# 基于注意力的蒙古语说话人特征提取方法

朱方圆 [1], 马志强 [1,2*], 刘志强 [1], 宝财吉拉呼 [1], 王洪彬 [1]
[1] 内蒙古工业大学, 呼和浩特, 010000
[2] 内蒙古自治区基于大数据的软件服务工程技术研究中心, 呼和浩特, 010000
mzq_bim@imut.edu.cn

## 摘要

说话人特征提取模型提取到的说话人特征之间区分性低，使得蒙古语声学模型无法学习到区分性信息，导致模型无法适应不同说话人。提出一种基于注意力的说话人自适应方法，方法引入神经图灵机进行自适应，增加记忆模块存放说话人特征，采用注意力机制计算记忆模块中说话人特征与当前语音说话人特征的相似权重矩阵，通过权重矩阵重新组合成说话人特征 s-vector，进而提高说话人特征之间的区分性。在 IMUT-MCT 数据集上，进行说话人特征提取方法的消融实验、模型自适应实验和案例分析。实验结果表明，对比不同说话人特征 s-vector、i-vector 与 d-vector，s-vector 比其他两种方法的 SER 和 WER 分别降低 4.96%、1.08%；在不同的蒙古语声学模型上进行比较，提出的方法相对于基线均有性能提升。

**关键词：** 说话人特征提取；注意力机制；神经图灵机；说话人自适应；蒙古语语音识别

# Attention based Mongolian Speaker Feature Extraction

**Zhu Fangyuan[1], Ma Zhiqiang[1,2*], Liu Zhiqiang[1], Bao Caijilahu[1], Wang Hongbin[1]**
[1] Inner Mongolia University of Technology, Huhhot, 010000
[2] Inner Mongolia Autonomous Region Software Service Engineering Technology Research Center Based on Big Data, Huhhot, 010000
mzq_bim@imut.edu.cn

## Abstract

The speaker features extracted by the speaker feature extraction model have low discrimination, which makes the Mongolian acoustic model unable to learn the discrimination information, resulting in the model unable to adapt to different speakers. A speaker adaptation method based on attention is proposed. The method introduces neural Turing machine for adaptation, adds a memory module to store the speaker features, uses the attention mechanism to calculate the similarity weight matrix between the speaker features in the memory module and the current voice speaker features, and recombines the weight matrix into the speaker features s-vector, so as

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。

to improve the discrimination between the speaker features. On the IMUT-MCT dataset, the ablation experiment, model adaptation experiment and case analysis of speaker feature extraction method are carried out. The experimental results show that comparing s-vector, i-vector and d-vector with different speaker characteristics, s-vector reduces SER and WER by 4.96% and 1.08% respective-ly compared with the other two methods; By comparing different Mongolian acoustic models, the performance of the proposed method is improved compared with the baseline.

**Keywords:** Speaker Feature Extraction , Attention Mechanism , Neural Turing Machine , Speaker Adaptation , Mongolian speech recognition

## 1 引言

蒙古语说话人自适应 (Mongolian Speaker Adaptation, MSA) 方法是解决训练数据和测试数据中的说话人不匹配问题。根据自适应的对象是特征还是模型，将说话人自适应方法分为基于模型域的和基于特征域的说话人自适应方法 (Bell et al., 2020)。在基于模型域的说话人自适应方法中，Stadermann and Rigoll (2005) 提出基于模型参数的说话人自适应方法，是只更新声学模型的部分参数，这些参数通常兼具鲁棒性和有效性。Samarakoon and Sim (2016) 提出对隐藏层进行分解，分解后的奇异值分解层插入线性层。利用奇异值分解方法对模型权重矩阵进行更新，比直接插入线性层的方法更能减少参数量，减轻过拟合问题。Swietojanski and Renals (2014) 将给定自适应数据的情况下学习说话人特定的隐藏单元分布，为解决在只对神经网络的某一层输出特征进行变换时，对所有层的输出特征均进行变换导致自适应参数数量成倍增加的问题。Dong et al. (2013) 在声学模型自适应过程中增加正则化项，通过限制原始模型和调整后的模型之间的距离，进而防止调整后的模型参数偏离原始模型参数。根据使用方法的不同将基于特征域的说话人自适应方法分为基于特征变换和基于辅助特征的说话人自适应 (朱方圆等, 2021)。在基于特征变换的说话人自适应方法中，Neto et al. (1995) 提出对输入特征进行线性变换的方法，通过对神经网络的输入特征或者隐藏层特征进行变换来实现自适应。在基于辅助特征的说话人自适应方法中，Saon et al. (2013) 通过在声学特征中添加特定说话人信息并利用新特征训练声学模型实现自适应。Abdel-Hamid and Jiang (2013) 提出说话人编码方法，即给定一个说话人编码，将每一个说话人的特征映射到一个说话人无关的特征空间，对于一个新的说话人学习其相应的说话人编码是容易的，且用反向算法时不会改变神经网络的模型参数。其中基于辅助特征的说话人自适应方法是极为重要的方法，辅助特征包括 i-vector (Saon et al., 2013)、d-vector (Variani et al., 2014) 等。该类方法直接提取训练集的辅助特征和声学特征拼接在一起进行训练，提高声学模型的适应性。测试时，提取测试集的辅助特征和声学特征拼接在一起进行测试。

在蒙古语中词根和后缀的存在差异造成蒙古语读音存在差异，而且不同内蒙地区的说话人发音特点也存在差异，因此不同地区说话人口音包含明显的说话人特性信息，所以提取具有区分性的蒙古语说话人特征是蒙古语说话人自适应方法的难点。目前区分性的蒙古语音频属于少量有标记数据，进而提取的区分性说话人特征较少，使用大量区分性很差的说话人特征和少量具有区分性的说话人特征进行声学模型训练，使得蒙古语声学模型学习到少量的区分性信息，导致训练的蒙古语声学模型自适应效果较差。然而基于辅助特征的说话人自适应方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

在蒙古语语音识别 (Mongolian Speech Recognition, MSR) 中采用神经图灵机 (Neural Turning Machine, NTM) (Graves et al., 2014) 进行说话人自适应，通过增加记忆模块来存放说话人特征信息，使用用注意力机制来计算记忆模块的说话人特征与当前语音的说话人特征的相似权重矩阵，将记忆模块中相似的说话人特征通过权重矩阵进行重新组合成一个说话人特征向量 (即 s-vector)，提高各内蒙地区说话人特征之间的区分性，蒙古语声学模型利用区分性说话人特征减小说话人的差异性。本文贡献为：(1) 设计一种基于注意力的说话人特征提取方法，提高说话人特征之间的区分性；(2) 在 IMUT-MCT 数据集上进行验证实验，比较不同说话人特征提取方法，将 s-vector 作为说话人特征来附加声学特征上，验证比直接使用 i-vector 获得更好的性能。

## 2 相关工作

基于辅助特征的说话人自适应方法是基于特征域的说话人自适应方法中一种重要方法，通过在声学特征中添加特定说话人信息组成新特征，然后利用新特征训练声学模型，进而提高模型的自适应性。按照使用辅助特征的不同，将基于辅助特征的说话人自适应方法分为基于 i-vector 的改进方法和基于 d-vector 的改进方法。

基于 i-vector 的改进方法是在利用说话人特征 i-vector 的基础上改进，进而提高模型的适应性的方法。Saon et al. (2013) 通过将说话人身份向量 i-vectors 和语音识别的声学特征同时作为网络的输入特征，使深度神经网络 (Deep Neural Network, DNN) 声学模型适应目标说话人。Cardinal et al. (2015) 提出将瓶颈 (Bottleneck, BN) 特征，BN 特征提取通过使用 DNN 从语音语料库中提取区分性特征来完成，在阿拉伯语广播新闻任务上词错误率降低了 1.2%。Cui et al. (2017) 提出一种基于辅助特征的说话人自适应训练方法，将说话人特征向量 i-vector 通过控制网络映射到每个隐藏层的仿射变换，以规范化主网络隐藏层输出处的内部特征表示，比使用具有说话人适应输入特征的 i-vector 方法具有更好的性能。然而基于 i-vector 的说话人自适应方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

基于 d-vector 的改进方法是在利用说话人特征 d-vector 的基础上改进，进而提高模型的适应性的方法。Variani et al. (2014) 提出 d-vector 作为说话人特征，首先训练可以在帧级别对说话人进行分类的 DNN，然后训练的 DNN 可以从最后一个隐藏层中提取说话人特定的特征。Veselý et al. (2016) 采用摘要网络产生输入特征的序列级概要，辅助特征由序列摘要神经网络 (Sequence Summarizing Neural Network, SSNN) 产生一个“摘要向量”，它表示话语的声学摘要，该神经网络将与声学模型相同的特征作为输入，并通过对输出的时间平均进行嵌入，i-vector 和 SSNN 说话人自适应方法都在 AMI 会议数据上进行了比较，两者性能相当。Sari et al. (2019) 使用长短期记忆网络 (Long Short-Term Memory, LSTM) 作为辅助网络，d-vector 作为辅助特征进行说话人自适应，在 HUB4 数据集上，与未适应模型和对抗模型相比，该方法实现了更高的 senone 分类准确度和更低的单词错误率，绝对词错误率降低了高达 2%。同样基于 d-vector 的改进方法的辅助特征的提取需要用到整句话的信息，不适用于实时语音识别。

说话人特征 i-vector 是表示帧级别特征分布模式的特征，其提取本质上是高斯混合模型超向量的降维，并且模型框架假定 i-vector 具有高斯分布。使用 DNN 提取的 d-vector，是来自该 DNN 的最后隐藏层的均值，没有假设任何有关特征分布。但是在 d-vector 说话人特征提取中，加权取平均的方式无法突出区分性说话人信息。本文在基于 i-vector 的基础上引入 NTM，将记忆模块存放各地域说话人特征向量，使用注意力机制代替 NTM 的余弦相似度算法，采用注意

力机制计算说话人特征与记忆模块中说话人特征的权重矩阵，通过权重矩阵计算出说话人特征，从而提升说话人特征之间区分性，提高蒙古语语音识别系统的适应性。

## 3 方法

### 3.1 问题描述

在蒙古语说话人自适应语音识别过程中，内蒙古 A 地区的说话人 $A$ 的一句蒙古语语音序列为 $X = \{x_1, x_2, ..., x_t, ..., x_l\}$，$x_t \in R^D$ 其中表示语音序列 $X$ 中的第 $t$ 个语音帧的特征向量，由一个 D 维向量构成，则语音的声学特征矩阵 $F(X) = [f_1, f_2, ..., f_t, ..., f_l]$，$F(X)$ 是以帧数为行，维度为列。使用声学特征得到对应说话人特征向量 $S = (s_1, s_2, ..., s_n)$，同时内蒙古 B 地区的说话人 $B$ 的一句蒙古语语音序列为 $H = \{h_1, h_2, ..., h_r\}$，同理可得对应说话人特征向量 $D = (d_1, d_2, ..., d_n)$，其中 $S$ 和 $D$ 是两个说话人特征矩阵，将两者通过余弦相似度方法，经过计算可得 $S \cong D$，此时可以说两个地区的说话人特征是不具有区分性。说话人特征提取模型 E 中的说话人特征集合为 $M$，将 $S$ 和 $D$ 分别通过模型 E 重新提取分别获得说话人特征 $S^*$ 和 $D^*$，由于模型中的特征 $M$ 不发生变化，因此通过算法提取到的两个特征是 $S^* \cong D^*$。

### 3.2 模型架构

蒙古语语音识别说话人自适应模型架构主要包括说话人自适应模型和声学模型两个部分，如图 1 所示。说话人自适应模型主要是将输入的蒙古语语音映射到声学特征序列和说话人特征序列，并将两种特征融合成一个特征序列。采用时延神经网络-隐马尔可夫模型 (Time Delay Neural Network-HMM, TDNN-HMM) (Waibel et al., 1989) 作为声学模型，声学模型主要是将每一个单词与基本的发音单位对应起来，根据特征序列直接得到最匹配的字符串。蒙古语声学模型采用 TDNN，优点是动态适应时域特征变化和参数较少，相较于传统的 DNN 的输入层与隐含层互相连接，TDNN 在这里做了一点改变，即隐含层的特征不仅与当前时刻的输入有关，而且还与未来时刻的输入有关。该网络的每一个隐藏层的当前时刻输出值与其向前和向后时间节点的输出值的拼接，作为下一个隐藏层的输出。

给定输入蒙古语语音 $X = \{x_1, x_2, ..., x_t, ..., x_l\}$，$x_t$ 表示第 $t$ 个语音帧，$x_t$ 通过分帧加窗和处理后得到的声学特征向量 $F = (f_1, f_2, ..., f_m)$，经过说话人特征提取模型得到说话人特征向量 $S = (s_1, s_2, ..., s_n)$。TDNN 网络第一个隐藏层的输出计算如公式 (1) 所示。

$$h_t = g(w_f x_t(c) + b_f) \tag{1}$$

其中 $h_t$ 表示第 $t$ 帧输出的隐藏层向量，$x_t(c)$ 表示第 $t$ 帧相邻输入特征拼接的向量，$w_f$ 和 $b_f$ 分别表示 TDNN 网络的权重矩阵和偏置向量，$g(\cdot)$ 表示 ReLU 非线性激活函数。本文的第 t 帧输入特征是将声学特征与说话人特征通过公式 (1) 得到的特征 $R$，$x_t(c)$ 也表示第 $t$ 帧相邻输入特征 $R$ 拼接的向量。后面的时延隐藏层的计算与第一个隐藏层类似，以上一隐藏层的输出 $h_t$ 作为输入进行计算。基于辅助特征的蒙古语语音识别在线说话人自适应模型不需要对声学模型增加自适应参数，蒙古语声学模型训练过程即是说话人自适应过程。

### 3.3 基于注意力的说话人特征提取单元

基于注意力的说话人特征提取单元是由注意力模块和记忆模块组成的，其中记忆模块的主要功能是存放说话人特征向量，注意力模块的主要功能是选择记忆模块中说话人特征进行组合成新的说话人特征，基于注意力的说话人特征提取单元结构图如图 2 所示。

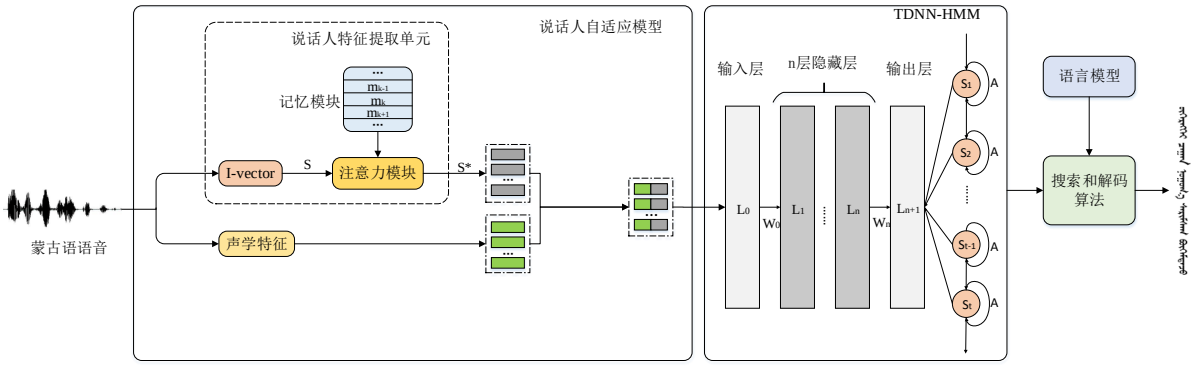



图 1. 蒙古语语音识别说话人自适应模型架构

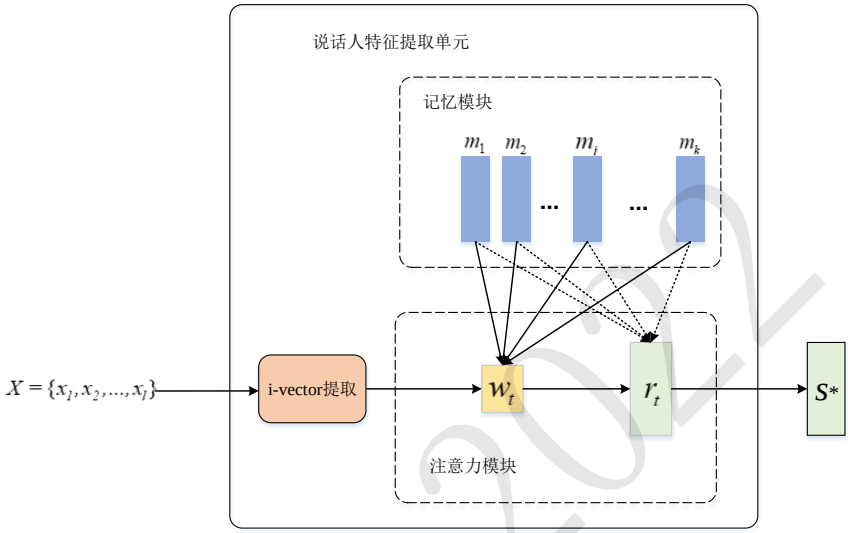



图 2. 基于注意力的说话人特征提取单元结构

记忆模块中存放说话人特征向量是通过训练好的 i-vector 说话人特征提取模型得到的，首先提取各个地区训练集的句子级 i-vector，然后使用 K-means 算法将每个蒙古语地区的说话人特征向量进行聚类，每个地区的 i-vector 的聚类数目相同，最后将聚类的所有地区的 i-vector 特征向量组合成记忆模块。由于受蒙古语语料库的影响，目前蒙古语语料库只有 8 个地区的语料比较丰富，因此记忆模块中包含 8 个蒙古地区的说话人特征向量。

记忆模块可以表示为 $M = \{m_1, m_2, ..., m_k\}$，其中 $m_k$ 表示第 $k$ 个说话人特征向量。给定一个 $t$ 时刻的说话人特征向量 $s_t$，首先通过注意力机制计算说话人特征向量 $s_t$ 与记忆模块 $M$ 的相似程度矩阵，如下公式：

$$K(s_t, m_i) = vtanh(Ws_t + Um_i) \tag{2}$$

其中，$K(s_t, m_i)$ 表示的是说话人特征向量 $s_t$ 与第 $i$ 个说话人特征向量相似程度，而矩阵 $W$、$U$ 和向量 $v$ 是注意力模块的参数。

通过 softmax 激活函数将相似程度 $K(s_t, m_i)$ 的输出映射为概率表达，得到锐化后的权重 $w_t(i)$，由说话人特征向量 $s_t$ 和标量 $\gamma_t$ 和说话人特征向量 $s_t$ 与记忆模块 $M$ 之间相似程度的确定，如下所示：

$$w_t(i) = \frac{e^{\gamma_t K(s_t, m_i)}}{\sum_{j=1}^{k} e^{\gamma_t K(s_t, m_j)}} \tag{3}$$

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。

将权重 $w_t(i)$ 和记忆模块块中所有的蒙古语说话人特征向量的进行加权求和后，即可得到新的说话人向量 $r_t$，具体表示为：

$$r_t = \sum_{n=1}^{k} w_t(i) m_n \tag{4}$$

其中，$r_t$ 表示第 $t$ 帧基于所有地区的说话人向量集合所构建的新的说话人向量，将区分性说话人特征 $r_t$ 表示为 s-vector。记忆模块中说话人特征向量是来自从不同内蒙地区和不同说话人的语音数据中提取得到的，因此经过注意力模块的计算得到的说话人特征向量之间具有区分性。

## 4 实验

### 4.1 实验设置

实验选用的语料库为 IMUT-MC (刘志强等, 2021)，由本实验室构建的一个针对蒙古语语音识别任务的语音语料库，其中 IMUT-MC2 和 IMUT-MC3 的数据来自于蒙古语的日常对话，文本语句都比较简短。录音人员来自内蒙古自治区 8 个地区的 210 人，说话人每人重复录制 200 句，年龄分布在 18 岁到 24 岁之间。蒙古语语音数据均为 16KHz 采样率、16bit 比特率、单声道的格式。

训练集是将 IMUT-MC2 和 IMUT-MC3 的数据集整合在一起，简称为 IMUT-MCT 数据集，其中选取包含 8 个内蒙古地区的 33200 个语音音频，共覆盖了 166 个说话人，其中 84 个男性和 84 个女性。验证集从 IMUT-MCT 数据集中选取包含 8 个内蒙古地区的 3000 个语音音频，共覆盖 15 个说话人，其中 8 个男性和 7 个女性。将训练集与测试集的地区说话人员的交集比例，依次按照 0%，30%，50%，70%，100% 构建 5 组测试集，分别命名为 Test1、Test2、Test3、Test4 和 Test5。其中 Test1 测试集是从 IMUT-MCT 数据集中选取包含 8 个内蒙古地区的 4000 个语音音频，共覆盖了 20 个说话人，其中 10 个男性和 10 个女性。

将通过 3 个实验来验证该方法的有效性。第一个是消融实验，为了说明在蒙古语在线说话人自适应任务中，基于记忆的说话人特征单元的有效性，将蒙古语 s-vector 特征与蒙古语 i-vector 特征和蒙古语 d-vector 特征在不同的测试集上进行对比。第二个是说话人特征提取方法对比实验，旨在探索基于注意力的说话人特征方法对于不同蒙古语语音识别系统的适应性，主要是从不同蒙古语说话人特征 i-vector、d-vector 和 s-vector 与不同蒙古语声学模型 DNN-HMM、LSTM-HMM 和 End-to-End 上进行对比实验。第三个是案例分析验，通过蒙古语语音样例来展示真实的识别结果，测试语料使用 i-vector 与 s-vector 两种说话人特征提取方法进行识别，验证本文提出方法的有效性。

### 4.2 评价指标

蒙古语语音识别模型的评价指标包括有词错率 (Word Error Rate, WER)、句错率 (Sentence Error Rate, SER) 和错误下降率 (Error Drop Rate, EDR)。

(1)WER 是指所有错误词的和所占总词数的百分比，计算公式为：

$$WER = \frac{S_w + D_w + I_w}{N_W} * 100\% \tag{5}$$

其中 $S_w$ 表示替换错误的词数，$D_w$ 表示删除错误的词数，$I_w$ 表示插入错误的词数，$N_W$ 表示数据集中的总词数。

(2)SER 是指所有识别结果与对应文本不能正确匹配的测试语音所占总语音数的百分比，计算公式为：

$$SER = \frac{N_error}{N_S} * 100\% \tag{6}$$

其中 $N_error$ 表示识别错误的蒙古语音频的个数，$N_S$ 表示数据集中蒙古语音频的总个数。

(3)EDR 是现在方法的错误率与原来方法的错误率相差的值和原来方法的错误率的百分比值，计算公式为：

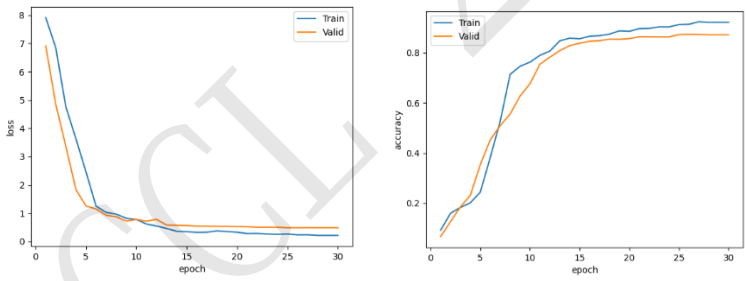
$$EDR = \frac{|B - A|}{B} * 100\% \tag{7}$$

其中 $A$ 表示现在方法的错误率，$B$ 表示原来方法的错误率，错误率可以是该方法的 WER 或者 SER。

### 4.3 实验结果与分析

#### 4.3.1 收敛性实验

训练过程在 IMUT-MCT 数据集上对使用基于注意力的说话人特征提取单元的蒙古语语音识别模型开展，收敛情况如图 3 所示。为了保证基于注意力的蒙古语语音识别模型的收敛效果，实验采用训练集和验证集的损失值和准确率来验证模型收敛。由图 3a) 中可知，基于注意力的蒙古语语音识别模型训练集和验证集上的损失函数呈指数下降，并于 25 轮左右使 Loss 趋于平缓，表明模型能够收敛。由图 3b) 中可知，在训练集和验证集上的准确率不断上升，最终在 25 轮左右识别准确率趋于平缓。综上所述，使用基于注意力的说话人提取单元提取说话人特征作为声学模型的输入来训练得到的模型可以收敛且无过拟合现象，可以学习到了训练集的数据分布特征。




(a) 训练的 loss 值的变化　　(b) 训练过程中识别准确率

图 3. 基于注意力的蒙古语语音识别模型的收敛情况

#### 4.3.2 说话人特征单元消融实验

说话人特征单元消融实验在不同测试集上将提出的蒙古语 s-vector 特征与蒙古语 i-vector、d-vector 特征进行对比。实验对不同说话人特征在蒙古语语音识别系统的 WER 和 SER 进行验证，其中采用特征拼接方法将蒙古语声学特征 Fbank 和说话人特征进行融合，实验结果如表 1 所示。

分析表 1 实验结果可知：不采用蒙古语说话人特征进行蒙古语说话人自适应时，Test5 测试集的蒙古语语音识别系统取得了比 Test1 测试集的蒙古语语音识别系统更好的识别效果；当采用不同的说话人特征进行说话人自适应时，Test2 测试集的 WER 和 SER 优于 Test1 测试集的 WER 和 SER，训练集与测试集的地区说话人数的交集比例逐步增大，WER 和 SER 也在随之

降低。对比不同的说话人特征，当说话人无论是否包含在训练集或者自适应集中时，s-vector 特征进行说话人自适应得到的 WER 均优于蒙古语 i-vector、d-vector 特征，比如在测试集为 Test1 和测试集为 Test5 的实验中，采用蒙古语 s-vector 特征比蒙古语 i-vector、d-vector 特征进行蒙古语语音识别说话人自适应的 WER 分别降低了 4.96%、1.08% 和 5.39%、1.2%。虽然在测试集为 Test1 时 d-vector 方法比 s-vector 方法的 SER 低了 1.01%，但是在其他测试集上蒙古语 s-vector 特征的 SER 优于蒙古语 i-vector。实验结果验证提出蒙古语 s-vector 特征的有效性。

| 说话人特征 | WER(%) | | | | | SER(%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Test1** | **Test2** | **Test3** | **Test4** | **Test5** | **Test1** | **Test2** | **Test3** | **Test4** | **Test5** |
| 无 | 36.54 | 33.87 | 30.26 | 29.23 | 28.87 | 58.37 | 56.18 | 54.83 | 52.48 | 50.36 |
| i-vector | 30.33 | 28.23 | 27.18 | 26.08 | 25.97 | 43.16 | 40.73 | 39.95 | 36.48 | 35.38 |
| d-vector | 26.45 | 25.07 | 23.88 | 23.53 | 21.78 | **36.77** | 35.25 | 35.78 | **32.53** | 31.47 |
| **s-vector** | **25.37** | **23.28** | **22.28** | **21.19** | **20.58** | 37.78 | **34.83** | **34.47** | 33.58 | **31.23** |

表 1. 说话人特征提取单元的消融实验

| 说话人特征 | 声学模型 | WER(%) | | | | | SER(%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Test1** | **Test2** | **Test3** | **Test4** | **Test5** | **Test1** | **Test2** | **Test3** | **Test4** | **Test5** |
| 无 | DNN-HMM | 38.56 | 36.86 | 34.66 | 33.46 | 31.34 | 64.16 | 62.73 | 60.46 | 58.86 | 56.83 |
| | TDNN-HMM | 36.54 | 33.87 | 30.26 | 29.23 | 28.87 | 58.37 | 56.18 | 54.83 | 52.48 | 50.36 |
| | LSTM-HMM | 26.54 | 25.89 | 24.33 | 23.23 | 22.96 | 42.53 | 40.45 | 38.16 | 36.10 | 35.85 |
| | End-to-End | 21.45 | 20.23 | 19.78 | 18.86 | 17.23 | 35.52 | 34.46 | 32.21 | 31.58 | 30.97 |
| i-vector | DNN-HMM | 33.46 | 31.78 | 30.89 | 28.72 | 26.34 | 58.45 | 56.36 | 55.86 | 54.58 | 54.13 |
| | TDNN-HMM | 30.33 | 28.23 | 27.18 | 26.08 | 25.97 | 43.16 | 40.73 | 39.95 | 36.48 | 35.38 |
| | LSTM-HMM | 24.23 | 22.15 | 22.89 | 21.76 | 20.45 | 36.63 | 35.25 | 34.03 | 32.58 | 30.56 |
| | End-to-End | 19.89 | 18.86 | 17.45 | 16.23 | 15.99 | 30.56 | 28.06 | 26.75 | 25.36 | 25.45 |
| d-vector | DNN-HMM | 30.23 | 28.18 | 26.59 | 25.45 | 24.56 | 39.36 | 38.34 | 37.98 | 36.43 | 34.21 |
| | TDNN-HMM | 26.45 | 25.07 | 23.88 | 23.53 | 21.78 | 36.77 | 35.25 | 35.78 | 32.53 | 31.47 |
| | LSTM-HMM | 23.56 | 24.78 | 24.88 | 22.26 | 20.08 | 34.23 | 33.03 | 32.74 | 30.47 | 29.31 |
| | End-to-End | 18.97 | 17.36 | 16.56 | 16.23 | 15.56 | 29.89 | 28.73 | 26.16 | 25.83 | 24.86 |
| s-vector | DNN-HMM | 29.89 | 28.56 | 27.72 | 25.34 | 24.02 | 39.89 | 38.78 | 36.56 | 35.23 | 34.12 |
| | TDNN-HMM | 25.37 | 23.28 | 22.28 | 21.19 | 20.58 | 37.78 | 34.83 | 34.47 | 33.58 | 31.23 |
| | LSTM-HMM | 23.78 | 22.55 | 21.45 | 20.26 | 19.87 | 33.57 | 32.75 | 31.48 | 30.33 | 29.15 |
| | End-to-End | 18.70 | 17.45 | 16.16 | 15.95 | 14.89 | 29.36 | 28.87 | 26.54 | 25.13 | 24.05 |

表 2. 不同说话人特征提取方法在不同声学模型上的 WER

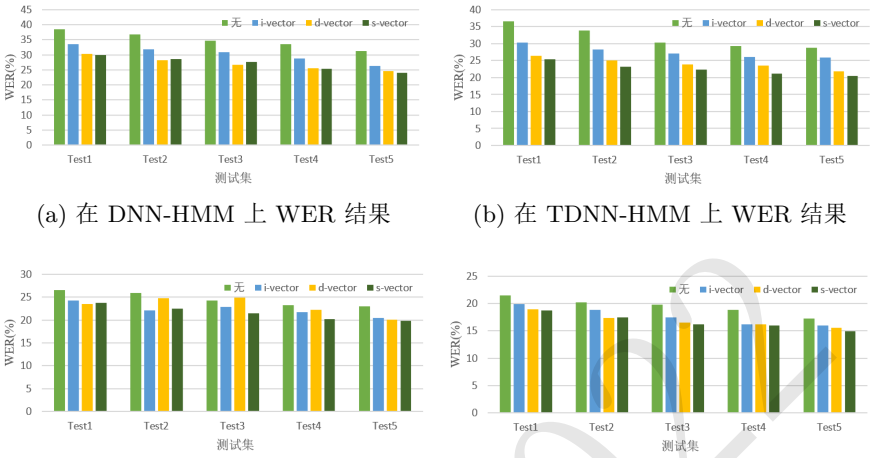
#### 4.3.3 模型适应性实验

为了验证不同说话人特征提取方法在不同语音识别模型上的性能，在不同测试集下开展模型适应性实验。实验分别对比 DNN-HMM、LSTM-HMM 和 End-to-End 声学模型。本文提出的蒙古语 s-vector 特征与蒙古语说话人特征 i-vector、说话人特征 d-vector 进行对比。实验对不同说话人特征在不同的蒙古语语音识别系统的 WER 和 SER 进行验证，实验结果如表 2 所示。

分析表 2 实验结果可知：对比不同说话人特征提取方法和在不同模型上性能，在这四种模

型中可以看出 End-to-End 的识别 WER 和句 SER 均优于 DNN-HMM、TDNN-HMM、LSTM-HMM，比如在说话人特征提取方法为 d-vector、测试集为 Test1 和提取方法为 s-vector、测试集为 Test1 的实验中，采用 End-to-End 比 DNN-HMM、TDNN-HMM 和 LSTM-HMM 进行蒙古语语音识别说话人自适应的 WER 分别降低了 11.26%、7.48%、4.59% 和 11.19%、6.67%、5.08%，实验结果验证了在不同模型中蒙古语 s-vector 特征对蒙古语语音识别系统的有效性。

为了更加直观在不同声学模型上使用不同说话人特征提取方法所对应结果，如图 4 所示。在图 4 中，横坐标为不同蒙古语测试集，纵坐标为词错率。




(a) 在 DNN-HMM 上 WER 结果　(b) 在 TDNN-HMM 上 WER 结果

(c) 在 LSTM-HMM 上 WER 结果　(d) 在 End-to-End 上 WER 结果

图 4. 不同说话人特征提取方法在不同声学模型上的 WER

分析图 4 实验结果可知：在图 4(a) 中，可以看出使用说话人自适应方法比没有使用说话人自适应取得了更好的识别准确率，而且 s-vector 和 d-vector 在不同测试集上均优于 i-vector，也可以在图中看出 s-vector 和 d-vector 性能相当。说话人特征 s-vector 是基于 i-vector 的改进方法，比 i-vector 有更好的性能。在图 4(b) 中，s-vector 和 d-vector 在不同测试集上均优于 i-vector，但是在不同的测试集 s-vector 性能优于 d-vector。在图 4(c) 和图 4(d) 中，得到了与图 34(a) 类似的结论，验证了 s-vector 在不同声学模型的适应性能力。


(a) 在 WER 上的结果　(b) 在 SER 上的结果

图 5. 说话人提取单元的错误下降率

为直接验证说话人特征提取单元对蒙古语语音识别系统的影响，本文还采用 EDR 评价指标对不同模型进行评价，其中是采用 s-vector 特征方法和没有说话人特征方法之间的错误下降率，如图 5 所示。在图 5 中，横坐标为不同蒙古语测试集，纵坐标为 ESR(%)。

从图 5(a) 可以看出，在不同测试集下，下降率最高的为 TDNN-HMM 模型，其次才是 DNN-HMM 模型，最低的为 End-to-End 模型。但是从图 5(b) 可以看出，在不同测试集下，下降率最高的为 DNN-HMM 模型，其次才是 TDNN-HMM 模型，最低的为 End-to-End 模型。综上所述，可以看出本文的方法比较适用于 TDNN-HMM 模型和 DNN-HMM 模型，对于这个两种模型影响较大。

### 4.3.4 案例分析

为了验证说话人特征提取单元的有效性，通过样例来展示真实的识别结果。在蒙古语语音中，例如语音：00001373-F-M-19.wav，其中 00001373 表示语句编号；F 为地区编码，数据集中收集了 10 个地区，F 表示兴安盟地区；M 为性别编码，M 表示男，F 表示女；19 表示年龄。表 3 展示了蒙古语语音识别的样例，1-3 是包含在记忆模块地区和不包含在训练集中，4-7 是不包含在记忆模块地区和不包含在训练集中，其中选取的语音数据的时长集中在 4-7s 之间。

| 编号 | 语音数据编号 | 标签数据 | 备注 |
|---|---|---|---|
| 1 | 00000016-F-M-19.wav | ᠨᠠᠷ᠋ ᠬᠦᠮᠦᠨ ᠳᠤ ᠪᠢ ᠤᠨ ᠨᠥᠬᠥᠳᠡᠢᠮᠢᠨ ᠪᠠᠶᠢᠴᠤᠯᠵᠤ ᠪᠠᠢᠨ᠋ᠠ ᠤᠤ | 他把朋友介绍给您了吗？ |
| 2 | 00000140-F-F-20.wav | ᠨᠢᠭᠡᠨ ᠳᠤ᠌ᠬᠢᠳᠠᠭᠠ ᠭᠣ ᠮᠣᠳᠣᠯᠢᠨ᠋ ᠬᠣᠶᠠᠷ ᠬᠣᠨᠣᠭ ᠪᠤᠶᠤ ᠭᠤᠷᠪᠠᠨ ᠰᠠᠷ᠋ᠠ ᠶᠢᠨ᠋ ᠬᠤᠭᠴᠠᠨ᠋ | 第一次机遇十多天或者一个月左右。 |
| 3 | 00000206-C-F-19.wav | ᠥᠷ᠌ ᠭᠣ ᠬᠥᠭᠵᠢᠮ᠌ ᠪᠠᠶᠢᠭᠤᠯᠬᠤ ᠬᠡᠷᠡᠭ ᠪᠢᠴᠢᠭᠰᠡᠨ ᠦᠭᠡ ᠬᠤᠭᠤᠢ ᠮᠠᠭᠠ᠋ᠵᠤ | 附近有的话，好坏无所谓。 |
| 4 | 00000220-H-M-23.wav | ᠮᠠᠨᠠᠢᠪᠠ ᠨᠠᠮᠢᠨᠴᠠᠭ᠋ᠠᠨ᠋ ᠪᠠ ᠪᠢᠳᠡ ᠬᠤᠭᠤᠴᠠᠭ᠋ᠠᠨ᠋ ᠳᠡᠯᠬᠦᠦ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ | 那么我们俩进二手衣服店吧。 |
| 5 | 00000292-H-F-19.wav | ᠠᠴ ᠠᠭᠤᠯᠠᠬᠤ ᠤᠨ ᠤᠨᠵᠠᠭ᠋ᠠ ᠤᠤ ᠵᠠᠭᠰᠠᠭ᠋ᠠᠯᠲ ᠪᠠᠢ᠋ᠨ᠋ᠠ ᠤᠤ | 您带几箱子东西？ |
| 6 | 00000306-G-M-20.wav | ᠬᠡᠷᠡᠭᠲᠡᠢ ᠴᠠᠭ ᠪᠠ ᠬᠡᠯᠡᠵᠤ ᠬᠡᠯᠡᠭᠰᠡᠨ ᠴᠠᠭ ᠳᠤ ᠪᠠ ᠬᠡᠯᠡ | 该求的时候要求，该要的时候该要。 |
| 7 | 00000843-G-F-19.wav | ᠮᠣᠩᠭᠣᠯ ᠬᠦᠮᠦᠨ ᠤ ᠲᠠᠪᠤᠨ ᠠᠮᠢᠲᠠᠨ ᠶᠠ ᠭᠡᠵᠦ ᠠᠯᠢᠪᠠ ᠬᠤᠨ᠌ ᠮᠣᠷᠢ᠌ ᠬᠥᠬᠡ ᠬᠤᠯᠠᠭᠠᠨ ᠲᠠᠪᠢᠨ᠋ | 蒙古人的五畜是指羊、山羊、牛、马、骆驼。 |

表 3. 蒙古语测试样例数据

将上述的测试样例数据采用基于注意力的说话人特征提取方法进行识别，结果如表 4 所示。其中具有下划线的词、“***”和字体加双下划线的词分别表示替换词、删除词和插入词。

| 编号 | 识别标签结果 | 备注 | 替换词错误 | 删除词错误 | 插入词错误 |
|---|---|---|---|---|---|
| 1 | ᠨᠠᠷ᠋ ᠬᠦᠮᠦᠨ ᠳᠤ ᠪᠢ ᠬᠡᠨᠳᠡ ᠤᠨ᠌ ᠤᠨ᠌ ᠨᠥᠬᠥᠳᠡᠢᠮᠢᠨ᠋ ᠪᠠᠶᠢᠴᠤᠯᠵᠤ ᠪᠠᠢᠨ᠋ᠠ ᠤᠤ | 请朋友介绍这两家业余的事 | 1 | 0 | 1 |
| 2 | ᠨᠢᠭᠡᠨ ᠳᠤ᠌ᠬᠢᠳᠠᠭᠠ ᠴᠢᠭᠠᠭ᠋ ᠬᠣᠨᠣᠭ᠋ᠤᠨ᠋ ᠤᠨᠵᠠᠭ᠋ᠠ ᠤᠤᠷ᠋ ᠶᠢᠨ᠋ ᠬᠤᠭᠴᠠᠨ᠋ | 第一次机遇十多天偶尔一个月左右。 | 1 | 0 | 0 |
| 3 | ᠥᠷ᠌ ᠭᠣ ᠬᠥᠭᠵᠢᠮ᠌ ᠪᠠᠶᠢᠭᠤᠯᠬᠤ ᠬᠡᠷᠡᠭ ᠬᠤᠭᠤᠢ ᠮᠠᠭᠠ᠋ᠵᠤ | 附近有的话，坏无所谓。 | 1 | 0 | 0 |
| 4 | ᠮᠠᠨᠠᠢᠪᠠ ᠨᠠᠮᠢᠨᠴᠠᠭ᠋ᠠᠨ᠋ ᠪᠠ ᠪᠢᠳᠡ ᠬᠤᠭᠤᠴᠠᠭ᠋ᠠᠨ᠋ ᠳᠡᠯᠬᠦᠦ ᠶᠠᠪᠤᠭᠠᠷᠠᠢ | 那么我们俩进二手衣服店什么的。 | 0 | 0 | 1 |
| 5 | ᠠᠴ ᠠᠭᠤᠯᠠᠬᠤ ᠤᠨ ᠤᠨᠵᠠᠭ᠋ᠠ ᠤᠤ ᠵᠠᠭᠰᠠᠭ᠋ᠠᠯᠲ ᠪᠠᠢ᠋ᠨ᠋ᠠ ᠤᠤ | 您带几箱子东西？ | 0 | 0 | 0 |
| 6 | ᠬᠡᠷᠡᠭᠲᠡᠢ ᠴᠠᠭ ᠪᠠ ᠬᠡᠯᠡᠵᠤ ᠬᠡᠯᠡᠭᠰᠡᠨ ᠴᠠᠭ ᠳᠤ ᠪᠠ ᠬᠡᠯᠡ | 该求的时候要求，该要的时候该要。 | 0 | 0 | 0 |
| 7 | ᠮᠣᠩᠭᠣᠯ ᠬᠦᠮᠦᠨ ᠤ *** ᠠᠮᠢᠲᠠᠨ ᠶᠠ ᠭᠡᠵᠦ ᠠᠯᠢᠪᠠ ᠬᠤᠨ᠌ ᠮᠣᠷᠢ᠌ ᠬᠥᠬᠡ ᠬᠤᠯᠠᠭᠠᠨ ᠲᠠᠪᠢᠨ᠋ | 蒙古人的 *** 畜是指羊、山羊、牛、马、骆驼。 | 0 | 1 | 0 |

表 4. 识别结果据

分析表 4 中识别结果可知：通过错误次数进行分析，使用 s-vector 得到的识别结果优于 i-vector，表明该方法得到的区分性说话人特征有利于声学模型建模。但是使用本文方法识别结果还存在错误，说明该方法仍需要改进进而提高准确率。

## 5 结论

围绕基于辅助特征的说话人自适应方法在蒙古语语音识别任务的适应性问题上开展研究。设计基于注意力的说话人特征提取单元，增加提取到蒙古语说话人特征之间的区分性，并将其用于声学模型的建模，降低蒙古语语音识别系统的 WER 和 SER。通过说话人特征单元的消融实验，可知使用 s-vector 特征比 d-vector 特征的 WER 降低 1.08%。实验结果间接表明蒙古语 s-vector 特征获得了较高的说话人特征区分性，并且提升蒙古语语音识别系统的识别准确率，但是部分结果中证明区分性的说话人特征对于声学模型建模也会产生不好的影响。未来研究可以尝试将不同类型的说话人特征存入记忆模块进行对比，如 x-vector (Snyder et al., 2018) 或 d-vector。

第二十一届中国计算语言学大会论文集，第344页-第354页，南昌，中国，2022年10月14日至16日。

## 参考文献


Abdel-Hamid O and Jiang H. 2013. *Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition based on Discriminative Learning of Speaker Code.* Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, May 26-31, pages 7942-7946.

Bell P, Fainberg J, Klejch O, et al. 2020. *Adaptation Algorithms for Speech Recognition: An Overview.* IEEE Open Journal of Signal Processing, 2:33-66.

Cardinal P, Dehak N, Zhang Y, et al. 2015. *Speaker Adaptation using the I-vector Technique for Bottleneck Features.* Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Sep 6-10, pages 2867-2871.

Cui X D, Goel V, Saon G, et al. 2017. *Embedding-based Speaker Adaptive Training of Deep Neural Networks.* Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Aug 20-24, pages 122-126.

Dong Y, Yao K S, Hang S, et al. 2013. *KL-Divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition.* Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, May 26-31, pages 7893-7897.

Graves A, Wayne G, and Danihelka I. 2014. *Neural Turing Machines.* https://arxiv.org/abs/1410.5401 2014-12-10

Neto J P, Almeida L B, Hochberg M, et al. 1995. *Speaker-Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System.* Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, Sep 18-21, pages 2171-2174.

Stadermann J and Rigoll G. 2005. *Two-Stage Speaker Adaptation of Hybrid Tied-Posterior Acoustic Models.* Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, Mar 18-23, pages 977-980.

Saon G, Soltau H, Nahamoo D, et al. 2013. *Speaker Adaptation of Neural Network Acoustic Models using I-vectors.* Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Dec 8-12, pages 55-59.

Swietojanski P and Renals S. 2014. *Learning Hidden Unit Contributions for Unsupervised Speaker Adaptation of Neural Network Acoustic Models.* Proceedings of the 2014 IEEE Spoken Language Technology Workshop, South Lake, Dec 7-10, pages 171-176.

Samarakoon L and Sim K C. 2016. *Factorized Hidden Layer Adaptation for Deep Neural Network based Acoustic Modeling.* IEEE Transactions on Audio, Speech and Language Processing, 24(12): 2241-2250.

Snyder D, Garcia-Romero D, SELL G, et al. 2018. *X-Vectors: Robust DNN Embeddings for Speaker Recognition.* Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Apr 15-20, pages 5329-5333.

Sari L, Thomas S, Hasegawa-Johnson M A. 2019. *Embedding-based Speaker Adaptive Training of Deep Neural Networks.* Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Aug 20-24, pages 122-126.

Variani E, Lei X, McDermott E, et al. 2014. *Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification.* Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, May 4-9, pages 4052-4056.

Veselý K, Watanabe S, Zmolíková K, et al. 2016. *Sequence Summarizing Neural Network for Speaker Adaptation.* Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, Mar 20-25, pages 5315-5319.

Waibel A, Hanazawa T, Hinton G, et al. 1989. *Phoneme Recognition using Time-Delay Neural Networks.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(3):328-339.

刘志强, 马志强, 张晓旭, 等. 2021. *IMUT-MC*：一个针对蒙古语语音识别的语音语料库. 中国科学数据, 2021.(2021-12-29). DOI: 10.11922/11-6035.csd.2021.0096.zh

朱方圆，马志强，陈艳, 等. 2021. 语音识别中说话人自适应方法研究综述. 计算机科学与探索, 2021, 15(12):2241-2255.