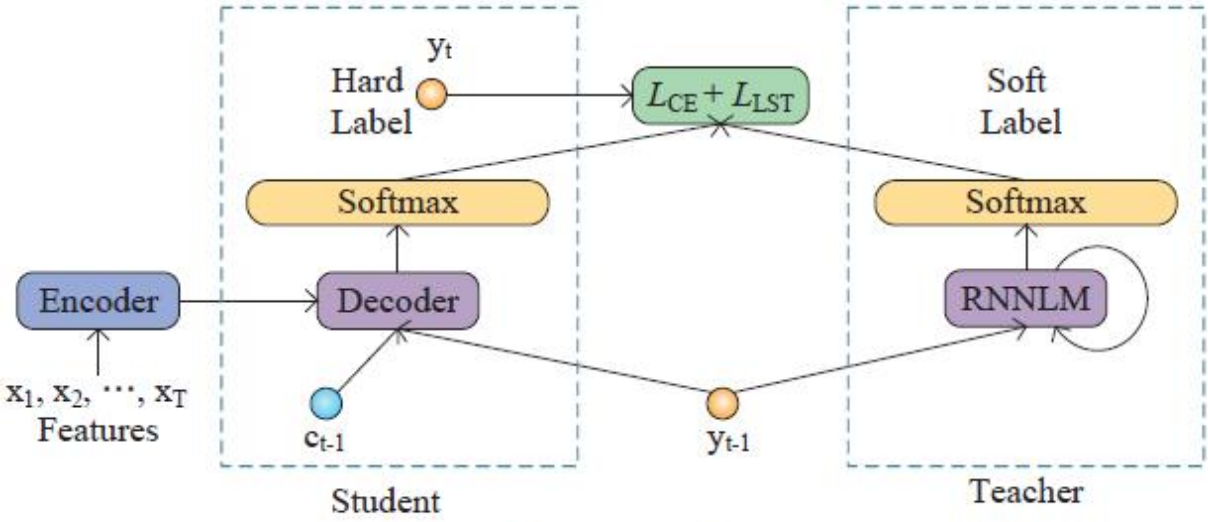
科研动态

智能交互团队在语音识别方向获新进展

发表日期：2019-07-25 【大 中 小】【打印】【关闭】

我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端处理等几个方面进行持续攻关，在语音识别方面获新进展，相关成果将在全球语音顶级学术会议INTERSPEECH2019发表。

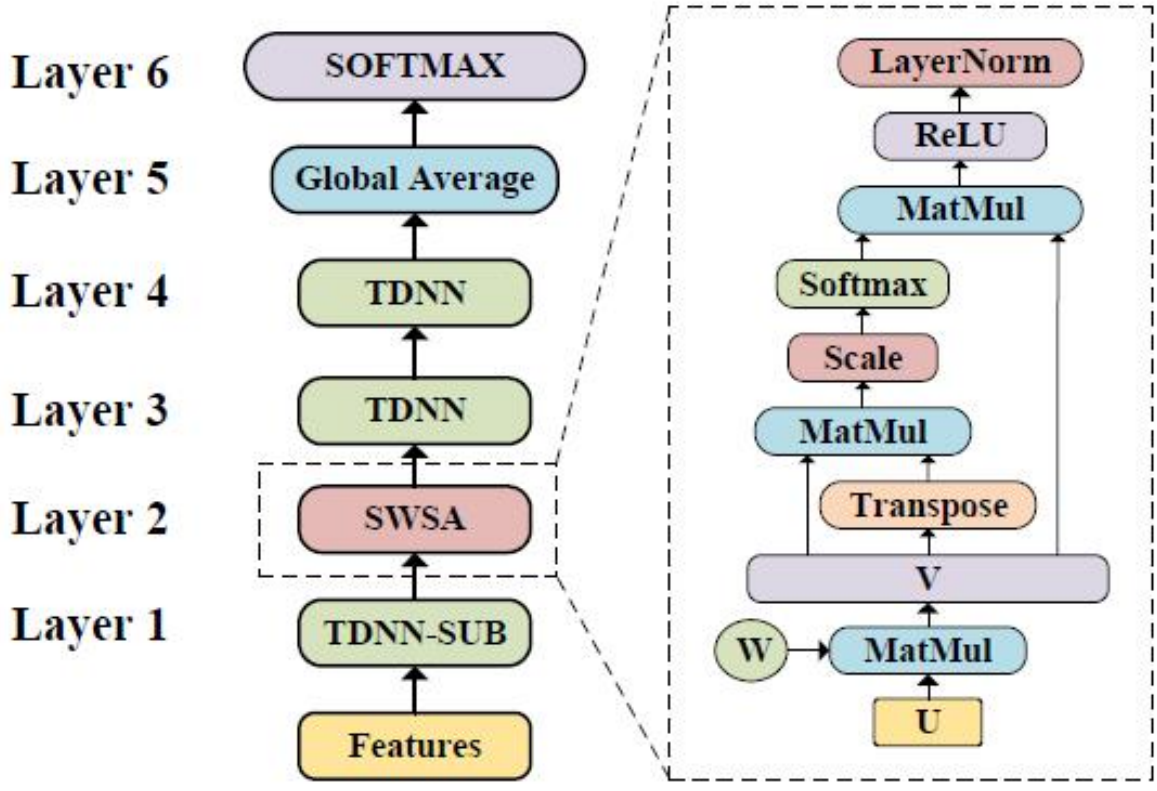
现有端到端语音识别系统难以有效利用外部文本语料中的语言学知识，针对这一问题，陶建华、易江燕、白烨等人提出采用知识迁移的方法，首先对大规模外部文本训练语言模型，然后将该语言模型中的知识迁移到端到端语音识别系统中。这种方法利用了外部语言模型提供词的先验分布软标签，并采用KL散度进行优化，使语音识别系统输出的分布与外部语言模型输出的分布接近，从而有效提高语音识别的准确率。



(b) *Learn Spelling from Teachers (LST)*

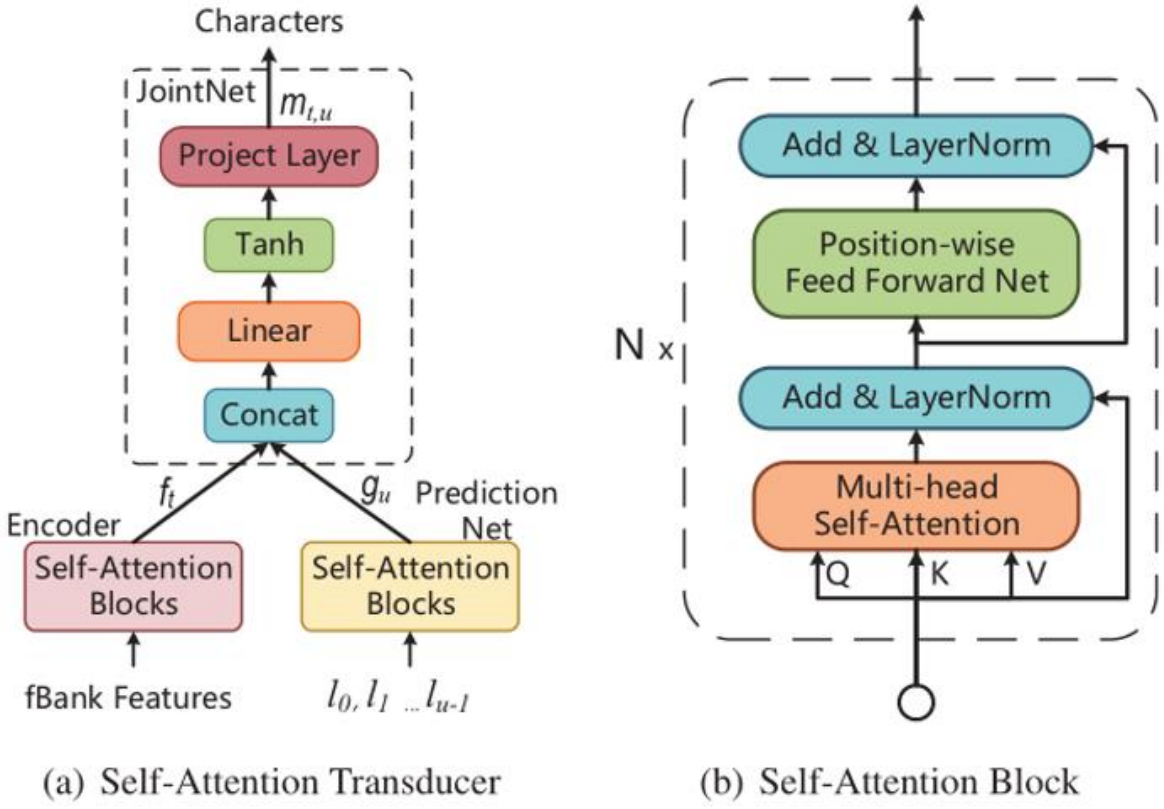
基于知识迁移的端到端语音识别系统

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词检测对算法的时间复杂度和空间复杂度有着很高的要求。当前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。 为了解决这一问题，陶建华、易江燕、白烨等人提出基于共享权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降采样，通过自注意力机制捕获时序相关性；并采用共享权值的方法，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型的尺寸。与目前的性能最好的基于残差神经网络的语音关键词检测模型相比，我们提出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1/20，有效降低了算法复杂度。



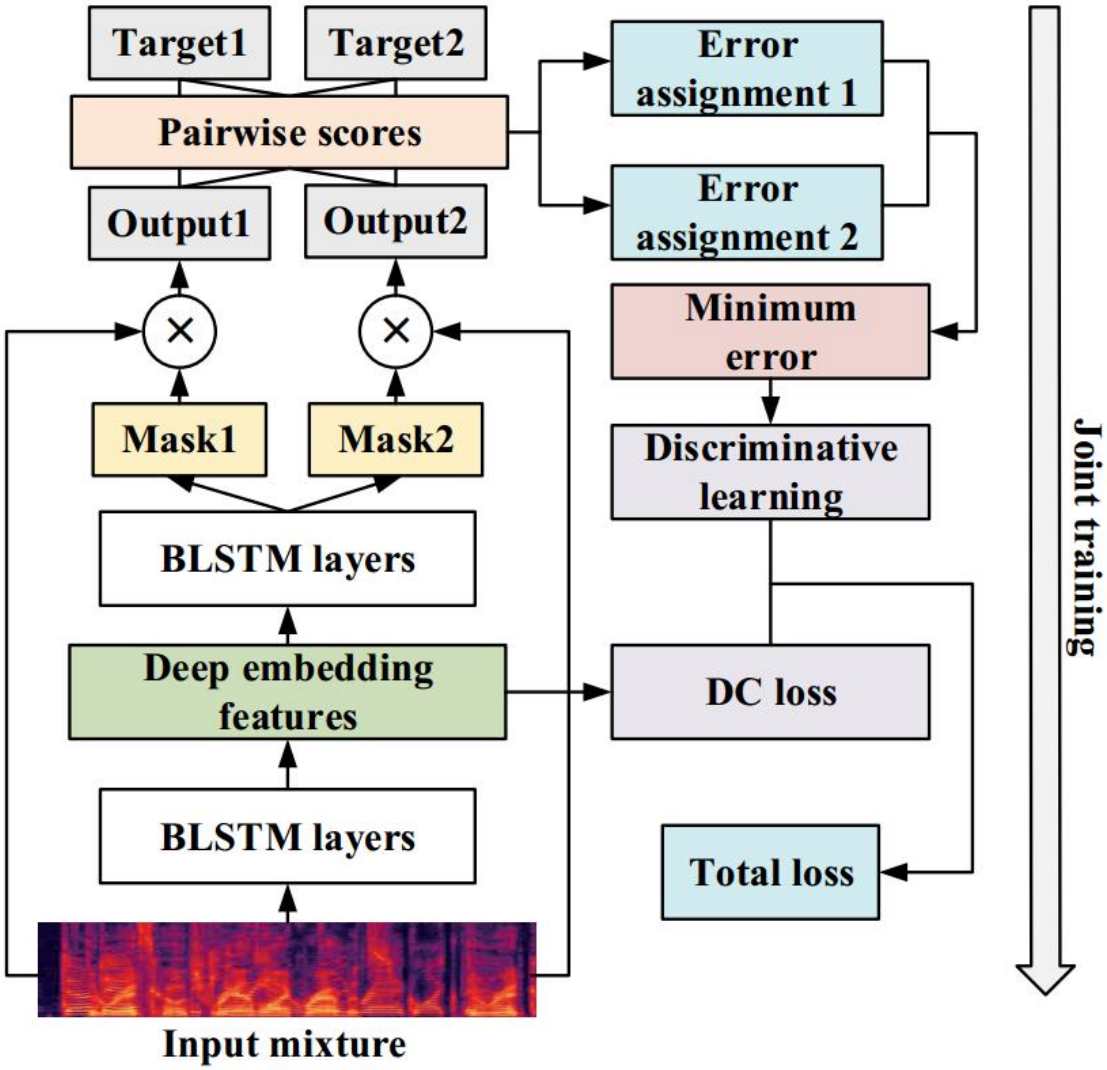
基于共享权值自注意力机制和时延神经网络的轻量级语音关键词检测

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、易江燕、田正坤等人提出了一种Self-attention Transducer (SA-T)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，有效提高了模型训练的速度； （2）为了使SA-T能够进行流式的语音识别和解码，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模，并通过堆叠多层网络对长距离依赖信息进行建模； （3）受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 Regularization(PAR)，通过先验知识引入一条可行的对齐路径，在训练过程中重点优化该路径。 经验证，上述改进有效提高了模型训练速度及识别效果。



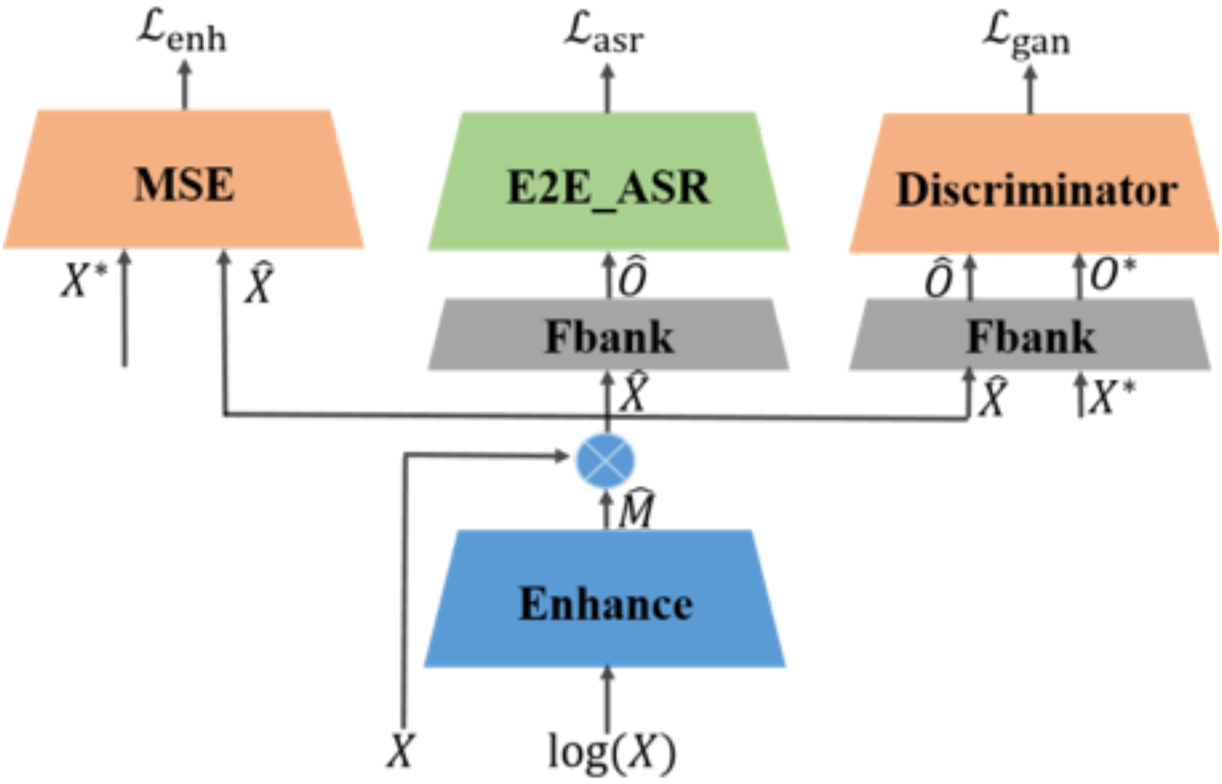
基于自注意力机制的端到端语音转写模型

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不同说话人的信号。当一段语音中同时含有多个说话人时，会严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（DC, deep clustering）算法和排列不变性训练（PIT, 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语音作为目标，性能受限于k-means聚类算法；而PIT算法其输入特征区分性不足。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学习和深度嵌入式特征的语音分离方法。首先，利用DC提取一个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，为了增大不同说话人之间的距离，减小相同说话人之间的距离，引入了区分性学习目标准则，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得较大的性能提升。



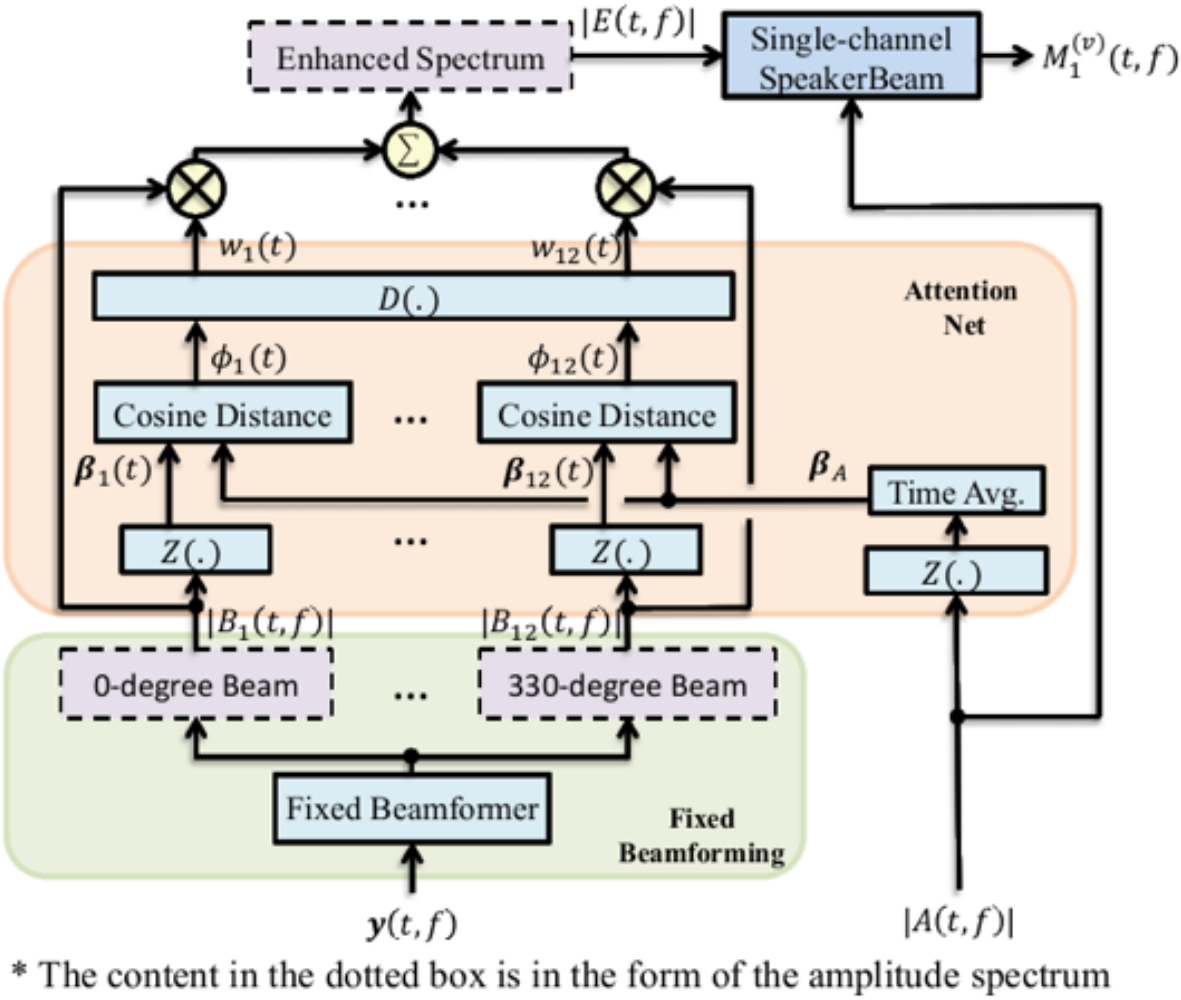
基于区分性学习和深度嵌入式特征的语音分离方法总体框图

端到端系统在语音识别中取得了重大的突破。然而在复杂噪声环境下，端到端系统的鲁棒性依然面临巨大挑战。针对端到端系统不够鲁棒的问题，刘文举、聂帅、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，使用一个基于mask的语音增强网络、基于注意力机制的的端到端语音识别网络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谱，可以引导语音增强网络的输出更加接近纯净语音分布。通过联合优化识别、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aishell-1数据集上面取得了较大的性能提升。



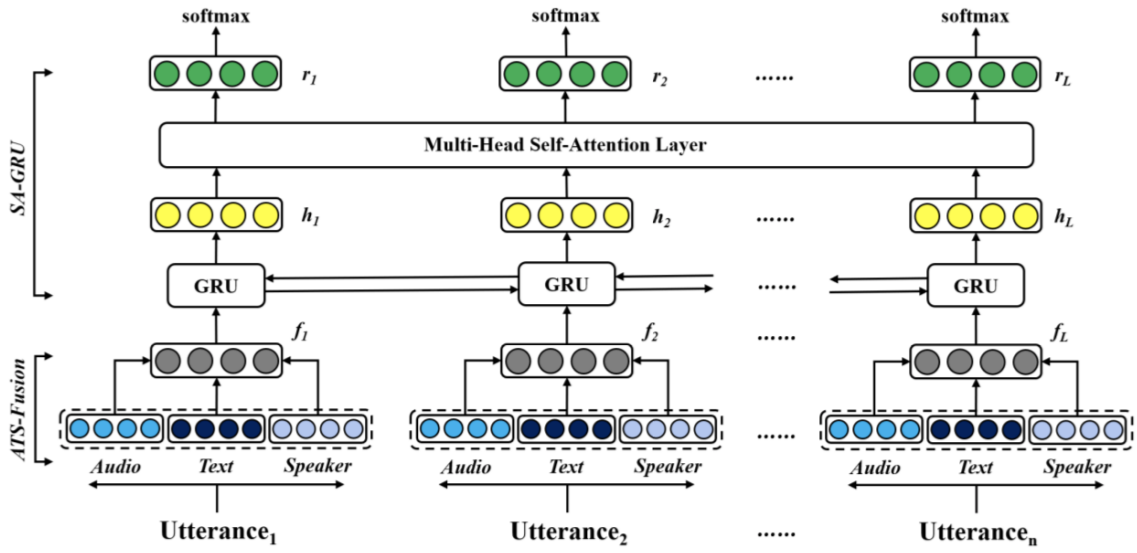
基于联合对抗增强训练的鲁棒性端到端语音识别总体框图

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音频中所有说话人的声音，而只关注某一特定说话人。目前主流的说话人提取方法是：说话人波束（SpeakerBeam）和声音滤波器（Voice filter）。这两种方法都只关注声音的频谱特征，而没有利用多通道信号的空间特性。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空间区分性，说话人提取系统可以更好地估计目标说话人。为了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方法。首先多通道的信号先经过一组固定波束形成器，来产生不同方向的波束。进而DNN采用attention机制来确定目标信号所在的方向，来增强目标方向的信号。最后增强后的信号经过SpeakerBeam通过频谱线索来提取目标信号。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。



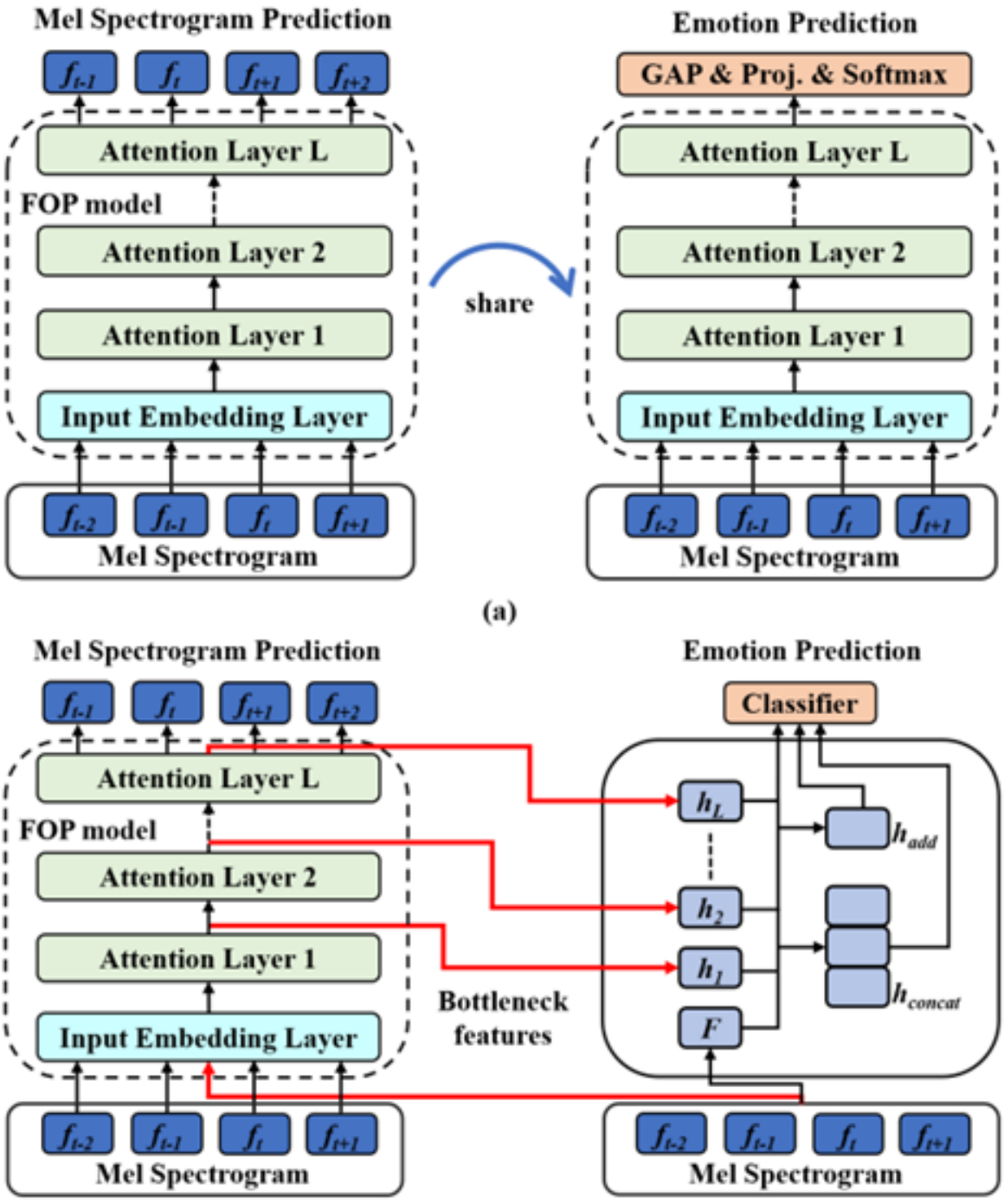
方向感知的多通道说话人提取方法框图

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文信息对于当前时刻情感状态的影响。针对这一问题，陶建华、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，采用注意力机制对文本特征和声学特征进行融合；在识别层，采用基于自注意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模式，引入身份编码向量作为额外的特征输入到模型，用于区分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，该方法相比现有最优基线方法，在情感识别性能上提升了2.42%。

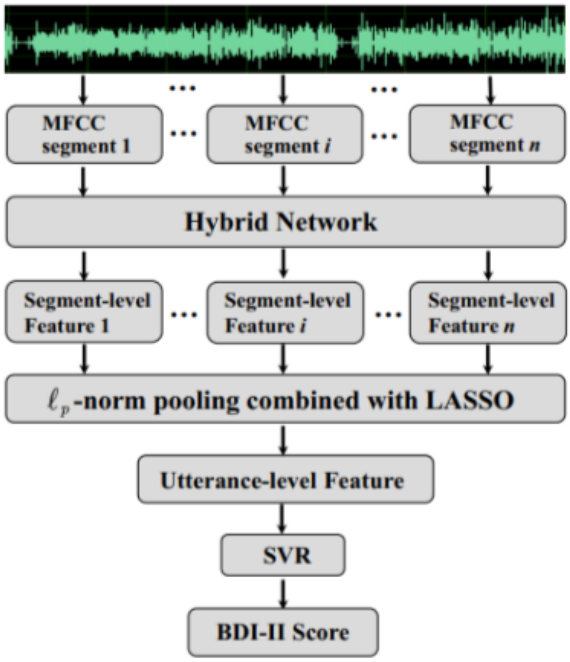


融合上下文信息的多模态情感识别

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法，将其他领域知识迁移到语音情感识别，可以在一定程度上缓解低资源的问题，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。针对这一问题，陶建华、刘斌、连政等人提出了一种基于未来观测预测（Future Observation Prediction, FOP）的无监督特征学习方法。FOP采用自注意力机制，能够有效捕获长时信息；采用微调（Fine-tuning）和超列（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。该方法在IEMOCAP情感数据集上的性能超过了基于无监督学习策略的语音情感识别.



相关生理学研究表明，MFCC (Mel-frequency cepstral coefficient)对于抑郁检测来说是一种有区分性声学特征,这一研究成果使得不少工作通过MFCC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中反映抑郁程度的高表征特征；此外，针对抑郁检测这一问题，合适的特征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结合LASSO (least absolute shrinkage and selection operator)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度；然后将这些切分的片段输入到混合神经网络中以挖掘特征序列的空间结构、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别的特征；最后结合LASSO的lp范数池化将这些段级别的特征进一步聚合为表征原始语音句子级的特征。



基于混合神经网络结合ℓp范数池化的自动抑郁检测算法

相关文献：

Learn Spelling from Teachers: Integrating Language Models into Sequence-to-Sequence Models

A Time Delay Neural Network with Shared Weight Self-Attention for Small-Footprint Keyword Spotting

Self-Attention Transducers for End-to-End Speech Recognition

Discrimination Learning for Monaural Speech Separation Using Deep Embedding Features

Jointly Adversarial Enhancement Training for Robust End-to-End Speech Recognition

Direction-aware Speaker Beam for Multi-channel Speaker Extraction

Conversational Emotion Analysis via Attention Mechanisms

Unsupervised Representation Learning with Future Observation Prediction for Speech Emotion Recognition

Automatic Depression Level Detection via Lp-norm Pooling