# 基于混合密度网络的苗语语音合成方法

蔡 姗[1,2],郭 胜[1,2],王 林[1,2]

(1. 贵州民族大学 数据科学与信息工程学院;2. 贵州省模式识别与智能系统重点实验室,贵州 贵阳 550025)

**摘 要:** 苗语语音合成研究对民族文化的传承、保护和发展具有重要意义。针对苗语存在文字缺失、电子资源匮乏及数据难以获取导致其语音合成研究滞后的问题,提出一种基于混合密度网络的苗语语音合成方法。该方法根据持续时间来学习文本与语音间的对齐,解决了根据注意力机制学习对齐时容易出现的漏词、重复等问题。利用混合密度网络提取文本真实的持续时间,并与持续时间预测器联合训练,不需要额外的外部对齐器或自回归模型来指导模型进行对齐学习,简化了模型训练过程。以自建苗语语音合成语料库Hmong_data为基准数据,与先进方法进行对比实验。实验结果显示,该方法的平均意见得分为3.89,较Tacotron2方法提升了0.41,且产生的对齐图更清晰、平滑,合成的语音是可理解和正确的。

**关键词:** 苗语;语音合成;混合密度网络;语料库

**DOI:** 10.11907/rjdk.231900 **开放科学(资源服务)标识码(OSID):**

**中图分类号:** TP391;TN912.33 **文献标识码:** A **文章编号:** 1672-7800(2024)004-0031-07

# Mixure Density Network-Based Hmong Language Text-to-Speech Method

CAI Shan[1,2], GUO Sheng[1,2], WANG Lin[1,2]

(1. *College of Data Science and Information Engineering, Guizhou Minzu University*;
2. *Key Laboratory of Pattern Recognition and Intelligent System of Guizhou Province, Guiyang* 550025, *China*)

**Abstract:** The research on Hmong language text-to-speech is of great significance for the inheritance, protection, and development of ethnic culture. In response to the problems of missing text, lack of electronic resources, and difficulty in obtaining data for Hmong language, a mixure density network-based Hmong language speech synthesis method is proposed. This method learns the alignment between text and speech based on duration, addressing issues such as missing words and repetitions that may occur during alignment learning with attention mechanism. The mix density network is used to extract the real duration of the text and jointly trained with the duration predictor, eliminating the need for additional external aligners or autoregressive models to guide alignment learning, simplifying the complexity of model training. Using the self-built Hmong language text-to-speech corpus, Hmong_data, as the benchmark data, comparative experiments are conducted with advanced methods. The experimental results shows that the proposed method achieves an average opinion score of 3.89, which is a 0.41 improvement over the Tacotron2 method. The generated alignment graphs are clearer and smoother, and the synthesized speech is considered understandable and correct.

**Key Words:** Hmong language; text-to-speech; mixure density network; corpus

## 0 引言

语音合成(Text-to-Speech,TTS)[1]是一种从任何给定的文本中生成相应语音的技术,在人机语音交互中有着广泛应用,如手机语音助手、无障碍辅助及车载导航等。苗语是苗族人交流使用最广泛的语言,其使用人口约为900万,具有重要的文化、经济和社会价值。因此,对苗语的语音合成研究有助于促进苗族语言文化的保护和传承,让更多人了解和使用苗语,为苗族的文化遗产保护和发展作出

**收稿日期:** 2023-08-21

**基金项目:** 贵州省科技计划项目(黔科合基础-ZK[2022]一般195,黔科合基础-ZK[2023]一般143);贵州省教育厅自然科学研究项目(黔教技[2023]061号,黔教技[2023]012号)

**作者简介:** 蔡姗(1996-),女,贵州民族大学数据科学与信息工程学院硕士研究生,研究方向为语音合成、图像处理;郭胜(1997-),男,贵州民族大学数据科学与信息工程学院硕士研究生,研究方向为模式识别与图像处理;王林(1965-),男,博士,贵州民族大学数据科学与信息工程学院教授、博士生导师,研究方向为模式识别与图像处理。本文通讯作者:王林。

贡献，同时为推动少数民族地区的经济发展提供支持[2]。

## 1 相关工作

传统的语音合成技术主要包括基于规则的方法和基于统计学习的方法。基于规则的方法是一种基于人工规则和知识库的语音合成技术，能针对特殊口音、方言等特定情况进行优化，但需要人工参与进行语音合成规则的编写，因而耗时耗力，且合成的语音也不够自然[3]。基于统计学习的方法则是利用机器学习技术，根据语音数据来学习语音合成模型，不需要人工编写规则，但需要大量的语音数据和计算资源[4]。

随着深度学习技术的发展，基于神经网络的语音合成方法取得了很大进展，成为语音合成的一个重要研究方向[5]。为减少语音合成对语言学知识的依赖，端到端的生成方式在语音合成中得到了广泛应用。其主要采用带有注意力机制的编码器—解码器结构，注意力机制用来学习输入文本与语音之间的隐式对齐，其中最具代表性的有Tacotron[6]、Tacotron2[7]模型。Tacotron是一种基于序列到序列的端到端生成式语音合成模型，该模型以字符为输入，输出梅尔谱图，并以Griffin-Lim方法为合成器直接从梅尔谱图合成语音[8]。Tacotron2使用简单的卷积层和双向LSTM作为编码器，以更好地对上下文信息进行提取，采用的位置敏感注意力机制不但综合了内容方面的信息，而且关注了位置特征，同时舍弃了Tacotron中简单的Griffin-Lim算法，采用改进的WaveNet进行波形重建[9]。上述方法采用递归的方式从文本进行逐帧的梅尔谱图预测和波形生成，推理速度较慢，且容易出现注意力的崩溃，导致合成的语音出现漏词、跳词及重复等问题。为此，Ren等[10-11]提出FastSpeech和FastSpeech2来并行生成语音，提高了语音合成速度，且不采用注意力机制，而是采用文本持续时间来学习文本与梅尔谱图之间的对齐。但由于这些模型依赖外部对齐器或预先训练的自回归模型来提取持续时间，增加了训练的复杂性。因此，为了提高合成语音质量，非自回归TTS模型有必要设计出更合适的方法来直接学习对齐[12]。

目前，汉语、英语等主流语言的语音合成研究相对成熟[13]，小语种、少数民族语言及地区方言等低资源语言的语音合成研究也逐渐受到越来越多学者的关注[14-15]。但苗语存在文字缺失、电子资源匮乏及数据难以获取等问题，阻碍了苗语语音合成的发展进程。针对此问题，本文提出基于混合密度网络的苗语语音合成方法（Mix density network-based Hmong language Text-to-Speech，MHTTS），该方法以并行方式生成梅尔谱图，其训练不需要其他自回归TTS模型的指导。MHTTS由前馈Transformer[16]、持续时间预测器和混合密度网络[17]组成。前馈Transformer是一个前馈网络，用于将文本转换为梅尔谱图，其中需要持续时间预测器预测的每个字符持续时间来调节推理中的对齐。最后，采用HiFi-GAN声码器将梅尔谱图转化为语音波形[18]。实验结果表明，MHTTS合成的语音具有较高的保真度和自然性。通过对齐损失来训练混合密度网络，以提取真实的持续时间，没有引入额外的模型，简化了训练的复杂性。

## 2 数据收集与预处理

为了开发一个好的语音合成系统，需要收集大量文本数据，并转录相应内容。由于目前缺少公开可用的苗语语音数据集，需从零开始准备一个标准的苗语语音合成语料库。

### 2.1 文本语料库

据文献[19]所述，构建的文本语料库应保持音素平衡，尽量覆盖苗语中所有可能的发音，以避免出现集外词的情况。本文的研究对象为中部苗语，也称黔东苗语，表1给出了所构建的文本语料库的统计信息。

**Table 1 Statistical information of Hmong text corpus**

**表1 苗语文本语料库统计信息**

| 分类 | 统计 | |
|---|---|---|
| 总句子 | 3 545 | |
| 音节（bangx） | 总和 | 25 661 |
| | 最短句子 | 3 |
| | 最长句子 | 18 |
| 子音节（b、angx） | 总和 | 48 161 |
| | 最短句子 | 5 |
| | 最长句子 | 35 |

### 2.2 语音数据

在文本语料库准备完成后，选择一名母语为黔东苗语的女大学生为录音者，按照中部苗语的标准音进行发音，在安静环境下用专业麦克风、以正常语速进行录制。语音文件的存储格式为.wav，采样率为44.1KHz，单声道。

### 2.3 文本预处理

文本规范化是预处理的一部分，旨在将原始文本转换为其发音形式，使模型能准确学习输入文本的对应发音。现代苗文的书写形式是“声母+韵母+声调”，是一种以拉丁字符为基础的拼音型文字，共有32个声母、26个韵母和8个声调[20]。表2、表3分别展示了苗语声韵母及声调信息。

由于目前先进的语音合成系统是针对英语开发的，并以字符作为输入基元进行训练来学习输入文本对应的发

**Table 2 Vocal vowels of Hmong language**

**表2 苗语声韵母**

| 声母 | 韵母 |
|---|---|
| b, p, m, hm, f, hf, w, d, t, n, hn, dl, hl, l, z, c, s, hs, r, j, q, x, hx, y, g, k, ng, v, hv, gh, kh, h | i, e, a, o, u, ai, ei, ia, io, ie, iu, ang, en, ong, in, iang, iong, ee, ao, iee, iao, ui, ua , uai, un,uang |

**Table 3　Tone of Hmong language**

**表 3　苗语声调**

| 调类 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 调值 | 33 | 55 | 35 | 11 | 44 | 13 | 53 | 31 |
| 调符 | b | x | d | l | t | s | k | f |
| 例词 | dab | dax | dad | dal | dat | das | dak | daf |
| 译文 | 回答 | 来 | 长短 | 丢失 | 早晨 | 死 | 翅膀 | 搭 |

音。但由表 2 和表 3 可知,苗语中声调的调符与部分声母相同,且不同的声调表示不同的含义。故以字符作为模型的训练基元可能无法表示苗语的正确发音,混淆同形异音的部分声母与声调,从而导致合成语音的发音错误。此外,由于苗语是单音节语言,以音节本身作为输入基元虽然保留了文本对应的发音信息,但在构建词典库时会产生过大的字符集,使特征提取器提取文本特征向量时产生过大的编码维度,导致在后续的特征学习过程中因维度降低而损失部分有效的文本特征信息,且需要很大的语料库才能包含所有的发音组合。因此,一个可执行的方式是以声母和带声调的韵母作为训练基元,称为子音节,在保证发音和声调信息的同时降低编码维度。表 4 显示了一些对苗文进行预处理操作的例子,由于现代苗文中字母的大小写对发音没有影响,故在训练时所有文本全部统一为小写。

**Table 4　Text preprocessing**

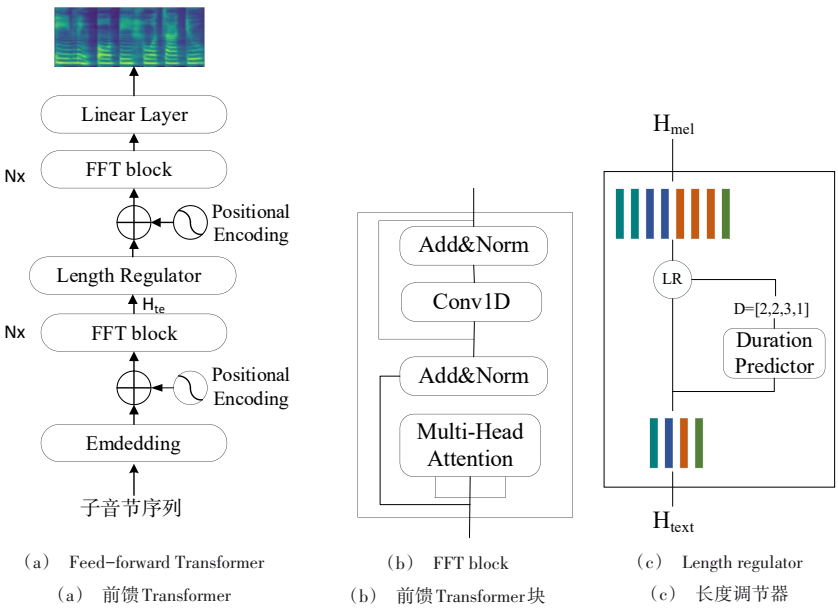
**表 4　文本预处理**

| 输入文本 | 训练基元 | 译文 |
|---|---|---|
| det liax eb diul jil | d et l iax eb d iul j il | 杨柳枝下垂 |
| ax gid dlab lal naix yel | ax g id dl ab l al n aix y el | 不要欺骗人了 |
| jib daib xangt dlangd dud | j ib d aib x angt dl angd d ud | 小孩放风筝 |
| bet nongd maix laib zaid | b et n ongd m aix l aib z aid | 这里有栋房子 |
| wil zaid ax maix diux | w il z aid a x m aix d iux | 我家没有门 |
| ib hnaib ghuk ib had | ib hn aib gh uk ib h ad | 一天省一口 |

## 3　苗语语音合成模型

苗语属于汉藏语系苗瑶语支,现代苗语采用拉丁字符的书写形式。尽管目前关于英语、汉语等主流语言的语音合成研究取得了较多成就,但对于少数民族语言的语音合成研究还相对较少。这是由于苗语存在文字缺失、语料难以获取使得可用训练样本少等问题,导致其语音合成研究滞后。

针对上述问题,提出一种基于混合密度网络的苗语语音合成方法,模型结构如图 1、图 2 所示。MHTTS 使用持续时间来调节输入文本与梅尔谱图之间的对齐,以学习文本准确的发音,其包含前馈 Transformer、持续时间预测器和混合密度网络 3 个模块,下面将详细描述每一个模块。




(a) Feed-forward Transformer　(b) FFT block　(c) Length regulator

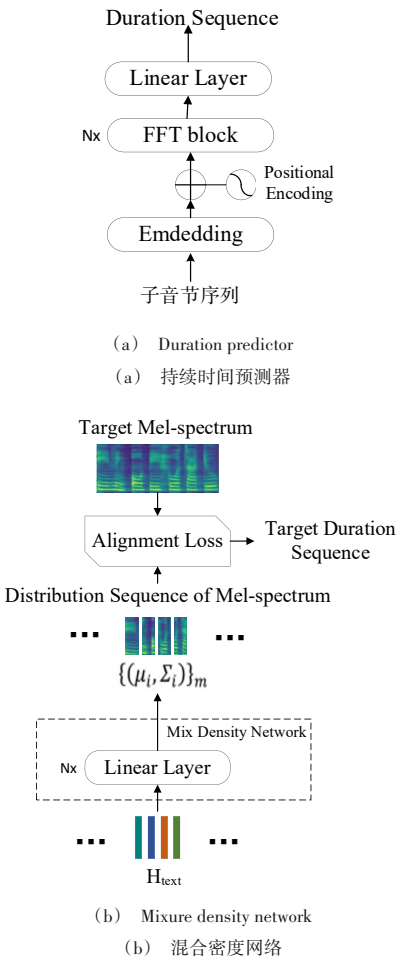
(a) 前馈 Transformer　(b) 前馈 Transformer 块　(c) 长度调节器

**Fig. 1　Framework of MHTTS model**

**图 1　MHTTS 模型框架**

### 3.1　前馈 Transformer

前馈 Transformer(Feed-Forward Transformer,FFT)是一个前馈网络,以并行的方式从文本预测出梅尔谱图。如图 1(a)所示,FFT 包含一个嵌入层(Embedding)、多个 FFT 块、一个长度调节器(Length Regulator,LR)和一个线性层。多个 FFT 块被 LR 划分为两部分,其中 LR 下方的多个 FFT 块用于编码提取输入文本特征,LR 上方的多个 FFT 块用于解码预测梅尔谱图。如图 1(b)所示,FFT 块由来自 Transformer 的多头注意力机制和一维卷积组成,并采用残差结构来防止梯度消失和过拟合问题。此外,LR 根据预测的持续时间序列来调节文本与梅尔谱图之间的对齐,如图 1(c)所示,即根据持续时间序列对文本特征进行长度的复




（a） Duration predictor

（a） 持续时间预测器

（b） Mixure density network

（b） 混合密度网络

**Fig. 2 Submodules of MHTTS**

**图 2 MHTTS 子模块**

制，以匹配输出的梅尔谱图长度。

### 3.2 持续时间预测器

由于输入文本序列远比输出梅尔谱图序列短，故定义持续时间为每个字符对应的梅尔谱图帧数。如图 2(a)所示，持续时间预测器(Duration Predictor，DP)由一个 Embedding 层、多个 FFT 块和一个线性层组成。DP 通过输入子音节序列，预测输出对应的持续时间序列。在推理阶段，DP 将预测的持续时间序列用于 LR 以调整文本与梅尔谱图之间的对齐。

### 3.3 混合密度网络

为了提取 FFT 和 DP 训练所需的文本与梅尔谱图之间的正确对齐，设计了混合密度网络(Mix Density Network，MDN)来学习对齐。与以往采用非自回归提取真实持续时间的方法不同，这里不需要额外的自回归教师模型和外部对齐器来产生持续时间。如图 2(b)所示，MDN 由多个线性层堆叠而成，每个层都经过层归一化、Relu 激活函数和 dropout。最后一个线性层输出多维高斯分布的均值和方差向量，表示每个子音节的梅尔谱图分布。MDN 仅用于训练阶段，可以在推理中删除。

### 3.4 对齐损失

根据 Baum-Walch 算法思想设计一种对齐损失来训练 MDN 和学习文本与梅尔谱图之间的准确对齐。令 $y = \{y_1, y_2, \cdots, y_n\}$ 表示梅尔谱图序列，n 为梅尔谱图的帧数，$z = \{z_1, z_2, \cdots, z_m\}$ 表示 MDN 输出的多维高斯分布序列，其中 $z_i = (\mu_i, \Sigma_i)$ 为均值向量和方差矩阵，m 为输入子音节序列的长度。则第 $i$ 帧梅尔谱相对于多维高斯分布序列的第 $j$ 个分布概率可用多维高斯函数进行计算，如式(1)所示。

$$p(y_i \mid z_j) = N(y_i \mid \mu_j, \Sigma_j) \tag{1}$$

令 $l$ 表示对齐，$j = \mathrm{Align}(i, l)$ 表示第 $i$ 帧梅尔谱对齐到分布序列的第 $j$ 个元素。目标函数可表示为：

$$p(y, y \mid z) = \prod_{i=0}^{n} p(y_i \mid z_{\mathrm{Align}(i,l)}) \tag{2}$$

然而，由于文本与梅尔谱图之间的对齐是未知的，不能准确地计算等式(2)，故利用 Baum-Walch 算

法来解决此问题。Baum-Walch 算法考虑了所有可能的对齐，并累加这些对齐概率，最终的目标函数如式(3)所示。

$$p(y \mid z) = \sum_{l} \prod_{i=0}^{n} p(y_i \mid z_{\mathrm{Align}(i,l)}) \tag{3}$$

由此得到的损失函数如式(4)所示。

$$l_{\mathrm{align}}(z, y) = -\log p(y \mid z) \tag{4}$$

## 4 实验与分析

### 4.1 实验设置

为评估所提方法的有效性，在自建的单说话人苗语语音数据集 Hmong_data 上进行实验分析。Hmong_data 包含 3 545 句音频片段，采样率为 44.1 kHZ，录音者为女性，时长为 3.6 h。文本语料库的内容主要是日常生活用语，考虑了苗语的常见发音组合。实验中随机将语料库划分为 3 个子集：训练集(90%)、验证集(5%)、测试集(5%)。

前馈 Transformer 中长度调节器两侧的 FFT 块都为 6 个，注意力头的数量为 2。所有实验在 Linux 系统上使用 tensorflow 的深度学习框架完成，批量大小为 32，使用 Adam 优化器，固定学习率为 1e-4，总共训练 80 000 步。重新在 Hmong_data 数据上训练 HiFi-GAN 作为所有方法的声码器，从而将预测生成的梅尔谱图转化为语音波形。

### 4.2 不同方法语音质量评估

为验证 MHTTS 的有效性，将 MHTTS 与 Tacotron、Tacotron2 及真实语音(GT)进行比较。从测试集中随机选择 20 句文本作为不同模型的输入，以获得合成的语音。将合成语音与真实语音混合在一起，邀请 10 名母语为黔东苗语的青年学生对这些语音进行评分。收集所有评分后计算平均意见得分(Mean Opinion Score，MOS)作为语音合成方法的评价指标。MOS 采用五级评分标准(1—极差，2—一般，3—可接受，4—良好，5—非常好，精度为 0.5)，实验结

果如表5所示。

由表5可知,基于持续时间的MHTTS方法显著优于基于注意力机制的方法。Tacotron与Tacotron2的MOS分别为3.04和3.48,MHTTS的MOS为3.89,较前两者分别提高了0.85和0.41,表明该方法合成语音的可懂度、自然性和流畅性更好,且与真实语音的MOS值仅有0.51之差,取得不错的效果。

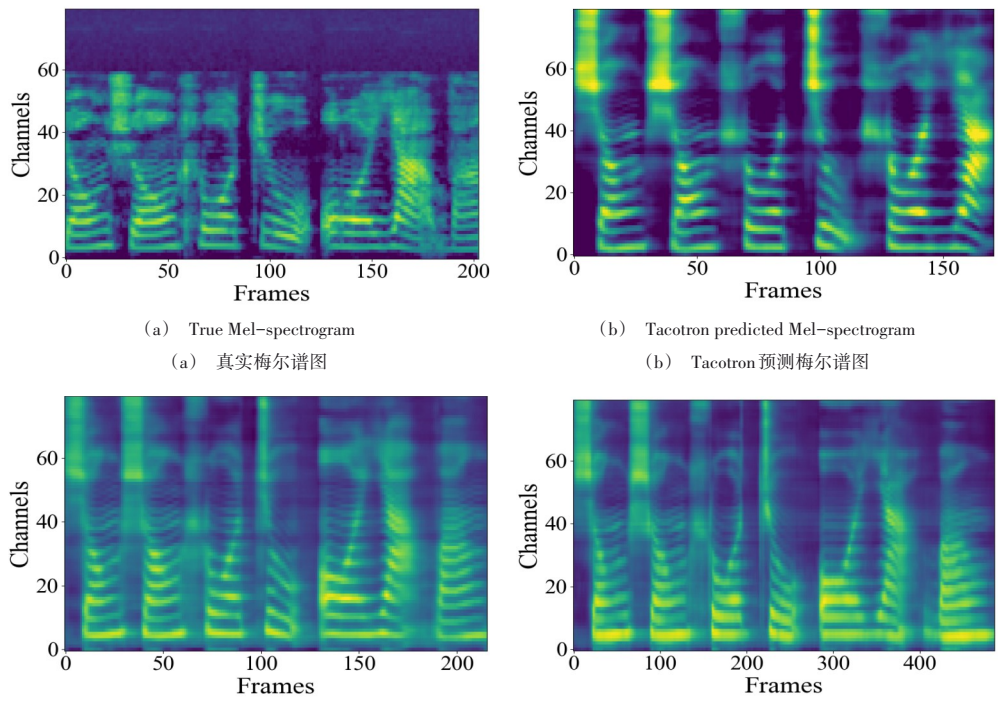
**Table 5　MOS score results of different methods**

**表5　不同方法的MOS评分结果**

| 评测员 | Tacotron | Tacotron2 | MHTTS | GT |
|---|---|---|---|---|
| 评测者1 | 3.20 | 3.55 | 4.20 | 4.50 |
| 评测者2 | 2.90 | 3.35 | 3.95 | 4.60 |
| 评测者3 | 2.65 | 3.60 | 3.85 | 4.35 |
| 评测者4 | 3.25 | 3.25 | 3.60 | 4.50 |
| 评测者5 | 3.35 | 3.70 | 4.15 | 4.20 |
| 评测者6 | 2.70 | 3.45 | 3.80 | 4.30 |
| 评测者7 | 2.80 | 3.50 | 3.70 | 4.15 |
| 评测者8 | 3.15 | 3.40 | 3.90 | 4.45 |
| 评测者9 | 3.45 | 3.60 | 3.75 | 4.30 |
| 评测者10 | 2.90 | 3.35 | 4.00 | 4.65 |
| MOS | 3.04 | 3.48 | **3.89** | 4.40 |

### 4.3　不同方法预测分析

为直观分析MHTTS的有效性,对不同方法预测的梅尔谱图特征及对齐图进行可视化。图3为不同方法预测的梅尔谱图,Tacotron(见图3(b))和Tacotron2(见图3(c))预测的梅尔谱图中部分谐波与真实梅尔谱图(见图3(a))差异较大,如实线框部分所示,而图3(d)所示MHTTS预测的梅尔谱图更接近真实梅尔谱图。图4展示了各方法在测试时学习到的文本序列与梅尔谱图间的对应关系,与Tacotron(见图4(a))和Tacotron2(见图4(b))两种方法产生的对齐图相比,MHTTS可学习到更平滑、清晰的对齐(图4(c)),表明MHTTS能更准确地建立文本与语音间的对齐,从而合成出更高保真、准确的苗语语音。




(a)　True Mel-spectrogram

(a)　真实梅尔谱图

(b)　Tacotron predicted Mel-spectrogram

(b)　Tacotron预测梅尔谱图

(c)　Tacotron2 predicted Mel-spectrogram

(c)　Tacotron2预测梅尔谱图

(d)　MHTTS predicted Mel-spectrogram

(d)　MHTTS预测梅尔谱图

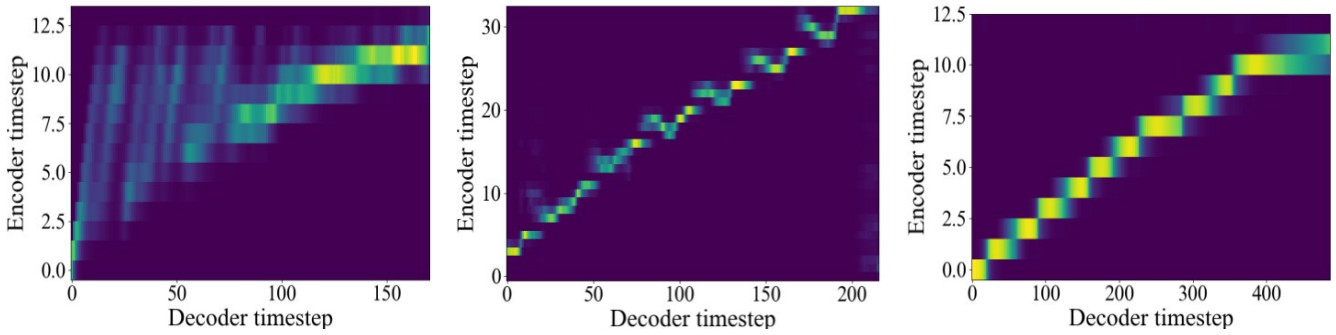
**Fig. 3　Prediction of Mel-spectrogram features**

**图3　梅尔谱图特征预测情况**

### 4.4　鲁棒性分析

为验证MHTTS模型的鲁棒性,统计了各模型合成语音的词错误率(Word Error Rate,WER)。将MHTTS评估结果与Tacotron及Tacotron2模型进行比较,结果如表6所示。由表6可知,所提MHTTS方法的重词数为1,漏词数为0,错词数为2,最终的词错误率仅为1.44%。与Tacotron和Tacotron2相比,词错误率分别降低了25.48%和7.21%,表明MHTTS方法具有较强的鲁棒性,且有效减少了合成语音的重复、跳跃等问题,提高了语音合成的可理解性。

为验证MHTTS模型在其他公开数据集上的有效性,选择了标贝科技公司提供的中文数据集,因为中文的书写形式与苗语类似,都为“声母+韵母+声调”的形式。表7是在中文数据集上合成的实验结果。由表7可知,MHTTS模型提升了合成语音质量,并减少了合成的错误率,表明模型具有较好的鲁棒性,针对相同输入形式的数据集也同样表现优越。




（a） Tacotron alignment diagram
（a） Tacotron对齐图

（b） Tacotron2 alignment diagram
（b） Tacotron2对齐图

（c） MHTTS alignment diagram
（c） MHTTS对齐图

**Fig. 4 Alignment effect of different methods**

**图4 不同方法对齐效果**

**Table 6 Results of robustness evaluation**

**表6 鲁棒性评估结果**

| 方法 | 重词数 | 漏词数 | 错词数 | WER/% |
|---|---|---|---|---|
| Tacotron | 10 | 13 | 33 | 26.92 |
| Tacotron2 | 4 | 2 | 12 | 8.65 |
| MHTTS | **1** | **0** | 2 | **1.44** |

**Table 7 Evaluation results of Chinese dataset synthesis**

**表7 中文数据集合成评估结果**

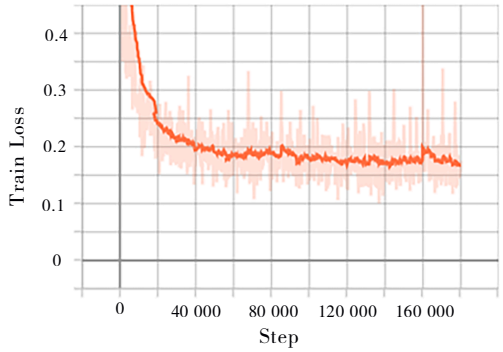
| 方法 | MOS | WER/% |
|---|---|---|
| Tacotron | 3.52 | 18.73 |
| Tacotron2 | 3.69 | 6.58 |
| MHTTS | **3.81** | **0.94** |

### 4.5 稳定性分析

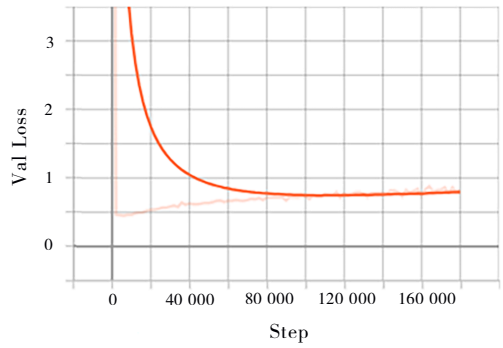
针对MHTTS模型的稳定性和收敛性分析，给出了模型在训练和验证阶段的损失图，如图5所示。由图5可知，MHTTS模型在训练80 000步之后逐渐趋于稳定，损失值呈平缓状态，只在小范围内波动，验证损失在一定步数之后也逐渐稳定在0.5附近，说明模型具有较好的稳定性和收敛性。

## 5 结语

本文针对苗语存在文字缺失、语料难以获取使得可用训练样本少，导致其语音合成研究滞后的问题，提出基于混合密度网络的苗语语音合成方法（MHTTS）。MHTTS利用持续时间来学习文本与梅尔谱图间的对齐，可合成出比注意力机制更准确的苗语语音。同时，根据苗语的发音规则提出用子音节作为训练基元，以准确地学习文本对应的发音。通过与不同方法的主观指标对比实验，结果显示，MHTTS可以合成出更高质量的语音。但由于苗语电子资源匮乏等原因，目前构建的苗语语音合成语料库规模还较小，也未能很好地学习其中的韵律，在未来的研究中将持续扩充语料库，并探索更好的合成方法。




（a） Training loss diagram
（a） 训练损失图




（b） Validation loss diagram
（b） 验证损失图

**Fig. 5 Loss diagram**

**图5 损失图**

**参考文献：**

[1] YASUDA Y, TODA T. Text-to-speech synthesis based on latent variable conversion using diffusion probabilistic model and variational autoencoder [C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2023: 1-5.

[2] LIU R, KANG S Y, GAO G L, et al. MonTTS: a fully non-autoregressive real-time and high-fidelity mongolian text-to-speech synthesis model [J]. Journal of Chinese Information Processing, 2022, 36(7): 86-97.
刘瑞，康世胤，高光来，等. MonTTS：完全非自回归的实时、高保真蒙古语语音合成模型[J]. 中文信息学报，2022，36(7)：86-97.

[3] HUNT A, BLACK A W. Unit selection in a concatenative speech synthesis system using a large speech database[C]//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1996: 373-376.

[4] TOKUDA K, NANKAKU Y, TODA T, et al. Speech synthesis based on hidden Markov models[J]. Proceedings of the IEEE, 2013, 101(5): 1234-1252.

[5] ZHANG K, GONG C, LU W, et al. Joint and adversarial training with ASR for expressive speech synthesis[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2022: 6322-6326.

[6] WANG Y, SKERRY R, STANTON D, et al. Tacotron: towards end-to-end speech synthesis [C]// Proceedings of the 18th Annual Conference of the International Speech Communication Association, 2017: 4006-4010.

[7] SHEN J, PANG R, WEISS R J, et al. Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2018: 4779-4783.

[8] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1983: 804-807.

[9] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[C]//Proceedings of the 9th ISCA Speech Synthesis Workshop, 2016: 125.

[10] REN Y, RUAN Y, TAN X, et al. FastSpeech: fast, robust and controllable text to speech[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 3165-3174.

[11] REN Y, HU C, TAN X, et al. Fastspeech 2: fast and high-quality end-to-end text to speech[C]// Proceedings of the International Conference on Learning Representations, 2020: 1-15.

[12] NGUYEN B, CARDINAUX F, UHLICH S. Autotts: end-to-end text-to-speech synthesis through differentiable duration modeling[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2023: 1-5.

[13] LEE M, LEE J, CHANG J H. Non-autoregressive fully parallel deep convolutional neural speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30(1): 1150-1159.

[14] ZU B, CAI R, CAI Z, et al. Research on Tibetan speech synthesis based on fastspeech2[C]//Proceedings of the 3rd International Conference on Pattern Recognition and Machine Learning, 2022: 241-244.

[15] LI G, LI G, DAI Y, et al. Research on end to end low resource speech synthesis based on meta learning[C]//Proceedings of the 4th International Academic Exchange Conference on Science and Technology Innovation, 2022: 1059-1064.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Conference on Neural Information Processing Systems, 2017: 5998-6008.

[17] ZEN H, SENIOR A. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2014: 3844-3848.

[18] KONG J, KIM J, BAE J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis [DB/OL]. https://arxiv.org/abs/2010.05646.

[19] RENZENG Z M, ZHU L P. Tibetan dialect speech synthesis dataset[J]. China Scientific Data (English Version), 2022, 7(2): 24-33.
仁曾卓玛, 朱丽平. 藏语方言语音合成数据集[J]. 中国科学数据(中英文网络版), 2022, 7(2): 24-33.

[20] ZHANG X W, WANG L, FENG F J, et al. Isolated word speech recognition of Hmong language based on convolutional neural network[J]. Software Guide, 2022, 21(2): 21-26.
张学文, 王林, 冯夫健, 等. 基于卷积神经网络的苗语孤立词语音识别[J]. 软件导刊, 2022, 21(2): 21-26.

(责任编辑:黄　健)