自动化所在语音识别研究中

2019-08-05 来源： 自动化研究所

中国科学院自动化研究所智能交互团队在环境鲁棒性、轻量级建模、自适应能力以及端到
关成果将在全球语音学术会议INTERSPEECH2019发表。

现有端到端语音识别系统难以有效利用外部文本语料中的语言学知识，针对这一问题，陶
模外部文本训练语言模型，然后将该语言模型中的知识迁移到端到端语音识别系统中。这种方
进行优化，使语音识别系统输出的分布与外部语言模型输出的分布接近，从而有效提高语音识

语音关键词检测在智能家居、智能车载等场景中有着重要作用。面向终端设备的语音关键
主流的基于残差神经网络的语音关键词检测，需要20万以上的参数，难以在终端设备上应用。
值自注意力机制和时延神经网络的轻量级语音关键词检测方法。该方法采用时延神经网络进行
方法，将自注意力机制中的多个矩阵共享，使其映射到相同的特征空间，从而进一步压缩了模
检测模型相比，他们提出的方法在识别准确率接近的前提下，模型大小仅为残差网络模型的1

针对RNN-Transducer模型存在收敛速度慢、难以有效进行并行训练的问题，陶建华、易
模型，主要在以下三个方面实现了改进：（1）通过自注意力机制替代RNN进行建模，有效提
和解码，进一步引入了Chunk-Flow机制，通过限制自注意力机制范围对局部依赖信息进行建
CTC-CE联合优化启发，将交叉熵正则化引入到SA-T模型中，提出Path-Aware Regularizati
重点优化该路径。经验证，上述改进有效提高了模型训练速度及识别效果。

语音分离又称为鸡尾酒会问题，其目标是从同时含有多个说话人的混合语音信号中分离出
重影响语音识别和说话人识别的性能。目前解决这一问题的两种主流方法分别是：深度聚类
permutation invariant training）准则算法。深度聚类算法在训练过程中不能以真实的干净
特征区分性不足。针对DC和PIT算法的局限性，陶建华、刘斌、范存航等人提出了基于区分性
具有区分性的深度嵌入式特征，然后将该特征输入到PIT算法中进行语音分离。同时，为了增
分性学习目标准则，进一步提升算法的性能。所提方法在WSJ0-2mix语音分离公开数据库上获

端到端系统在语音识别中取得突破。然而在复杂噪声环境下，端到端系统的鲁棒性依然面
刘斌等人提出了基于联合对抗增强训练的鲁棒性端到端语音识别方法。具体地说，使用一个基
和判别网络的联合优化方案。判别网络用于区分经过语音增强网络之后的频谱和纯净语音的频
联合优化识别、增强和判别损失，神经网络自动学习更为鲁棒的特征表示。所提方法在aishe

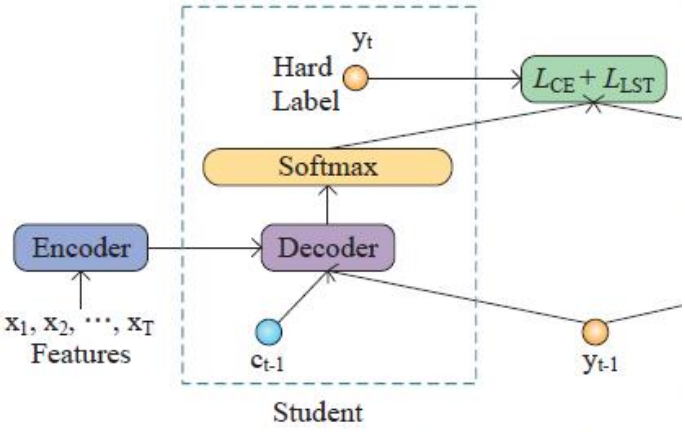
说话人提取是提取音频中目标说话人的声音。与语音分离不同，说话人提取不需要分离出
的说话人提取方法是：说话人波束（SpeakerBeam）和声音滤波器（Voice filter）。这两种
性。因为声源是有方向性的，并且在实际环境中是空间可分的。所以，如果正确利用多通道的
有效利用多通道的空间特性，刘文举、梁山、李冠君等人提出了方向感知的多通道说话人提取
同方向的波束。进而DNN采用attention机制来确定目标信号所在的方向，来增强目标方向的
目标信号。提出的算法在低信噪比或同性别说话人混合的场景中性能提升明显。

传统的对话情感识别方法通常从孤立的句子中识别情感状态，未能充分考虑对话中的上下
刘斌、连政等人提出了一种融合上下文信息的多模态情感识别方法。在输入层，采用注意力机
力机制的双向循环神经网络对长时上下文信息进行建模；为了能够有效模拟真实场景下的交互
交互过程中的身份信息。在IEMOCAP情感数据集上对算法进行了评估，实验结果表明，该方

由于情感数据标注困难，语音情感识别面临着数据资源匮乏的问题。虽然采用迁移学习方
解低资源的问题，但是这类方法并没有关注到长时信息对语音情感识别的重要作用。针对这一
（Future Observation Prediction, FOP）的无监督特征学习方法。FOP采用自注意力机制，
（Hypercolumns）两种迁移学习方法，能够将FOP学习到的知识用于语音情感识别。该方法
音情感识别。

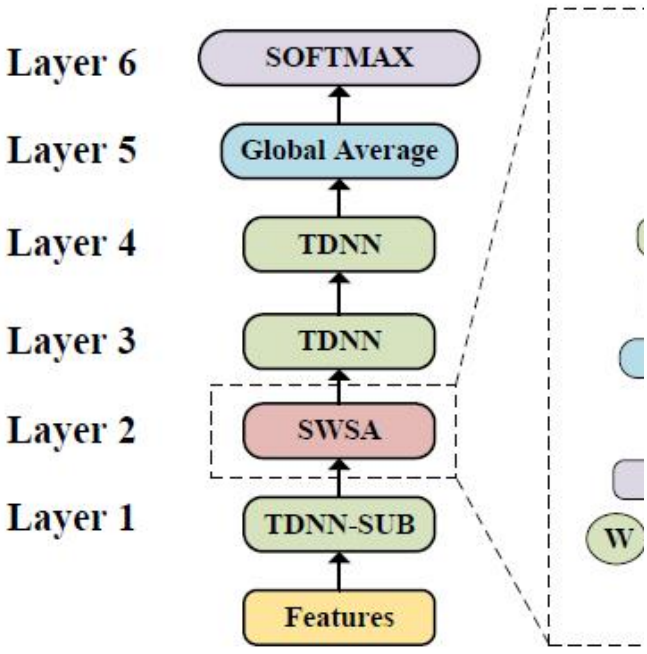
相关生理学研究表明，MFCC (Mel-frequency cepstral coefficient)对于抑郁检测来说是
来辨识个体的抑郁程度。但是，上述工作中很少使用神经网络来进一步捕获MFCC中反映抑郁
池化参数未能被有效优化。针对上述问题，陶建华、刘斌、牛明月等人提出了一种混合网络并

的lp范数池化方法来提升抑郁检测的性能。首先将整段音频的MFCC切分成具有固定大小的长
列的空间结构、时序变化以及区分性表示与抑郁线索相关的信息，并将所抽取的特征记为段级
一步聚合为表征原始语音句子级的特征。

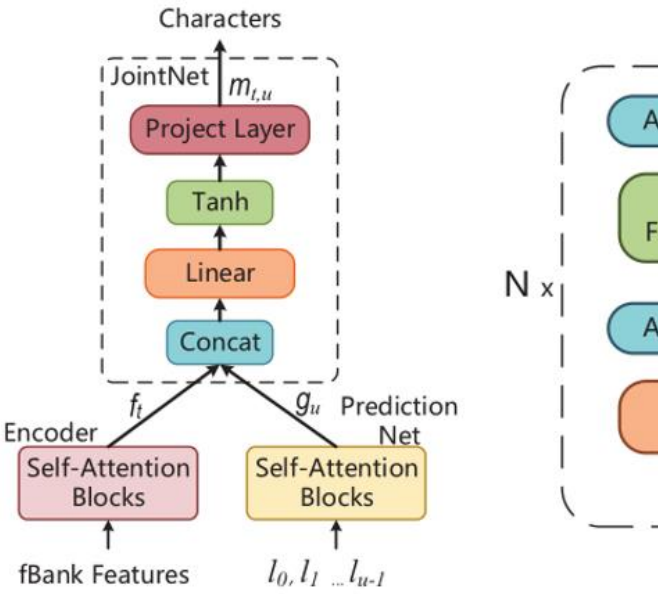


(b) *Learn Spelling from Teacher*

基于知识迁移的端到端语音识别

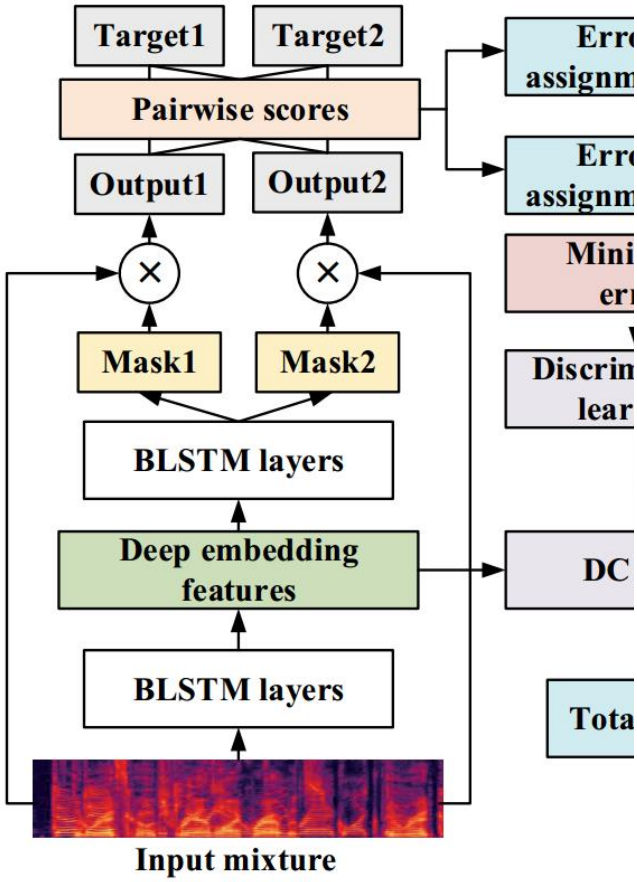


基于共享权值自注意力机制和时延神经网络的轻

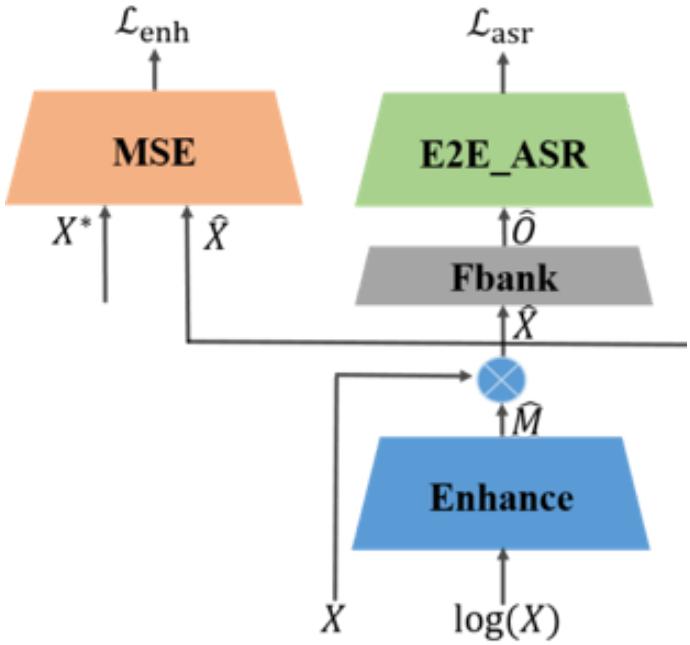


(a) Self-Attention Transducer (b) Sel

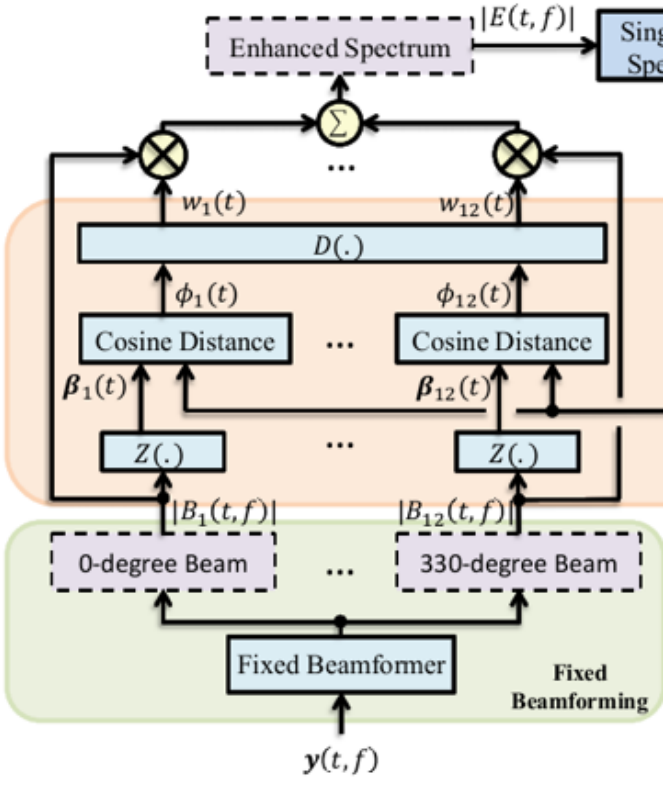
基于自注意力机制的端到端语音转



基于区分性学习和深度嵌入式特征的语音分

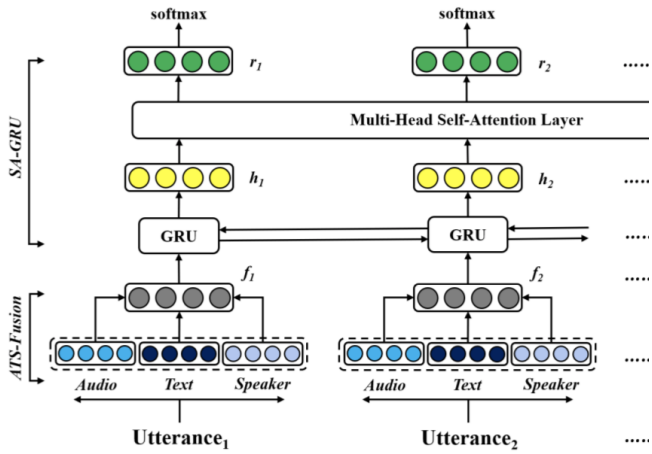


基于联合对抗增强训练的鲁棒性端到端语音

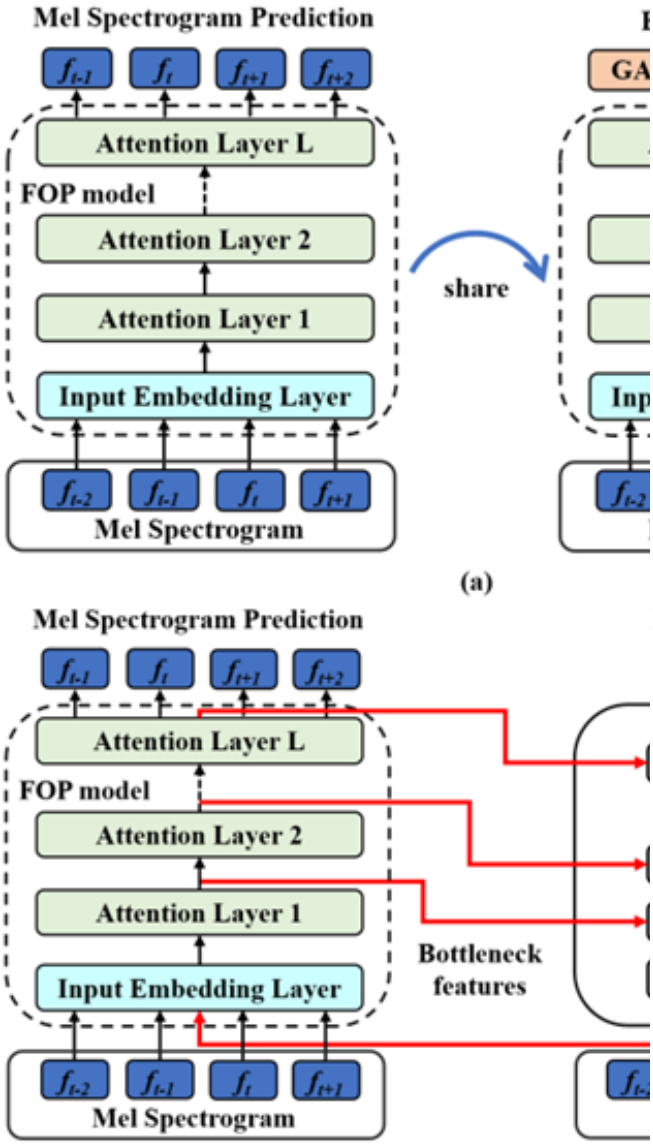


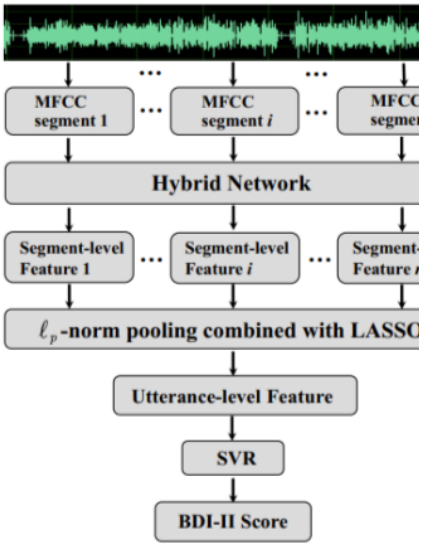
* The content in the dotted box is in the form of the

方向感知的多通道说话人提取方法



融合上下文信息的多模态情感识





基于混合神经网络结合ℓp范数池化的自

上一篇： 版纳植物园古大气二氧化碳浓度重建的代理指标研究获进展

下一篇： 地质地球所提出一种最小二乘全走时反演法用于浅层地震速度建模