Li、C. Song、X. Wei、Z. Wu、J. Jia和H. Meng，“迈向有声读物数据集上的跨说话人朗读风格迁移”，载于*Interspeech 2022*。ISCA，2022年9月，第5528–5532页。 [在线]。网址：https:// www.isca-speech.org/archive/interspeech 2022/li22ca interspeech.html [13] L. Sun、H. Wang、S. Kang、K. Li和H. M. Meng，“使用音素后验图的个性化跨语言文本到语音合成”，载于*INTERSPEECH*，2016年，第322–326页。[14] Y.-J. Chen、T. Tu̇、C.-c. Yeḣ和H.-Y. Lee，“通过跨语言迁移学习为低资源语言实现端到端文本到语音”，载于*Interspeech 2019*。ISCA，2019年9月，第2075–2079页。[在线]。网址：https://www.isca-speech.org/archive/interspeech 2019/chen19f interspeech.html [15] M. Chen、M. Chen、S. Liang、J. Ma、L. Chen、S. Wang和J. Xiao，“使用神经说话人嵌入的跨语言、多说话人文本到语音合成”，载于*Interspeec̄h*，2019年，第2105–2109页。

[16] A. ¨Oktem、M. Farr´us和A. Bonafonte，“机器配音的韵律短语对齐”，载于Interspeech 2019。ISCA，2019年9月，第4215–4219页。 [在线]。网址：https://www.isca-speech.org/archive/ interspeech 2019/oktem19 interspeech.html[17] J. Li、Y. Meng、X. Wu、Z. Wu、J. Jia、H. Meng、T. Qiao、Y. Wang和Y. Wang，“通过多尺度关系图卷积网络，从多模态对话语境中推断说话风格”，载于第30届ACM多媒体国际会议论文集，系列号MM ’22。葡萄牙里斯本：ACM，2022年。 [在线]。网址：http://doi.acm.org/ 10 1145 3503161 3547831 。/。W. Brannon、Y. Virkar和[18] B. Thompson，“配音实践：大规模人工本地化研究及其对自动配音的启示”，计算语言学协会会刊，第11卷，2023年5月。 [在线]。网址：https://transacl.org/ index.php/tacl/article/view/ 4447[19] J. Effendi、Y. Virkar、R. Barra-Chicote和 M. Federico，“用于自动配音的神经文本到语音时长建模”，载于ICASSP 2022 - 2022年IEEE声学、语音与信号处理国际会议（ICASSP），2022年，第8037–8041页。[20] M. Federico、Y. Virkar、R. Enyedi和R. Barra-Chicote，“自动配音韵律对齐的评估与优化”，载于INTER-SPEECH，2020年，第1481–1485页。[21] Y. Virkar、M. Federico、R. Enyedi和R. Barra-Chicote，“画外自动配音的韵律对齐”，载于Interspeech 2022。ISCA，2022年9月，第496–500页。 [在线]。网址：https://www. isca-speech.org/archive/interspeech 2022/virkar22 interspeech.html [22] J. Swiatkowski、D. Wang、M. Babianski、G. Coccia、P. Lumban Tobing、R. Vipperla、V. Klimkov和V. Pollet，“通过短语级跨语言韵律迁移实现表现力机器配音”，载于INTERSPEECH 2023。ISCA，2023年8月，第5546–5550页。[在线]。网址：https://www.isca-speech.org/archive/interspeech 2023/swiatkowski23b interspeech.html [23] Y. Virkar、M. Federico、R. Enyedi和R. Barra-Chicote，“自动配音韵律对齐的改进”，载于ICASSP 2021-2021年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2021年，第7543–7574页。[24] H. Zen、K. Tokuda和A. W. Black，“统计参数语音合成”，语音通信，第51卷，第11期，第1039–1064页，2009年。[25] A. v. d. Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior和K. Kavukcuoglu，“WaveNet：一种原始音频生成模型”，arXiv预印本 arXiv:1609.03499，2016年。[26] J. Kong、J. Kim和J. Bae，“HiFi-GAN：用于高效高保真语音合成的生成对抗网络”，神经信息处理系统进展，第33卷，第17 022–17 033页，2020年。[27] A. Graves和N. Jaitly，“迈向基于循环神经网络的端到端语音识别”，载于国际机器学习会议。PMLR，2014年，第1764–1772页。[28] K. Cho、B. van Merrienboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk和Y. Bengio，“用于统计机器翻译的RNN编码器–解码器短语表示”，载于2014年自然语言处理经验方法会议（EMNLP）论文集。计算语言学协会，2014年，第1724页。 [在线]。网址：https://cir.nii.ac.jp/crid/ 1364233270210557696[29] D. Bahdanau、K. H. Cho和Y. Bengio，“通过联合学习对齐与翻译实现神经机器翻译”，载于第三届国际学习表征会议，ICLR 2015 [ ] 2015 。在线。网址：https://nyuscholars.nyu.edu/en/publications/ neural-machine-translation-by-jointly-learning-to-align-and-trans-2[30] J. K. Chorowski、D. Bahdanau、D. Serdyuk、K. Cho和Y. Bengio，“用于语音识别的基于注意力的模型”，神经信息处理系统进展，第28卷，2015年。[31] Y. Ganin、E. Ustinova、H. Ajakan、P. Germain、H. Larochelle、F. Lavi-olette、M. Marchand和V. Lempitsky，“神经网络的领域对抗训练”，机器学习研究期刊，第17卷，第1期，第2096–2030页，2016年。[32] J. Li、Y. Meng、Z. Wu、H. Meng、Q. Tian、Y. Wang和Y. Wang，“NeuFA：基于双向注意力机制的神经网络端到端强制对齐”，载于ICASSP 2022-2022年IEEE声学、语音与信号处理国际会议（ICASSP）。IEEE，2022年，第8007–8011页。[33] Y. Jia、R. J. Weiss、F. Biadsy、W. Macherey、M. Johnson、Z. Chen和Y. Wu，“基于序列到序列模型的直接语音到语音翻译”，载于Interspeech 2019。ISCA，2019年9月，第1123–1127页。 [在线]。网址：https://www.isca-speech.org/archive/ interspeech 2019/jia19 interspeech.html

[34] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, “真实数据上的无文本语音到语音翻译”，载于《*2022*年北美计算语言学协会分会年会论文集：人类语言技术》，M. Carpuat、M.-C. de Marneffe 和 I. V. Meza Ruiz 编。西雅图，美国：计算语言学协会，2022年7月，第860–872页。 [ ] 2022 63 在线。可查阅：[35] https://aclanthology.org/ .naacl-main. G. Kikui、E. Sumita、T. Takezawa 和 S. Yamamoto，“创建用于语音到语音翻译的语料库”，载于第八届欧洲语音通信与技术会议，2003年。[36] F. Chaume，“配音”，载于帕尔格雷夫视听翻译与媒体无障碍手册，丛书：帕尔格雷夫翻译与口译研究丛书，L. Bogucki 和 M. Deckert 编。卡姆：施普林格国际出版，2020年，第103–132页。 [在线]。可查阅：https://doi.org/ 10 1007 978 3 030 42105 2 6 。 / - - - [37] L. Rabiner 和 B. Juang，“隐马尔可夫模型简介”，*IEEE ASSP* 杂志，第 3 卷，第1期，第4–16页，1986年。[38] M. Federico、R. Enyedi、R. Barra-Chicote、R. Giri、U. Isik、A. Krishnaswamy 和 H. Sawaf，“从语音到语音翻译到自动配音”，载于第*17*届国际口语翻译会议论文集。在线：计算语言学协会，2020年，第257–264页。 [在线]。可查阅：https://aclanthology.org/ 2020 1 31 .iwslt- . M. McAuliffe、M. Socolof、[39] S. Mihuc、M. Wagner 和 M. Sonderegger，“蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本-语音对齐”，*INTERSPEECH 2017*论文集，第498–502页，2017年。[40] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，载于《*2019*年北美计算语言学协会分会年会论文集：人类语言技术》，第*1*卷（长篇与短篇论文）。明尼阿波利斯，明尼苏达州：计算语言学协会，2019年6月，第 4171 4186 [ ] 19 1423 –页。在线。[41] 可查阅：https://aclanthology.org/N - A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，神经信息处理系统进展，第30卷，2017年。[42] Y. Ganin 和 V. Lempitsky，“通过反向传播进行无监督领域自适应”，载于国际机器学习会议。PMLR，2015年，第1180–1189页。[43] Y. Zhang、R. J. Weiss、H. Zen、Y. Wu、Z. Chen、R. Skerry-Ryan、Y. Jia、A. Rosenberg 和 B. Ramabhadran，“学会用外语流利地说话：多语言语音合成与跨语言声音克隆”，载于*Interspeech 2019*。ISCA，2019年9月，第2080–2084页。 [在线]。可查阅：https://www.isca-speech.org/ archive/interspeech 2019/zhang19e interspeech.html [44] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，载于《*2019*年北美计算语言学协会分会年会论文集：人类语言技术》，第*1*卷（长篇与短篇论文）。明尼阿波利斯，明尼苏达州：计算语言学协会，2019年6月，第 4171 4186 [ ] 19 1423 –页。在线。[45] 可查阅：https://aclanthology.org/N - A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga 等，“PyTorch：一种命令式风格的高性能深度学习库”，神经信息处理系统进展，第32卷，2019年。