文章编号：2095-4980(2023)09-1163-08

多智能体自组织语音识别

陈俊淇，张晓雷

(西北工业大学 航海学院，陕西 西安 710072)

摘　要：语音感知是无人系统的重要组成部分，已有的工作大多集中于单个智能体的语音感知，受噪声、混响等因素的影响，性能存在上限。因此研究多智能体语音感知，通过多智能体自组织、相互协作，提高感知性能非常必要。假设每个智能体输出一个通道的语音流条件下，本文提出一种多智能体自组织语音系统，旨在综合利用所有通道提高感知性能；并进一步以语音识别为例，提出能处理大规模多智能体语音识别的通道选择方法。基于Sparsemax算子的端到端语音识别流注意机制，将带噪通道权重置零，使流注意力具备通道选择能力，但Sparsemax算子会将过多通道权重置零。本文提出Scaling Sparsemax算子，只将带噪较强的通道权重置零；同时提出了多层流注意力结构，有效降低了计算复杂度。在30个智能体的无人系统环境下，基于conformer架构的识别系统实验结果表明，在通道数失配的测试环境下，提出的Scaling Sparsemax在仿真数据集上的文字差错率(WER)相比Softmax降低30%以上，在半真实数据集上降低20%以上。

关键词：多智能体语音识别；通道选择；注意力；Scaling Sparsemax算子

中图分类号：TP391.4　　**文献标志码：**A　　**doi：**10.11805/TKYDA2021247

Multi-agent ad-hoc speech recognition

CHEN Junqi，ZHANG Xiaolei

(School of Marine Science and Technology，Northwestern Polytechnical University，Xi'an Shaanxi 710072，China)

Abstract：Speech perception is an important part of unmanned systems. Most of the existing work focuses on the speech perception of a single agent, which is affected by factors such as noise and reverberation, and the performance has an upper limit. Therefore, it is necessary to study multi-agent speech perception, and improve perception performance through multi-agent self-organization and mutual cooperation. A multi-agent ad-hoc speech system is proposed under the assumption that each agent outputs a channel of speech stream. The multi-agent ad-hoc speech system aims to comprehensively utilize all channels to improve perception performance. Taking the speech recognition as an example, a channel selection method that can handle large-scale multi-agent speech recognition is proposed. Specifically, an end-to-end speech recognition stream attention mechanism based on Sparsemax operator is proposed to force the channel weights of noisy channels to zero, and make the stream attention bear the function of channel selection. Nevertheless, Sparsemax would punish the weights of many channels to zero harshly. Therefore, Scaling Sparsemax is proposed, which punishes the channels mildly by setting the weights of strong noise channels to zero only. At the same time, a multilayer stream attention structure is proposed to effectively reduce computational complexity. Experimental results in an unmanned system environment with up to 30 agents under the conformer speech recognition architecture show that the Word Error Rate(WER) of the proposed Scaling Sparsemax is lower than that of Softmax by over 30% on simulation data sets, and by over 20% on semi-real data sets, in test scenarios with mismatched channel numbers.

Keywords：multi-agent speech recognition；channel selection；attention；Scaling Sparsemax

收稿日期：2021-06-14；**修回日期：**2021-08-09

近年来，智能家居等无人系统发展迅猛，语音感知是其中一个重要的组成部分。但大多数语音感知相关工作只针对单个智能体研究，每个智能体通常只有一个麦克风或一个麦克风阵列进行拾音[1-3]。在现实场景下，当说话人移动到与单个智能体距离较远的位置时[4]，接收到的语音质量会急剧下降，并且由于受到噪声、混响等因素的影响，其语音感知性能极其有限[5]。因此，为了使说话人在一定区域内移动都能获得较好的语音感知性能，有必要引入多智能体系统。

多智能体系统是分布式人工智能领域中的一个热点，是多个可以相互协作的简单智能体为完成某些全局或局部目标使用相关技术组成的分布式智能系统[6]。目前，多智能体系统在许多领域都得到了充分的应用，如云计算、智能交通、机器人集群、智能电网等[7]。近来，智能家居等对语音感知需求较大的应用受到越来越多的关注，而基于多智能体下的语音感知技术还没有深入的研究。本文提出一种多智能体自组织语音系统，考虑每个智能体包含一个麦克风，输出一个通道的语音流，多智能体自组织语音系统旨在使多个随机分布的智能体协同工作，综合利用所有通道的信息以提升语音感知性能。

进一步以语音识别为例，构建多智能体自组织语音识别系统，且考虑单个智能体包含单个通道。针对多智能体的协同控制，需要考虑两大问题：多智能体系统的一致性[8]；用于协同控制的多智能体技术。后者的核心在于通道选择，即利用通道权重分配和通道选择，自动将说话人附近的智能体组织为一个局部的多智能体系统[9-10]，从而获得更好的语音识别性能。

已有的针对语音识别任务的通道选择和权重分配标准可以划分为两类：a) 基于信号层面的标准[11-14]，如信噪比等；b) 基于识别层面的标准[11-12,15-16]，如词错误率等。对于前者，基于信号层面的评价指标虽然与语音识别性能有较强的关联性，但优化语音质量指标并不能得到最优的语音识别性能；后者基于优化语音识别性能设计通道选择和通道融合。较早的方法考虑选择语音识别解码后输出似然概率最大的通道[11-12]。深度学习出现后，更多方法考虑在语音识别模型中加入通道选择模型，R Li 等[15]提出了流注意力模型，利用注意力机制对所有通道进行权重分配和融合；同时提出了一种两阶段的训练方法[16]：第一步采用所有多通道数据训练单个语音识别系统；第二步将训练好的识别系统的参数分享到其他通道的识别系统上并将参数固定，再利用多通道数据微调流注意力模型。然而，上述方法都只考虑少量智能体的情况，并没有探究大规模多智能体语音识别情况下通道选择算法的泛化能力，且没有考虑丢弃某些被噪声影响过大的通道。

为解决上述问题，本文搭建了基于 conformer[17]的多智能体自组织语音识别系统，并提出了 2 种新的通道选择算法。本文核心方法是将流注意力中的 Softmax 算子替换为 2 种新的算子，分别为 Spasemax 和 Scaling Sparsemax，这 2 种算子可以将对识别系统性能提升没有贡献的带噪通道权值置零。针对多智能体的时滞一致性，借鉴 R Li 等[15]的思想，搭建了基于 conformer 架构下的双层注意力模型，其中第一层注意力用于将各个智能体的输出语音流对齐，第二层则是用于通道选择的流注意力。针对多智能体的收敛一致性，本文对 R Li 等的训练方法[16]做出改进，首先用干净语音数据训练单个智能体的语音识别系统，使所有智能体被成功训练且收敛到同一状态；再用多通道带噪数据训练基于 Sparsemax 和 Scaling Sparsemax 的流注意，使多智能体可以协同工作。在包含至多 30 个智能体的自组织语音识别系统下的实验结果证明，本文提出的方法能有效用于仿真和半真实场景中。

1 基于 conformer 的自组织语音识别系统

考虑单个智能体输出一个通道的语音流，分别设计了单通道和多通道系统，图 1 为提出的 conformer 结构下的单通道和多通道语音识别系统。为表示清晰，省去了残差连接和位置编码模块。在多智能体自组织语音识别系统训练中，第一阶段训练单通道识别系统，第二阶段训练多通道识别系统。下文中一个通道指代单个智能体。

1.1 基于 conformer 的单通道语音识别系统

图 1(a)为干净语音下训练的单通道识别系统。给定一个语料的输入声学特征 $\boldsymbol{X}\in\mathbb{R}^{T\times D_x}$ 及其目标输出文本序列 $\boldsymbol{O}\in\mathbb{R}^{L\times D_v}$，其中 T 和 D_x 分别是输入 $\boldsymbol{X}$ 的长度和特征维度，L 和 D_v 分别是输出的长度和字典大小。首先，输入 $\boldsymbol{X}$ 经过卷积下采样层，得到下采样后的输入 $\tilde{\boldsymbol{X}}\in\mathbb{R}^{\tilde{T}\times D_x}$；然后经过编码器 $\mathrm{Enc}(\cdot)$ 和解码器 $\mathrm{Dec}(\cdot)$：

$$\boldsymbol{H}=\mathrm{Enc}_{1:N_1}(\tilde{\boldsymbol{X}}) \tag{1}$$

$$\boldsymbol{c}_l=\mathrm{Dec}_{1:N_2}\left[\boldsymbol{H},\mathrm{Emb}(\boldsymbol{y}_{1:l-1})\right] \tag{2}$$

式中：$\boldsymbol{H}\in\mathbb{R}^{\tilde{T}\times D_h}$ 是由编码器提取得到的高维表征；下标 N_1 和 N_2 分别代表编码器和解码器的块数量；$\mathrm{Emb}(\cdot)$ 代表线性变换和位置编码。给定当前解码时间步 l 之前的输出 $\boldsymbol{y}_{1:l-1}\in\mathbb{R}^{l-1\times D_v}$ 后，可以得到解码时间步 l 的语义向量

$\boldsymbol{c}_l \in \mathbb{R}^{D_h}$；最后，通过一个线性变换将$\boldsymbol{c}_l$映射为输出向量$\boldsymbol{y}_l$。基于conformer的自动语音识别(Automatic Speech Recognition，ASR)优化目标最大化如式(3)所示：

$$\mathcal{L}=\sum_{l=1}^{L}\log(\boldsymbol{y}_l^{\mathrm{T}}\boldsymbol{o}_l) \tag{3}$$

式中$\boldsymbol{o}_l$为输出文本序列$\boldsymbol{O}$的第l个时间步的文本向量。多头注意力(Multi-Head Attention，MHA)机制在编码器和解码器中都具有重要的作用，同时也是conformer结构相较于双向长短时记忆结构[16]的关键不同点。多头注意力机制表达为：

$$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_n)\boldsymbol{W}^O \tag{4}$$

式中：$\boldsymbol{Q}\in\mathbb{R}^{T_1\times D_h}$、$\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{T_2\times D_h}$分别为质询矩阵、键矩阵和值矩阵；Concat(·)为矩阵拼接操作；n为头的数量；$\boldsymbol{W}^O\in\mathbb{R}^{D_h\times D_h}$为可学习的变换矩阵。第$i$个头$\boldsymbol{U}_i$的运算可表达为：

$$\boldsymbol{U}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}^{q_i},\boldsymbol{K}\boldsymbol{W}^{k_i},\boldsymbol{V}\boldsymbol{W}^{v_i}) \tag{5}$$

$$\mathrm{Attention}(\hat{\boldsymbol{Q}},\hat{\boldsymbol{K}},\hat{\boldsymbol{V}})=\mathrm{Softmax}\left(\frac{\hat{\boldsymbol{Q}}\hat{\boldsymbol{K}}^{\mathrm{T}}}{\sqrt{D_k}}\right)\hat{\boldsymbol{V}} \tag{6}$$

式中：$\boldsymbol{W}^{q_i}$、$\boldsymbol{W}^{k_i}$、$\boldsymbol{W}^{v_i}\in\mathbb{R}^{D_h\times D_k}$都是可学习的变换矩阵，$D_k=D_h/n$是每个头的特征向量维度，上标$q_i$、$k_i$、$v_i$分别代表该参数属于第$i$个头的质询矩阵、键矩阵和值矩阵。

1.2 基于conformer的多通道语音识别系统

图1(b)为多通道系统，图中的编码器、除最后一层的其他解码器以及输出层由单通道识别系统预训练得到，并在训练多通道系统时，这些模块的参数将会固定且被所有通道共享，图中虚线直角矩形框内为流注意力模块，在第二阶段使用多通道带噪数据训练。

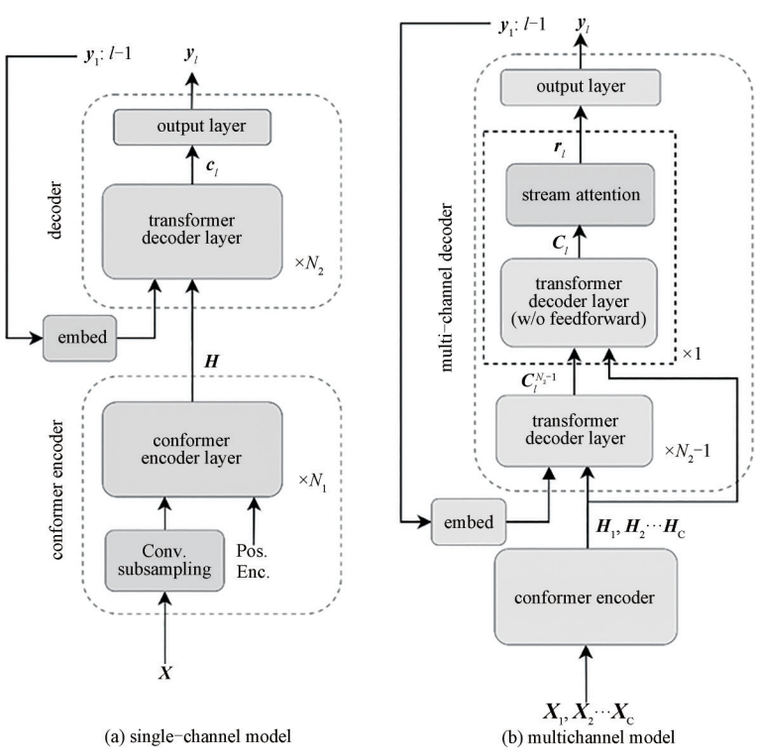

Fig.1 Conformer−based ad−hoc ASR systems
图1 基于conformer的自组织语音识别系统

多通道系统的结构描述如下：给定一个语料所有通道的输入声学特征$\boldsymbol{X}_k\in\mathbb{R}^{T\times D_x},k=1,2,\cdots,C$，其中下标$k$指定某个特定通道，$C$代表总输入通道数，可以分别得到每一个通道的高维表征$\boldsymbol{H}_k$：

$$\boldsymbol{H}_k=\mathrm{Enc}_{1:N_1}(\tilde{\boldsymbol{X}}_k),k=1,2,\cdots,C \tag{7}$$

然后将在时间步 l 得到的每个通道的语义向量进行拼接：

$$\boldsymbol{C}_l=\mathrm{Concat}(\boldsymbol{c}_{l,1},\boldsymbol{c}_{l,2},\cdots,\boldsymbol{c}_{l,C}) \tag{8}$$

式中：

$$\boldsymbol{c}_{l,k}=\widehat{\mathrm{Dec}}_{N_2}(\boldsymbol{c}_{l,k}^{N_2-1},\boldsymbol{H}_k,\boldsymbol{H}_k) \tag{9}$$

$$\boldsymbol{c}_{l,k}^{N_2-1}=\mathrm{Dec}_{1:N_2-1}(\boldsymbol{H}_k,\mathrm{Emb}(\boldsymbol{y}_{1:l-1})) \tag{10}$$

式中：$\boldsymbol{c}_{l,k}^{N_2-1}$ 为经过解码器第 1 到 N_2-1 层后的输出；$\widehat{\mathrm{Dec}}_{N_2}$ 代表缺少前向输出层的第 N_2 层解码器。同时，由之前时间步的输出向量提取导向矢量 $\boldsymbol{g}_l\in\mathbb{R}^{D_k}$：

$$\boldsymbol{g}_l=\mathrm{MHA}\left[\mathrm{Emb}(\boldsymbol{y}_{l-1}^{\mathrm{T}}),\mathrm{Emb}(\boldsymbol{y}_{1:l-1}),\mathrm{Emb}(\boldsymbol{y}_{1:l-1})\right] \tag{11}$$

导向矢量 $\boldsymbol{g}_l$ 一方面作为解码器块的输入，另一方面将作为流注意力的输入。

2 流注意力及其变体

2.1 流注意力简述

流注意力的计算定义为：

$$\text{Stream Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{Z}+\text{FeedForward}(\boldsymbol{Z}) \tag{12}$$

式中：$\boldsymbol{Z}=\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$；FeedForward(·)是前向输出模块。

流注意力将每个通道的高维语义向量 $\boldsymbol{C}_l$ 和式(11)中定义的导向矢量 $\boldsymbol{g}_l$ 作为输入，得到融合语义向量 $\boldsymbol{r}_l$：

$$\boldsymbol{r}_l=\text{Stream Attention}(\boldsymbol{g}_l^{\mathrm{T}},\boldsymbol{C}_l,\boldsymbol{C}_l) \tag{13}$$

最后，将融合语义向量 $\boldsymbol{r}_l$ 通过输出层，获取当前时间步的输出向量 $\boldsymbol{y}_l$。

图 2(a)为基于 Softmax 的流注意力结构，该结构将 R Li 等[16]提出的循环神经网络架构更换为 conformer 架构。

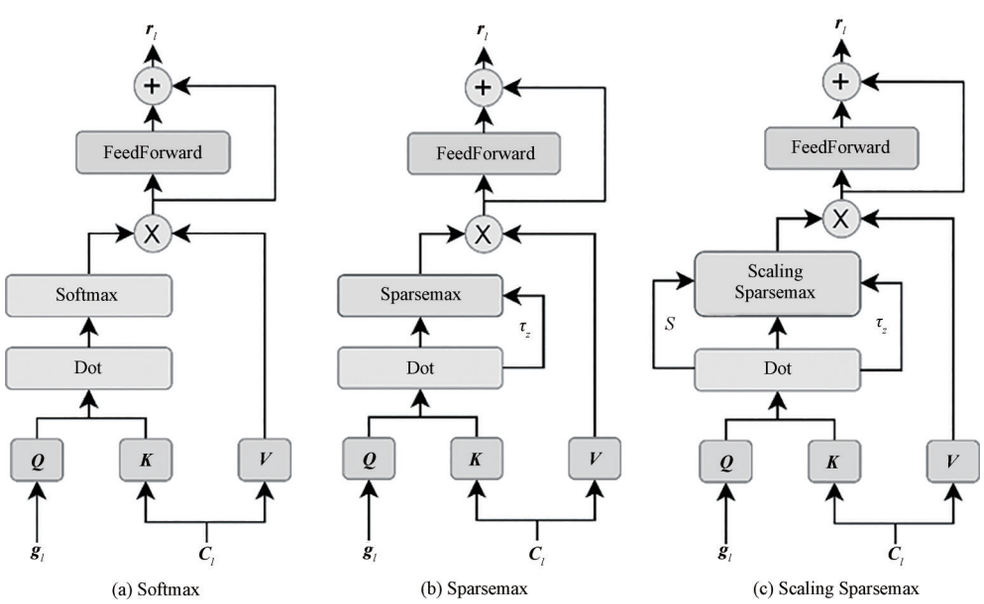

Fig.2 Three kinds of stream attention architectures

图 2 三种流注意力结构

2.2 基于 Sparsemax 的流注意力

基于 Softmax 的流注意力在多智能体自组织语音识别任务下存在一定的局限：对于任意的 $\boldsymbol{z}$ 和 i，$\mathrm{Softmax}_i(\boldsymbol{z})\neq 0$，导致这个方法不能进行通道选择。为解决这个问题，本文提出了基于 Sparsemax[18]的流注意力，

其结构如图2(b)所示，其中Sparsemax的定义如下：

$$\text{Sparsemax}(\boldsymbol{z})=\text{argmin}_{\boldsymbol{p}\in\Delta^{K-1}}\left\|\boldsymbol{p}-\boldsymbol{z}\right\|^2 \tag{14}$$

式中 $\Delta^{K-1}=\{\boldsymbol{p}\in\mathbb{R}^K\left|\sum_{i=1}^{K}p_i=1,p_i\geqslant 0\right.\}$ 代表一个 $K-1$ 维的单纯形，$\boldsymbol{p}$ 是定义在 K 维空间中的向量。Sparsemax本质上是将输入向量 $\boldsymbol{z}$ 投影到设定的单纯形上，由于单纯形的特性，可以得到一个稀疏的输出向量。同时，这个输出向量有一个近似解：

$$\text{Sparsemax}_i(\boldsymbol{z})=\max(z_i-\tau(\boldsymbol{z}),0) \tag{15}$$

式中 $\tau:\mathbb{R}^K\to\mathbb{R}$ 代表一个搜寻软阈值的函数。

2.3 基于Scaling Sparsemax的流注意力

从上一节得知，Sparsemax的输出与输入向量和单纯形的维度密切相关。由于麦克风位置的随机性，导致输入向量的值会在较大范围内波动。同时，单纯形的维度与输入通道数相关，也是一个变量。因此，在某些情况下，Sparsemax的泛化能力可能会下降。为此，本文提出了Scaling Sparsemax，如图2(c)所示。

加入一个可训练的可缩放因子 s，重新标定Sparsemax：

$$s=1+\text{ReLU}\left\{\text{Linear}\left(\left[\|\boldsymbol{z}\|,C\right]^{\mathrm{T}}\right)\right\} \tag{16}$$

式中：$\|\boldsymbol{z}\|$ 为输入向量的L2范数；Linear(·)为两层可学习线性变换，其维度分别为 2×2 和 1×2。

Scaling Sparsemax算子的流程如下。当 $s=1$ 时，Scaling Sparsemax退化为Sparsemax。

Require：$\boldsymbol{z}$,s
  Sort $\boldsymbol{z}$ as $\boldsymbol{z}(1)\geqslant\cdots\geqslant\boldsymbol{z}(K)$
  Initialize $k\leftarrow K$
  while $k>0$ do
    if $\boldsymbol{z}(K)\geqslant\left(\sum_{i=1}^{k}\boldsymbol{z}(i)-s\right)/k$ then
      $\tau(\boldsymbol{z}):=\left(\sum_{i=1}^{k}\boldsymbol{z}(i)-s\right)/k$
      Break
    end if
    $k\leftarrow k-1$
  end while
Ensure: $\boldsymbol{p}$ where $p_i=\max(\boldsymbol{z}(i)-\tau(\boldsymbol{z}),0)/s$

2.4 多层流注意力

在多智能体的应用中，降低计算复杂度也是一个重要的研究方向。实验中发现Scaling Sparsemax学习的可缩放因子具备对不同通道和不同向量L2范数的泛化能力，图3展示了它们的关系。从图中可以看出，缩放因子大小与通道数量成正比，缩放因子越小，最终保留的通道也越少。

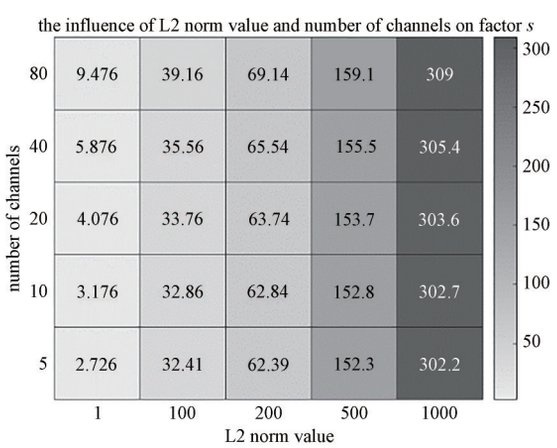

Fig.3 Visualization of scaling factor s
图3 缩放因子 s 可视化

基于上述关系，本文设计了一个多层的流注意力结构。相对于单层流注意力通过 N 层解码器后再进行1次通道丢弃操作，多层流注意力结构在 N 层解码器内部进行 N 次通道丢弃操作。具体而言，若当前层通道分配的权重为0，则直接丢弃而不再输入下一层，使通道数逐层减少，同时缩放因子大小也会逐层减小，达到每一层逐渐丢弃通道并降低下一层计算量的效果。实验表明，该结构可以一定程度地降低计算复杂度和加快解码速度。

3 实验与分析

3.1 实验设置

实验使用了3个数据集：Librispeech语音识别语料库[19]、基于Librispeech仿真的自组织麦克风阵列多通道数据(Libri-adhoc-simu)、基于Librispeech回放的真实环境下40通道自组织麦克风阵列数据(Libri-adhoc40)[20]。在Libri-adhoc-simu和Libri-adhoc40数据集中，考虑每一个智能体都只使用一个麦克风进行拾音，因此通道数量就指代多智能体的数量。Librispeech包含1 000 h，来自2 484个说话人的英文朗读语音。实验中，选择了960 h的数据去训练单通道的语音识别系统，选择了10 h的数据进行验证。

Libri-adhoc-simu使用Librispeech的一个包含100 h数据的子集“train-clean-100”作为训练集，选取包含10 h数据的子集“dev-clean”和“dev-other”作为验证集，选取分别包含5 h数据的子集“test-clean”和“test-other”作为2个测试集。对于每一句语料，都随机生成一个房间，房间的长和宽在[5,25] m内均匀选取，高在[2.7,4] m均匀选取。房间中放置多个麦克风以及一个扬声器。限制声源到墙壁的距离大于0.2 m，声源到麦克风的距离大于0.3 m。使用理想源模型仿真混响环境，并在[0.2,0.4] s范围内选择混响时间。同时，使用背景噪声模型仿真不相关的背景噪声，训练集和验证集的噪声来自一个大规模的噪声数据库，其包含超过20 000条噪声[21]，测试集的噪声来自CHiME-3数据集[22]和NOISEX-92语料库[23]。在训练集中随机生成16个通道的数据，在测试集中随机生成10、16和20个通道的数据。

Libri-adhoc40是在一个大房间中，用Librispeech的子集“train-clean-100”，“dev-clean”和“test-clean”回放得到[20]。录音的环境是一个包含40个麦克风和一个扬声器的真实办公室环境，该环境下存在较强混响和较弱的加性噪声。在训练、验证和测试集中，扬声器和麦克风的摆放位置都有区别，扬声器的摆放共有17个不同的位置，其中训练集占9个，验证集占4个，测试集占4个，扬声器和麦克风的距离在[0.8,7.4] m范围内。随机选择20个通道用于训练和验证，随机选取16、20和32个通道以得到3种不同的测试场景。

表1为声学特征和模型的结构。在训练阶段，首先用干净的Librispeech数据训练了一个基于conformer的单通道语音识别系统。模型训好后，将其参数固定并送到基于conformer的多通道语音识别系统的每一个通道中。最后，用多通道带噪数据训练流注意力。在测试阶段，采用不使用语言模型的贪婪解码，用词错误率作为评价指标。

表1 声学特征和模型结构介绍

Table1 Descriptions of the acoustic feature and the model structure

module name	parameter name and value			
acoustic feature	type:fbank	# dim (D_x): 80	augmentation: SpecAugment[24]	
conformer	Enc. blocks (N_1):12	Dec. blocks (N_2): 6	vocabulary size(D_x): 5 000	# Feed. dim: 2 048
multi-head attention	# head:8	# dim (D_h): 512		
stream attention	# head:1	# dim (D_h): 512		

对基于Softmax的流注意力与基于Sparsemax和Scaling Sparsemax的流注意力的性能进行比较。同时，构造了一个称为理想最优通道(oracle one best)的基线，该基线人为选取了离扬声器最近的麦克风作为单通道语音识别系统的输入获得结果(该基线需要知道麦克风和扬声器位置的先验信息)。

3.2 实验结果分析

表2为Libri-adhoc-simu数据上各个方法的性能比较。从表中可以发现：a) 3种流注意力模型在2种测试集中都表现出优良的性能。其中模型在16通道下训练，分别在适配的16通道和失配的10通道和20通道下测试，可以发现，即使在失配通道数的情况下，模型依旧表现出强大的泛化能力。同时可以发现，20通道环境下的性能优于10通道和16通道，印证了在多智能体自组织语音识别下，增加多智能体数量的优势。b) 基于Sparsemax和Scaling Sparsemax的流注意力模型相比Softmax，性能上获得了显著提升。如在20通道环境的“test-clean”数据集下，Scaling Sparsemax相比Softmax，词错误率降低了30.9%；在20通道环境的“test-other”数据集下，词错误率降低了22.7%。

表3为Libri-adhoc40半真实数据集上的实验结果。从表中可以看出，在实际场景下，模型同样具备强大的泛化能力(模型在20通道训练，分别在16、20、32通道下测试)，在失配的32通道环境下达到了最佳效果。同时本文提出的Scaling Sparsemax方法性能最佳，相比于“理想最优通道”的词错误率，其在16通道环境下降低了17.9%，在32通道环境下，降低了15.8%。

表2 Libri-adhoc-simu数据集上结果比较(文字差错率(%))

Table2 Comparison results on Libri-adhoc-simu(WER(%))

method	test-clean			test-other		
	10-channel	16-channel	20-channel	10-channel	16-channel	20-channel
Oracle one best	19.3	14.3	13.1	35.7	30.1	28.4
Softmax	19.5	15.4	13.9	38.5	33.7	32.1
Sparsemax(proposed)	16.0	11.5	10.0	33.9	27.5	26.1
Scaling Sparsemax(proposed)	14.7	10.7	9.6	31.6	26.5	24.8

表3 Libri-adhoc40数据“test-clean”子集上结果比较(文字差错率(%))

Table3 Comparison results on Libri-adhoc40(WER(%))

method	16-channel	20-channel	32-channel
oracle one best	31.2	28.2	22.2
Softmax	31.7	29.7	24.9
Sparsemax(proposed)	38.8	33.7	29.9
Scaling Sparsemax(proposed)	25.6	23.3	18.7

表4 仿真与半真实数据“test-clean”子集下多层流注意力结果(文字差错率(%))

Table4 Multi-layer stream attention results under “test-clean”subset of simulation and semi-real data(WER(%))

method	simulation data		semi-real data		computational complexity
	10-channel	20-channel	16-channel	32-channel	
Sparsemax	16.0	10.0	38.8	29.9	$\mathcal{O}(\tilde{C}\times L^2)$
+Scaling	14.7	9.6	25.6	18.7	$\mathcal{O}(\tilde{C}\times L^2)$
+multi-layer	15.5	9.6	25.3	19.9	$\mathcal{O}(\log \tilde{C}\times L^2)$

表4为多层流注意力在仿真与半真实数据“test-clean”子集上的结果，其中$\tilde{C}$表示所有层通道数的总和，L代表输出时间总步长。可以看出，由于逐层丢弃通道的策略，多层流注意力明显降低了计算复杂度。相对于单层结果，性能下降6.4%，且在16通道半真实数据上，多层流注意力性能获得提升。

图4为在Libri-adhoc40某一句语料上的通道选择可视化结果，其中每个点代表一个麦克风，正方形点旁的数字代表通道权重在所有帧上的均值。从图中可以看出，Softmax只考虑了通道权重分配，但并没有进行选择；尽管Sparsemax能做到通道选择，但是将过多通道置零；Scaling Sparsemax仅将带噪较强的通道置零，且尽可能保留了通道的信息，获得了最佳的性能。最后可以发现，距离并不是判断语音质量的最佳准则，近处的通道不一定能获得最佳识别效果，证明了基于识别系统训练通道选择算法的优势。

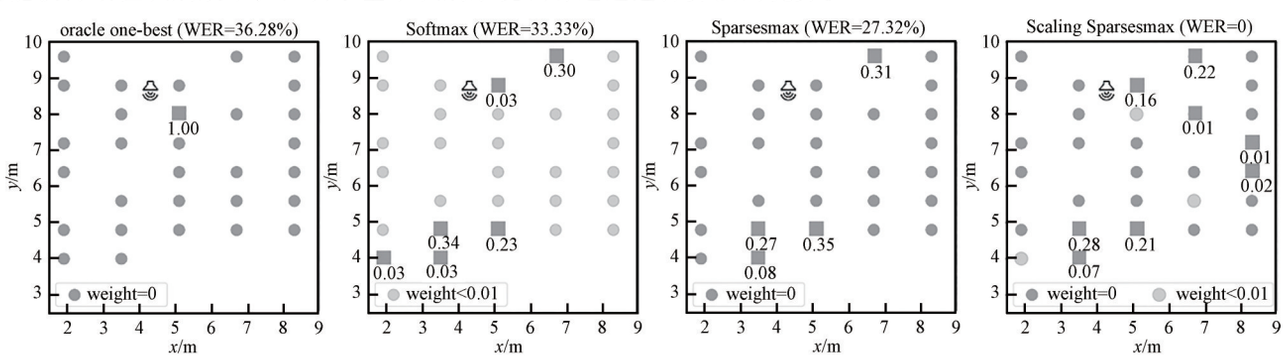

Fig.4 Visualization of the channel selection results on the utterance ID '6930-81414-0024' of Libri-adhoc40

图4 Libri-adhoc40语料“6930-81414-0024”的通道选择可视化

4 结论

本文构建了基于conformer的多智能体自组织语音识别系统并提出2种通道选择算法优化识别系统的性能，即将流注意力中的Softmax算子替换为Sparsemax，使其具备通道选择能力，但由于Sparsemax会将大多数通道权重置零，进一步提出了Scaling Sparsemax算子，仅将带噪严重的通道权重置零。最后，提出了一种多层的流注意力结构，在性能损失不明显的情况下降低了解码的计算复杂度。在带有加性噪声的仿真数据和带有高混响的半真实数据上验证本文的模型。实验结果证明，在仿真和半真实数据上，提出的Scaling Sparsemax流注意力优于Softmax流注意力和理想最优通道；提出的多层流注意结构在极低性能损失的情况下，有效降低了解码的计算复杂度。实验结果同时也证明了选择适合的通道选择算法对多智能体自组织语音识别进行协同控制的重要性。

参考文献：

[1] HEYMANN J, DRUDE L, HAEB-UMBACH R. Neural network based spectral mask estimation for acoustic beamforming[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Shanghai, China: IEEE, 2016: 196-200.

[2] XIAO Xiong, WATANABE S, ERDOGAN H, et al. Deep beamforming networks for multi-channel speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Shanghai, China: IEEE, 2016: 5745-5749.

[3] SAINATH T N, WEISS R J, WILSON K W, et al. Multichannel signal processing with deep neural networks for automatic speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(5): 965-979.

[4] HAEB-UMBACH R, HEYMANN J, DRUDE L, et al. Far-field automatic speech recognition[J]. Proceedings of the IEEE, 2021, 109(2): 124-148.

[5] HEYMANN J, BACCHIANI M, SAINATH T N. Performance of mask based statistical beamforming in a smart home scenario[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Calgary, AB, Canada: IEEE, 2018: 6722-6726.

[6] 李杨, 徐峰, 谢光强, 等. 多智能体技术发展及其应用综述[J]. 计算机工程与应用, 2018, 54(9): 13-21. (LI Yang, XU Feng, XIE Guangqiang, et al. Survey of development and application of multi-agent technology[J]. Computer Engineering and Applications, 2018, 54(9): 13-21.)

[7] DORRI A, KANHERE S S, JURDAK R. Multi-agent systems: a survey[J]. IEEE Access, 2018(6): 28573-28593.

[8] 陈磊, 李钟慎. 多智能体系统一致性综述[J]. 自动化博览, 2018, 35(2): 74-78. (CHEN Lei, LI Zhongshen. Literature review on the consistency of multi-agent systems[J]. Automation Panorama, 2018, 35(2): 74-78.)

[9] RAYKAR V C, KOZINTSEV I V, LIENHART R. Position calibration of microphones and loudspeakers in distributed computing platforms[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(1): 70-83.

[10] ZHANG Xiaolei. Deep ad-hoc beamforming[J]. Computer Speech & Language, 2021(68): 101201.

[11] COSSALTER M, SUNDARARAJAN P, LANE I. Ad-hoc meeting transcription on clusters of mobile devices[C]// The 12th Annual Conference of the International Speech Communication Association-INTERSPEECH 2011. Florence, Italy: ISCA, 2011: 2881-2884.

[12] WOLF M, NADEU C. Channel selection measures for multi-microphone speech recognition[J]. Speech Communication, 2014 (57): 170-180.

[13] BOEDDEKER C, HEITKAEMPER J, SCHMALENSTROEER J, et al. Front-end processing for the CHiME-5 dinner party scenario[C]// The 5th International Workshop on Speech Processing in Everyday Environments(CHiME 2018). Hyderabad, India: ISCA, 2018: 35-40.

[14] WATANABE S, MANDEL M, BARKER J, et al. CHiME-6 challenge: tackling multispeaker speech recognition for unsegmented recordings[EB/OL]. (2020-05-02). https://doi.org/10.48550/arXiv.2004.09249.

[15] LI Ruizhi, WANG Xiaofei, MALLIDI S H, et al. Multi-stream end-to-end speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020(28): 646-655.

[16] LI Ruizhi, SELL G, WANG Xiaofei, et al. A practical two-stage training strategy for multi-stream end-to-end speech recognition[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Barcelona, Spain: IEEE, 2020: 7014-7018.

[17] GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[EB/OL]. (2020-05-16). https://doi.org/10.48550/arXiv.2005.08100.

[18] MARTINS A F T, ASTUDILLO R F. From softmax to sparsemax: a sparse model of attention and multi-label classification[C]// Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA: JMLR.org, 2016: 1614-1623.

[19] PANAYOTOV V, CHEN Guoguo, POVEY D, et al. Librispeech: an ASR corpus based on public domain audio books[C]// 2015 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). South Brisbane, QLD, Australia: IEEE, 2015: 5206-5210.

[20] GUAN Shanzheng, LIU Shupei, CHEN Junqi, et al. Libri-adhoc40: a dataset collected from synchronized ad-hoc microphone arrays[EB/OL]. (2021-04-07). https://doi.org/10.48550/arXiv.2103.15118.

[21] TAN Xu, ZHANG Xiaolei. Speech enhancement aided end-to-end multi-task learning for voice activity detection[EB/OL].

（下转第 1187 页）