Computer Science and Application 计算机科学与应用, 2022, 12(1), 33-45
Published Online January 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.121005

Hans汉斯

基于语音参数自适应的缅甸语情感语音合成

刘奇云，杨　鉴*，谭婉琳

云南大学，信息学院，云南　昆明

收稿日期：2021年12月14日；录用日期：2022年1月10日；发布日期：2022年1月17日

摘　要

相比汉语和英语，缅甸语的语音合成技术发展相对滞后，合成的语音缺乏情感。情感语音合成使机器表达不再生涩，采用基于HMM声学模型的语音参数自适应方法，研究缅甸语情感语音合成。情感语音合成研究面临的一个困难是难以获取大规模的情感语音库，在低资源条件下提出了一种实现缅甸语情感语音合成的方法。首先在MFA (蒙特利尔强制对齐)平台进行缅甸语音子自动切分以训练语音声学模型，基于HTS平台采用中规模的缅甸语平静情感语音库，构建缅甸语语音合成基线系统。在此基础上，基于少量的高兴、悲伤、生气情感语音数据，采用语音参数自适应方法，构建缅甸语情感语音合成系统，并通过引入平均音模型和调整转换矩阵的方法进一步改进情感语音合成系统。实验结果表明，情感语音合成系统可合成出平静、高兴、悲伤、生气四种情感的缅甸语语音，EMOS平均评分可达3.40，证明了方法的有效性。

关键词

缅甸语，MFA音子自动切分，情感语音合成，语音参数自适应

Burmese Emotional Speech Synthesis Based on Speech Parameter Adaptation

Qiyun Liu, Jian Yang*, Wanlin Tan

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

Received: Dec. 14th, 2021; accepted: Jan. 10th, 2022; published: Jan. 17th, 2022

Abstract

Compared with Chinese and English, the development of Burmese speech synthesis technology is relatively lagging, and the synthesized speech lacks emotion. Emotional speech synthesis makes

*通讯作者。

文章引用: 刘奇云, 杨鉴, 谭婉琳. 基于语音参数自适应的缅甸语情感语音合成[J]. 计算机科学与应用, 2022, 12(1): 33-45. DOI: 10.12677/csa.2022.121005

the machine's expression not reproducible. Using the HMM acoustic model-based speech parameter adaptation method, the Burmese emotional speech synthesis is studied. One of the difficulties faced by the research of emotional speech synthesis is that it is difficult to obtain a large-scale emotional speech library. Under the condition of low resources, a method to realize the emotional speech synthesis of Burmese is proposed. Firstly, the Myanmar phonetic sub-segmentation is carried out on the MFA (Montreal Force Align) platform to train the voice acoustic model. Based on the HTS platform, a medium-scale Burmese calm emotion speech library is used to construct a Burmese speech synthesis baseline system. On this basis, based on a small amount of happy, sad, and angry emotional speech data, the Burmese emotional speech synthesis system is constructed using the method of speech parameter adaptation, and the emotional speech synthesis system is further improved by introducing the average sound model and adjusting the conversion matrix. Experimental results show that the emotional speech synthesis system can synthesize Burmese speech with four emotions: calm, happy, sad, and angry, with an average EMOS score of 3.40, which proves the effectiveness of the method.

Keywords

Burmese, MFA Automatic Phoneme Segmentation, Emotional Speech Synthesis, Speech Parameter Adaptation

1. 引言

随着语音合成技术的发展，汉语和英语的语音合成的自然度与可懂度有较大的提升，在语音转换、说话人迁移、情感语音合成等多样化语音合成方面有较大的发展与突破。语音合成是实现人机交互的重要环节，近年来端到端技术在语音合成上有较大的应用，合成语音的质量有明显的提升。在提升合成语音质量的前提下，建立实现任意说话人特征、任意情感特征的语音合成系统对于提升语音合成的自然度具有重要意义。

缅甸语是缅甸各民族的共同语，属于汉藏语系藏缅语族缅语支，使用人口超过 4800 万[1]。中缅两国长期以来建立了良好的战略合作伙伴关系，在“一带一路”战略背景下，中缅两国人民展开了广泛的经贸交流与合作，缅甸对于我国外交具有重要的战略地位。针对缅甸语的语音合成，有部分学者提出了不同的方法实现缅甸语的语音合成，如 2015 年 Ye 等人提出采用基于 HMM 的统计参数的方法实现缅甸语语音合成[2]，2017 年 Thida 等人提出基于音子拼接的方法[3]，2018 年 Hlaing 等人采用基于 DNN 的方法进行缅甸语语音合成研究[4]，但主要的研究还是集中于合成出中性情感的缅甸语音，合成的声音缺乏情感表现力。为了使合成出的缅甸语声音具有情感，更好地促进人机交流，本文进行缅甸语情感语音合成研究。

相比缅甸语，汉语和英语等通用语言的情感语料库获取更为方便，针对汉语和英语的情感语音合成研究，许多学者采用构建大规模情感语料库的方法进行特定情感的建模训练。近年来深度学习技术在情感语音研究上有了一定的应用，如 2018 年 Gao 等人将 GAN (生成对抗网络)深度学习框架应用于情感语音转换上取得了较好的实验效果[5]，但许多深度学习网络要得到较好的实验结果需要一定规模的语料库，而缅甸语等非通用语的电子化资源相对匮乏，难以获取较大规模的高质量特定情感语料库，且构建一份大规模的情感语料库所需成本较高，时间周期较长，因此难以将深度学习技术直接应用于缅甸语的情感

语音合成。

针对以上问题，本文采用基于 HMM 的语音合成技术并结合语音参数自适应的方法实现缅甸语的情感语音合成，并在 MFA (Montreal Force Align)平台进行音子强制对齐训练[6]，完成缅甸语文本分词、注音、音子自动切分、情感语音合成训练等工作。

本文的主要贡献为在构建缅甸语小规模的情感语料库条件下，实现不同情感特征的语音合成，采用 MFA 音子强制对齐训练方法，完成缅甸语的音子自动切分，改进了情感语音合成系统，并调整合成语音的情感状态。本文接下来的内容组织结构如下：第二部分完成了缅甸语情感语料库的构建；第三部分介绍本文的语音合成系统，并完成文本分析和音子自动切分工作；第四部分为对情感语音合成系统进行改进；第五部分进行缅甸语情感语音合成实验，并对所得到的实验结果进行评测与分析；最后一部分为本文的结束语。

2. 情感语音库的构建

本文采用自主构建语料库的方式进行情感语音合成研究，所需的语料库包括平静情感的语料库和请缅甸语专业播音员采用情感表演的方式录制的高兴、悲伤、生气情感语料库。语料库的质量对语音合成系统的性能有着较大的影响，本文所选用的平静情感的语料库综合考虑缅语文本长度、句子相似度、高频词等因素，由于构建平静情感的语料库为前期已完成工作，在此不进行过多赘述。

2.1. 挑选发音语料

由于构建一份大规模的高兴、悲伤、生气情感的语料库需要较高的时间与经济成本，因此本文在中规模的平静情感语料库的基础上构建一份小规模的高兴、悲伤、生气情感语料库，用于进行情感自适应。

本文邀请缅甸语专业播音员通过情感表演的方式录制情感语音进行情感语料库的构建，采用这样的方式可使情感表达更逼真，情感辨析度更高。为使播音员在录制情感语音时情感表达更自然，提升情感语料库的质量，本文先请缅甸语专业人员对缅甸语文本进行标注，挑选出不同方面的表示高兴、悲伤、生气情感色彩的缅语文本，由于句子的长度对语料库的质量有较大的影响，太长的文本在进行情感语音录制的时候会影响缅甸语播音员的情感表达，因此本文舍弃太长的缅语文本，最终得到的高兴、悲伤、生气情感色彩的缅语文本各 100 句，本文所选取的句子长短错落有致，以缅甸语音节数为统计单位，得到的待录制的缅甸语带情感的句子长度分布如图 1 所示。

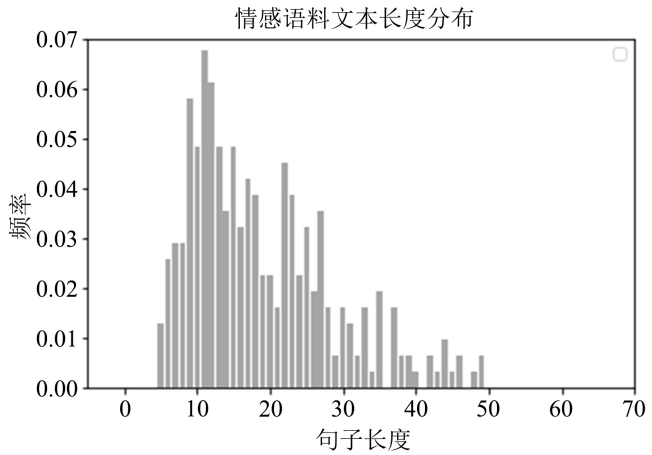

Figure 1. Sentence length distribution
图 1. 句子长度分布

经过文本校对处理，修正句子中存在的错误，再进行情感语音录制，使缅甸语播音员分别对表示高兴、悲伤、生气的句子采用对应的情感表达方式进行表演录音。相比对不带情感的中性文本采用表演进行情感语音录制的方式，播音员在进行录音的时候情感表达更自然，录制的情感语音更逼真。

2.2. 语音库的处理

在对挑选出的缅甸语情感文本采用表演的方式录制完情感语音后，需要对所录制的音频进行处理。在完成情感语音录制得到MP3格式的音频后，首先需要将MP3格式的音频通过音频处理平台转换为Wav格式音频，将每段音频前后置留0.5 s (秒)的静音段，再将音频采样率设置为48,000 Hz和16位单声道，经过校对处理得到每一段音频的<文本－语音>对。

本文的缅甸语情感语音合成系统包括平静、高兴、悲伤、生气四种基本情感，最终进行情感语音合成实验所用的语料库包括平静情感的4868句<文本－语音>对，以及高兴、悲伤、生气情感语音各100句<文本－语音>对。本文所构建的情感语料库的语音质量较高，情感表达清晰，具有较高的情感辨识度，可作为后续缅甸语情感语音合成研究的语料库。

3. 语音合成系统

3.1. 缅甸语文本分析

本文采用隐马尔可夫模型(Hidden Markov Model, HMM)作为声学模型进行声学参数建模，进行基于统计参数的情感语音合成训练[7]。传统的语音合成系统主要由文本分析和语音合成两部分组成[7]，本文不直接对缅甸语的文本字符进行语音合成训练，先将缅甸语文本进行分词、罗马化注音等文本分析工作，再将注音文本进行转换得到缅甸语韵律文本，本文所采用的缅甸语文本分析流程如图2所示。

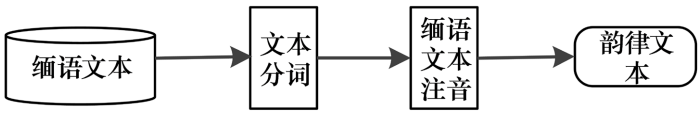

Figure 2. Text analysis process
图 2. 文本分析流程

首先将缅语文本去除文本中的非法字符等内容，由于缅甸语词与词之间没有空格表示，因此为进行后续的基于规则的缅甸语罗马化注音工作，需要进行缅甸语分词，本文针对缅语文本采用基于CRF条件随机场的分词算法[8]。

对经过CRF分词模型得到分词后的缅语文本再进行缅甸语罗马化注音，将得到的注音文本通过Python程序转化为后端情感语音合成训练所需的韵律文本。本文针对缅甸语文本结合注音词典采用基于规则的方式进行注音，基于规则的文本注音方案通过结合缅英词典进行，在得到基于CRF模型的分词文本后，对每句分词文本中的每个缅文词进行词典检索，若注音词典中存在的词则按照词典中的注音对缅文词进行文本注音。经过分词后的缅语文本中可能存在词典中不存在的词，即未登录词，以及缅甸语中存在一些叠字及变音变调的现象，对这部分则采用处理未登录词、叠字、变音变调的相应规则进行注音。

得到注音文本后，再将注音文本转为韵律文本。如下所示为一句缅语文本：

သုံးနာရီတောင်ထိုးနေပြီ၊ မြန်မြန်လုပ်ပါလား။

采用本文的方法得到的注音文本为：

thoun2 na3ji3 taun3htou2 nei3 pji3 mjan3mjan3 lout4pa3 la2

转换得到的韵律文本为：

thoun2*na3ji3*taun3htou2*nei3*pji3*mjan3mjan3*lout4pa3*la2

[th oun2][n a3/j i3][t aun3/ht ou2][n ei3][pj i3][mj an3/mj an3][l out4/p a3][l a2]

3.2. 音子强制对齐

进行声学模型训练前，需要提供带时间标注信息的训练数据，本文选用声韵母作为缅甸语合成基元，因此需要得到缅甸语声韵母持续时间的标注信息。

本文针对缅甸语的音子自动切分，采用在 MFA 平台下将平静情感语音和目标合成情感语音数据同时进行音子强制对齐训练。MFA 强制对齐平台基于 Kaldi 实现，采用三音子进行建模，适用于不同说话人的语音数据共同进行音子对齐训练[6]，因此本文在 MFA 平台下进行缅甸语音子强制对齐实验。如下所示为本文强制对齐所用的一句注音文本训练数据样例：

ta3ka3te2min2 ka1 le2 ba3 lou1 e2lou3 lout4 ja1 ta3 le2

注音后的缅语文本词与词之间通过空格分开，每个缅文词中的声韵母为本文需要得到的音子对齐信息。为进行后续的强制对齐训练，本文首先构建缅甸语的发音词典，对分词后的缅语文本完成文本注音后，可以得到每一个缅语词的对应注音，通过生成韵律文本得到每个词的声韵母信息，最后结合缅文词的对应注音和声韵母切分得到缅甸语的发音词典，其中包括训练数据中每个缅文词的文本注音及声韵母切分。本文最终构建的缅甸语发音词典包含 7000 余个缅甸语注音词及对应的声韵母，其中部分发音词典如表 1 所示，缅甸语是一种带声调的文字[9]，表中数字表示不同声调。

Table 1. Burmese pronunciation dictionary
表 1. 缅甸语发音词典

缅语词注音	声韵母
hpaut4the3cha1	hp aut4 th e3 ch a1
khet4tha2	kh et4 th a2
a3naut4hpet4	a3 n aut4 hp et4

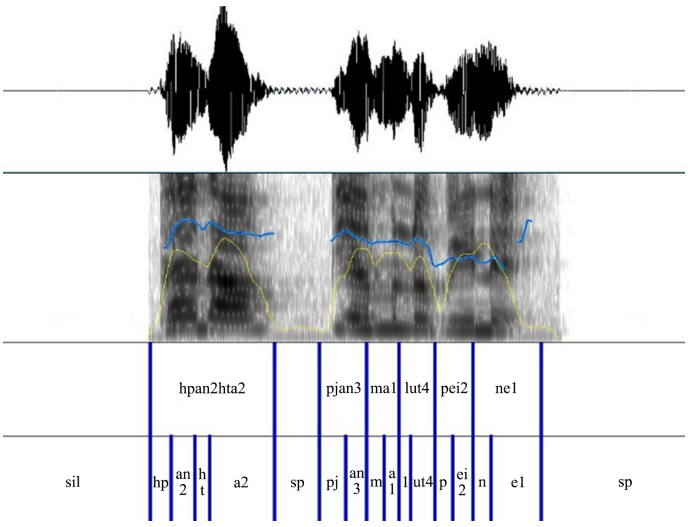

Figure 3. MFA Force alignment result
图 3. MFA 强制对齐结果

本文的强制对齐训练数据包括 4868 句平静情感<文本－语音>对，100 句高兴、悲伤、生气情感<文本－语音>对，再结合本文构建的缅甸语发音词典进行强制对齐训练。最终训练得到的结果包括每一句缅甸语注音文本的词边界信息和音子边界信息。

图 3 所示为对一句缅语文本采用 MFA 平台得到的强制对齐结果，从上到下依次为语音波形图、语谱图及对齐结果。

通过结合语音波形图和语谱图分析缅语文本的强制对齐结果可以得出，采用 MFA 平台进行的缅甸语声韵母自动切分所得到的正确率较高，音子边界切分清晰，因此本文将采用 MFA 平台得到的音子切分结果作为后续情感语音合成实验的训练数据。

3.3. 情感语音参数自适应

缅甸语是一种低资源语言，相比汉语和英语等通用语言构建语料库的成本较高，且难以针对不同情感的语音构建大规模的情感语料库。在情感语料库规模有限的情况下采用深度学习或者传统的基于统计参数的情感语音合成方法难以获得较好的实验结果，本文在缅甸语的低资源条件下，采用基于 HMM 的 Adaptation 自适应训练的方法所需的情感语料库的数据规模相对较小[10] [11]，这种方法基于某种特定的目标情感进行语音参数自适应训练，适用于缅甸语的低资源语言特点。

3.3.1. 系统结构

平静、高兴、悲伤、生气等不同情感的语音在发音速率、基频、谱参数、音强方面具有不同的声学特征参数表现，本文在平静情感语音的基础上对基频、谱参数和时长参数进行语音参数自适应调整。本文采用最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)作为自适应算法合成高兴、悲伤、生气的语音[12]，MLLR 算法根据平静情感的语料库与目标合成情感语音的语料库得到两者之间的差异，训练得到自适应转换矩阵。

本文基于 HMM 的语音参数自适应的缅甸语情感语音合成系统主要包括两个阶段，一是根据平静情感的语料库训练平静的语音合成系统，二是在合成平静语音的基础上引入语音参数自适应，合成不同情感特征的语音，系统框架如图 4 所示。

首先将平静情感的语料库单独进行训练，得到一个平静情感的缅甸语语音合成基线系统。在平静情感的语音合成系统基础上引入语音特征参数自适应，将需要合成的高兴、悲伤、生气特定缅甸语目标情感语音作为自适应训练目标，进行语音参数自适应变换。本文将自主构建的高兴、悲伤、生气的 100 句<文本－语音>对情感语料库作为自适应目标，训练得到缅甸语情感语音合成系统。

3.3.2. 情感自适应

在训练得到平静情感的语音合成系统后，利用目标合成情感的小规模语音数据进行情感自适应变换，训练得到缅甸语情感语音合成系统。采用 MLLR 算法进行情感自适应训练对目标合成情感语音的语料库规模没有特别严格的要求，使用较少规模的数据量也可得到相对较好的情感语音合成实验效果[10] [11]，相比采用深度学习的方法较大地减少了语料库规模的需求。本文对基频参数、谱参数与时长参数根据目标情感进行自适应变换。基于 MLLR 的自适应算法中，对基线系统的状态输出分布和时长分布进行线性变换来获得目标情感语音的状态和时长的分布，在状态为 i 时，状态矢量分布 $b_i(\boldsymbol{o})$ 如式(1)所示，时长分布 $p_i(d)$ 如式(2)所示：

$$b_i(\boldsymbol{o}) = \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\zeta}\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}, \Sigma_i\right) = \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{W}\boldsymbol{\xi}_i, \Sigma_i\right) \tag{1}$$

$$p_i(d) = \mathcal{N}\left(d; \chi m_i + v, \sigma_i^2\right) = \mathcal{N}\left(d; \boldsymbol{X}\boldsymbol{\phi}, \sigma_i^2\right) \tag{2}$$

其中，$\boldsymbol{W}=[\boldsymbol{\zeta},\boldsymbol{\varepsilon}]\in\mathcal{R}^{L\times(L+1)}$ 和 $\boldsymbol{X}=[\chi,V]\in\mathcal{R}^{1\times2}$ 分别表示从基线模型到缅甸语目标情感的状态和时长分布的转换矩阵，$\boldsymbol{\xi}_i=\left[\boldsymbol{\mu}_i^T,1\right]^T\in\mathcal{R}^{L\times1}$ 和 $\boldsymbol{\phi}_i=\left[m_i,1\right]^T\in\mathcal{R}^2$ 分别为状态和时长的扩展均值向量[12]，平静情感向目标合成情感语音的变化主要通过转换矩阵实现。

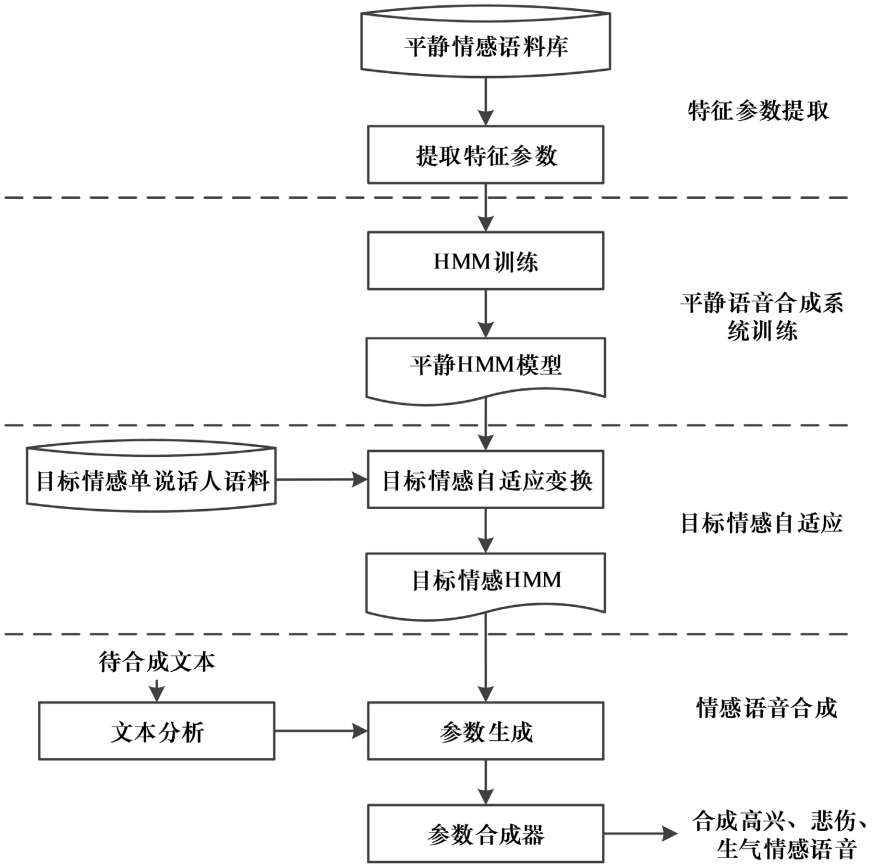

Figure 4. Emotional speech synthesis system framework
图 4. 情感语音合成系统框架

4. 情感语音合成系统的改进

4.1. 引入平均音模型

本文在平静情感语音合成系统的基础上采用语音参数自适应的方法合成高兴、悲伤、生气的情感语音。本文引入平均音模型来训练平静情感语音合成系统，平均音模型将不同说话人的语音数据共同进行训练，得到说话人无关的语音合成系统。本文的平均音模型如图 5 所示。

本文将平静情感语音数据与所需合成的目标情感语音数据共同训练，进行决策树聚类得到上下文相关 HMM 模型，训练得到平均音模型，并在此基础上合成平静情感的语音。由于训练平均音模型所使用的目标合成情感语音数据规模较小，训练得到的平均音模型合成的平静情感语音可保持与采用平静情感语音数据单独进行 HMM 模型训练所得到的语音合成系统高度相似的语音。

在得到平均音模型的基础上，通过语音参数自适应的方式合成高兴、悲伤、生气情感的语音，本文通过实验对比采用 HMM 基线模型和平均音模型合成平静情感语音的质量，并在此基础上进行语音参数自适应对比合成高兴、悲伤、生气情感语音的质量。

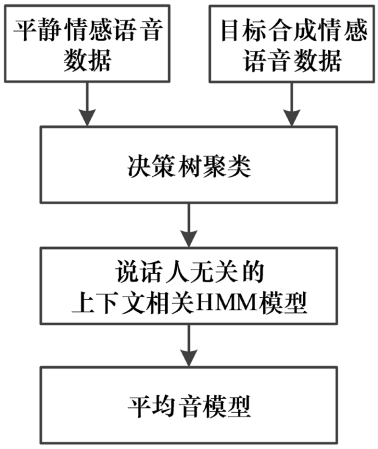

Figure 5. Average-voice Model
图 5. 平均音模型

4.2. 调整转换矩阵

本文采用情感分类的方式描述缅甸语情感，训练得到的情感语音合成系统包括四种类型的情感，但人类的情感表达是极度丰富和细腻的，即使在同一情感类别下，也有不同的情感状态差异。本文的情感语音合成系统通过训练得到的不同转换矩阵进行语音参数自适应合成高兴、悲伤、生气的情感语音，合成高兴、悲伤、生气的语音由平静向目标合成情感语音的转换矩阵实现，通过调整转换矩阵可以改变状态矢量和时长分布，进而调整生成的声学特征参数以合成不同情感特征的语音。

通过不同的目标合成情感语音数据进行自适应训练，可得到不同的状态和时长分布的转换矩阵。本文通过调整转换矩阵使合成的情感语音在平静情感与目标情感之间进行变换，具体的调整思路为通过将平静情感转换矩阵与目标情感转换矩阵赋予不同的权重值，两者的权重值相加等于 1，然后采用矩阵线性相加的方式调整转换矩阵。由于本文所训练的语音合成基线系统采用平静情感语料，基线系统所合成的声音即为平静情感状态，因此在平静情感状态的基础上向平静情感转换可采用单位矩阵与 0 偏置向量，而目标合成情感语音的转换矩阵则通过语音参数自适应训练得到。

由于本文的语音调整参数包括 F0 参数、谱参数和时长参数，因此本文对影响 F0 参数、谱参数和时长参数的状态和时长分布的转换矩阵均进行调整，具体的调整方式如式(3)和(4)所示：

$$\vec{\boldsymbol{W}}_j = (1-\eta)\boldsymbol{E}_w + \eta \boldsymbol{W}_j \tag{3}$$

$$\vec{\boldsymbol{X}}_j = (1-\eta)\boldsymbol{E}_x + \eta \boldsymbol{X}_j \tag{4}$$

其中 $\vec{\boldsymbol{W}}_j$ 和 $\vec{\boldsymbol{X}}_j$ 分别为状态和时长分布调整后的转换矩阵，$\boldsymbol{E}_w$ 和 $\boldsymbol{E}_x$ 分别为平静情感的状态和时长的转换矩阵，由单位矩阵和 0 偏置向量组成。$\boldsymbol{W}_j$ 和 $\boldsymbol{X}_j$ 分别为高兴、悲伤、生气情感的状态和时长转换矩阵，η 为转换矩阵的权重系数。在确定调整转换矩阵方式后，通过设置不同的转换矩阵权重系数进而调整合成情感语音。

5. 实验

5.1. 实验方案

5.1.1. 实验数据集

本文进行缅甸语情感语音合成实验所使用的语料库为自主构建的缅甸语情感语料库，其中包括 4868

句平静情感的<文本－语音>对，以及请缅甸语专业播音员采用表演方式录制的高兴、悲伤、生气情感语音各 100 句<文本－语音>对。

5.1.2. 训练与调整矩阵方案

HMM 建模参数配置如下：

音频参数设置：采样率为 48,000 Hz，16 位单声道，每段音频时长为 2~10 s，前后置留约 0.5 s 的静音段。

建模单元：以缅甸语声韵母为建模单元，采用上下文相关的三音子模型。

建模方式：对于频谱参数采用连续概率分布的建模方式，对于 F0 参数，由于在浊音部分有值，而清音部分没有值，则采用 MSD (多空间概率分布)对 F0 基频参数进行统一建模[13]。

对于合成为平静情感状态的语音，首先将平静情感的语料库进行传统 HMM 训练得到平静情感的语音合成基线系统，对于合成为高兴、悲伤、生气情感状态的语音，采用目标情感自适应的方式。在此基础上，通过将平静情感语音和目标合成语音一起训练平均音模型，用平均音模型合成平静情感语音，再通过目标合成情感语音进行自适应训练合成高兴、悲伤、生气的语音。本文所采用的训练策略如表 2 所示。

Table 2. Training strategy
表 2. 训练策略

训练数据集	训练模型	自适应目标	合成情感语音
4868 句平静情感语音	HMM 基线模型 平均音模型	无	平静
100 句高兴情感语音	HMM 基线模型 + 情感自适应 平均音模型 + 情感自适应	100 句高兴情感语音	高兴
100 句悲伤情感语音	HMM 基线模型 + 情感自适应 平均音模型 + 情感自适应	100 句悲伤情感语音	悲伤
100 句生气情感语音	HMM 基线模型 + 情感自适应 平均音模型 + 情感自适应	100 句生气情感语音	生气

完成缅甸语情感语音合成系统的构建后，进行转换矩阵调整，将平静情感转换矩阵与目标合成情感语音转换矩阵设置不同的权重值。本文在平均音模型所进行自适应训练所得到的转换矩阵的基础上进行转换矩阵的调整，并对比调整矩阵前后的合成语音情感表达。具体的转换矩阵权重系数设置如表 3 所示。

Table 3. Conversion matrix weight coefficient
表 3. 转换矩阵权重系数

矩阵训练方式	平均音模型 + 情感自适应				
权重系数	0.1	0.3	0.5	0.7	0.9

5.2. 实验结果评测与分析

本文对合成的缅甸语情感语音采用客观评测和主观评测相结合的方式，分析合成的声音效果。

5.2.1. 合成情感语音客观评测

图 6 和图 7 为对同一句缅甸语文本分别采用平静和生气的方式合成语音的语音波形图和语谱图，语谱图中蓝色曲线表示基频的变化，黄色曲线表示音强。

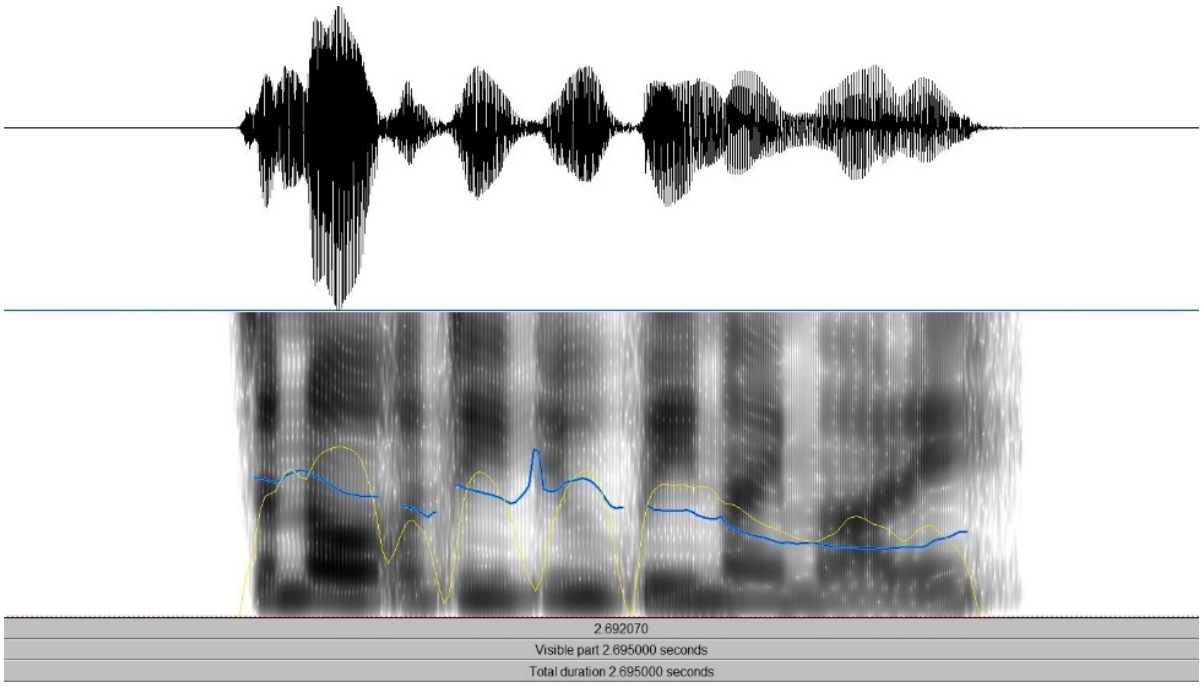

Figure 6. Synthesized calm emotion speech
图 6. 合成平静情感语音

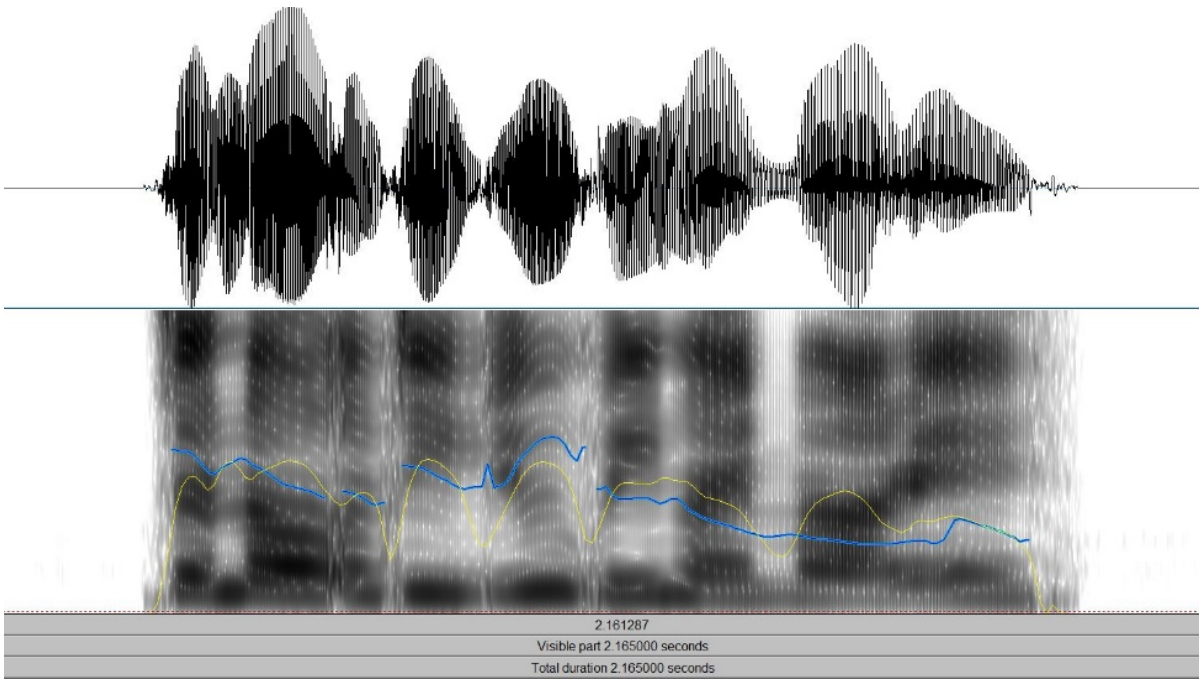

Figure 7. Synthesized angry emotion speech
图 7. 合成生气情感语音

通过观察语音波形图和语谱图可以得出，合成的语音音子边界比较清晰，对同一句文本采用不同情感的合成方式，在平静情感状态下合成的语音语速正常适中，语音的音强相较生气情感更弱，而在生气情感状态下，语音的音强在部分音子处表现更强烈，在生气情感状态下的语速相较于平静情感下更快，符合不同情感状态下的声学特征表现，证明了情感状态的变化。

5.2.2. 合成情感语音主观评测

只采用客观评测的方式具有一定的局限性，情感语音合成系统最终合成的语音质量还依赖于人耳所听到的主观感受，因此本文对采用不同训练方式合成的高兴、悲伤、生气、平静情感语音各 100 句进行主观评测打分，常用的语音主观评测方式为平均意见得分(Mean Opinion Score, MOS) [14] [15]，本文对合成的语音采用 EMOS (情感平均意见得分)的评测方式进行主观评分，其对应的情感表达程度及评测等级如表 4 所示，所得到的评测结果如表 5 所示。

Table 4. Emotional subjective evaluation level
表 4. 情感主观评测等级

评测等级	合成语音的情感表达程度
5 分	情感表达真实强烈，具有较高的可懂度和自然度
4 分	情感表达比较好，语句自然，声音能分辨
3 分	情感有些区分度，语句较自然，声音能分辨
2 分	情感基本无法分辨，非常不自然，存在区别
1 分	毫无情感表述，声音混沌杂乱，无区别度

Table 5. Evaluation results
表 5. 评测结果

情感语音	训练模型	平均分
平静	传统 HMM	3.42
	平均音模型	3.42
高兴	传统 HMM + Adaption	3.30
	平均音模型 + Adaption	3.35
悲伤	传统 HMM + Adaption	3.34
	平均音模型 + Adaption	3.40
生气	传统 HMM + Adaption	3.35
	平均音模型 + Adaption	3.41

通过采用主观评测的方式可以看到，本文所构建的情感语音合成系统合成的语音具有区分度，声音能分辨，语句较自然，所合成的平静情感的语音效果相较于高兴、悲伤、生气情感相对更好，通过引入平均音模型，合成的高兴、悲伤、生气的情感语音质量有所提高，合成语音的 EMOS 平均评分为 3.40。

在引入平均音模型进行情感自适应所得到的高兴、悲伤、生气情感语音合成系统基础上，通过调整转换矩阵，对不同转换矩阵下所合成的语音进行主观听测分析，通过设置不同的转换矩阵权重系数可以影响合成的情感语音，进而调整情感状态。当平静情感转换矩阵权重系数较大时，合成的情感语音更偏向于平静，当高兴、悲伤、生气情感转换矩阵权重系数较大时，合成的语音则更偏向高兴、悲伤、生气情感。虽然部分合成音频声音表达不太自然，情感表达有待加强，但证明了本文方法的有效性。

6. 结束语

本文围绕缅甸语情感语音合成研究，构建了一份包含高兴、悲伤、生气情感的缅甸语情感语料库，在 MFA 实验平台提出了一种缅甸语的音子强制对齐训练方法，并在所构建的规模较小的情感语料库的情况下采用基于 HMM 的语音参数自适应训练的方法实现了缅甸语情感语音合成，并通过引入平均音模型和调整转换矩阵来改进情感语音合成系统。在低资源的条件下能够感受到情感的表达与变化，合成的语音包括平静、高兴、悲伤、生气四种情感，基本达到情感语音合成的要求，证明了本文方法的可行性。

本文所构建的缅甸语情感语音合成系统只包括四种类型的情感，但是人类的情感是极度丰富和复杂的，还包括骄傲、痛苦、轻松等不同类别的情感，本文所构建的缅甸语情感语音合成系统合成的情感类别相对有限。对于合成的高兴、悲伤、生气情感的语音质量，仍有进一步提升的空间，后续可进一步扩充合成情感语音的类别，对于合成语音，可采用更有效的声学建模方式以提升合成语音的质量。

基金项目

国家自然科学基金(61961043)。

参考文献

[1] 钟智翔, 尹湘玲. 基础缅甸语(第一册) [M]. 广州: 世界图书出版广东有限公司, 2012.

[2] Thu, Y.K., Pa, W.P., Ni, J., Shiga, Y., Finch, A., Hori, C., et al. (2015) HMM Based Myanmar Text to Speech System. 16*th Annual Conference of the International Speech Communication Association*, Dresden, 6-10 September 2015, 2237-2241. https://doi.org/10.21437/Interspeech.2015-132

[3] Hlaing, C.S. and Thida, A. (2017) Myanmar Speech Synthesis System by Using Phoneme Concatenation Method. 2017 *International Conference on Signal Processing and Communication* (ICSPC), Coimbatore, 28-29 July 2017, 399-404. https://doi.org/10.1109/CSPC.2017.8305878

[4] Hlaing, A.M., Pa, W.P. and Ye, K.T. (2018) DNN Based Myanmar Speech Synthesis. *The* 6*th Workshop on Spoken Language Technologies for Under-Resourced Languages*, Gurugram, 29-31 August 2018, 142-146. https://doi.org/10.21437/SLTU.2018-30

[5] Gao, J., Chakraborty, D., Tembine, H. and Olaleye, O. (2019) Nonparallel Emotional Speech Conversion. 20*th Annual Conference of the International Speech Communication Association*, Graz, 15-19 September 2019, 2858-2862. https://doi.org/10.21437/Interspeech.2019-2878

[6] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. and Sonderegger, M. (2017) Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech* 2017: *Conference of the International Speech Communication Association*, Stockholm, 20-24 August 2017, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

[7] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K. (2013) Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE*, **101**, 1234-1252. https://doi.org/10.1109/JPROC.2013.2251852

[8] Ma, C.E. and Yang, J. (2018) Burmese Word Segmentation Method and Implementation Based on CRF. 2018 *International Conference on Asian Language Processing* (*IALP*), Bandung, 15-17 November 2018, 340-343. https://doi.org/10.1109/IALP.2018.8629163

[9] 汪大年. 缅甸语教程[M]. 北京: 北京大学出版社, 2012.

[10] 吴义坚. 基于隐马尔科夫模型的语音合成技术研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2006.

[11] Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T. (2001) Adaptation of Pitch and Spectrum for HMM-Based Speech Synthesis Using MLLR. *Proceedings of* 2001 *IEEE International Conference on Acoustics, Speech, and Signal*

Processing, Salt Lake City, 7-11 May 2001, 805-808. https://doi.org/10.1109/ICASSP.2001.941037

[12] Yamagishi, J. (2006) Average-Voice-Based Speech Synthesis. Tokyo Institute of Technology, Tokyo.

[13] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T. (1999) Hidden Markov Models based on Multi-Space Probability Distribution for Pitch Pattern Modeling. *Proceedings of* 1999 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, 15-19 March 1999, 229-232. https://doi.org/10.1109/ICASSP.1999.758104

[14] Viswanathan, M. and Viswanathan, M. (2005) Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale. *Computer Speech & Language*, **19**, 55-83. https://doi.org/10.1016/j.csl.2003.12.001

[15] Streijl, R.C., Winkler, S. and Hands, D.S. (2016) Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems*, **22**, 213-227. https://doi.org/10.1007/s00530-014-0446-1