SER 系统通常把语音作为**声学特征帧序列**来处理，而这些帧(frame-level)最终必须**汇聚**为一个话语级（utterance-level）的表示以供分类。早期模型（以及Emotion2Vec）采用简单的池化方式（常见如Max Pooling,Average Pooling,Statistics Pooling）,但这些池化方式通常会丢失很多信息，例如采用平均池化会使得所有帧的贡献相同，忽略**情绪关键帧**对情绪识别的突出贡献。

1. Max Pooling can be expressed as $\mathbf{Y} \in \mathbb{R}^{B \times K}$ where:

$$\mathbf{Y}_b = \max_{t \in \{1,\ldots,T\}} \left(\mathbf{X}_{b,t} \cdot M_{b,t}\right) \tag{1}$$

2. Average Pooling can be expressed as $\mathbf{Y} \in \mathbb{R}^{B \times K}$ where:

$$\mathbf{Y}_b = \boldsymbol{\mu}_b = \frac{\sum_{t=1}^{T} \mathbf{X}_{b,t} \cdot M_{b,t}}{\sum_{t=1}^{T} M_{b,t}} \tag{2}$$

3. Statistics Pooling is $\mathbf{Y} \in \mathbb{R}^{B \times 2K}$ where $\mathbf{Y}_b = [\boldsymbol{\mu}_b, \boldsymbol{\sigma}_b]$ and $(\cdot)^{\odot 2}$ denotes the element-wise squaring:

$$\boldsymbol{\sigma}_b = \sqrt{\frac{\sum_{t=1}^{T} M_{b,t} \cdot \left(\mathbf{X}_{b,t} - \boldsymbol{\mu}_b\right)^{\odot 2}}{\sum_{t=1}^{T} M_{b,t}}} \tag{3}$$

-

由此，研究中开始引入**注意力池化（Attention-Based Pooling）**，并由随着SER技术的发展，从在低级arousic feature的分类器加attention层，到引入CNN/RNN的频谱图的分类器加attention层，再到预训练模型embedding出的高维特征的分类器加attention层；同时，注意力池化的方式也由最简单的ASP（Attention Statistic Pooling）到多头MQMHA（Multi-Query Multi-Head Attention）等，以提升 SER 的性能与可解释性。

需注意的是，Attention Pooling和Transformer架构里的attention不一样

- Transformer中的attention：自注意力，使encode出的向量包含上下文信息
- Attention Pooling：给多个frame-level的帧向量按照重要性加权，整合成utterance-level的向量

注意力池化机制作为一种汇聚方式，能够让模型学习**话语中哪些帧或片段最具情绪显著性。**以下简单回顾关于 SER 注意力池化的关键工作，同时覆盖早期奠基性研究和近几年的发展，并整理了与金融/经济语音相关的研究。

一、SER 中的早期注意力机制

- Huang & Narayanan（Interspeech 2016） https://www.isca-archive.org/interspeech_2016/huang16b_interspeech.html
  - 受编码中的attention机制启发，最早提出可以将注意力引入话语级的池化当中，采用**内容注意力**
    - 在BLSTM后接注意力层，对所有时间步的隐向量h_t赋予注意力权重α_t，再加权求和形成最终的句级表示

$$\begin{aligned}\mathbf{h}_t &= \sigma_h(\mathbf{W}^{\mathbf{hx}}\mathbf{x}_t + \mathbf{W}^{\mathbf{hh}}\mathbf{h}_{t-1}),\\ \mathbf{y}_t &= \sigma_y(\mathbf{W}^{\mathbf{yh}}\mathbf{h}_t),\end{aligned}$$

    - 准确率小幅度提升
  - 可解释性
    - 通过绘制attention权重曲线，发现其与frame能量曲线（特别是隐向量能量）部分相关、但又不完全一致
    - 大多数句子的attention与能量曲线的相关系数较低（平均约为0.13），表明模型能关注到**非显性、高能量帧以外的重要情绪点**
  - 贡献：强调了attention权重的主动选择作用，有别于被动聚合（如全平均），并首次观察到attention权重与能量曲线的**非线性关系**，体现出emotion相关内容的**非均匀时序分布。**
- Mirsamadi 等（ICASSP 2017） https://sigport.org/sites/default/files/docs/icassp2017_1.pdf
  - 首次系统提出了**局部注意力机制（local attention)**
    - 对每帧 BLSTM 输出 y_t 用一个可学向量 u 做内积打分并 softmax 得权重 a_t，句级向量 $z = \sum_t \alpha_t y_t$
    - 加权求和
    - 沿时间做**1D 局部注意力**
    - 优势：参数少，适合小样本数据
  - 架构
    - 输入为**原始谱向量**和**手工 LLD**，用 **BLSTM** 学时序表征与聚合
    - 模型整体结构为：LLD特征 → BLSTM → 注意力加权池化 → Softmax
    - 分类头与训练策略
  - 可解释性
    - 注意力自动给静音/无关帧极小权重
    - 不仅关注能量，更关注“情绪密度”——能将高能但无情感的帧赋予低权重
  - 贡献：相比 Huang & Narayanan (2016) 使用的“内容注意力”，此文从更工程角度出发，给出一种**可解释、参数少、易训练的 logistic regression attention 模型**，并对比了多

种聚合策略，是 attention pooling 在 SER 中的重要里程碑

- Neumann & Vu（Interspeech 2017) https://www.isca-archive.org/interspeech_2017/neumann17_interspeech.pdf
  - 将注意力机制引入 CNN，提出了 **“Attentive CNN（ACNN）”结构**，将卷积特征图与注意力池化融合，用于建模情绪信号在时间上的异质性
  - $$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))}$$
  - 架构
    - CNN + max pooling → attention pooling → max pooled vector ⊕ attention vector → softmax；
    - 端到端训练 CNN + attention + 分类器
- Zhang 等（2018/2019） https://arxiv.org/pdf/1806.01506
  - **2D 自注意力**：同时在时间和频率上“挑重点”。
    - 对**时频格子** $F \times T \times C$ 里每个单元 a_i 用 MLP 打分 $e_i = u \top tanh(W_a i + b)$
    - 经 softmax（带温度 λ）得a_i，加权求和得句向量 c
    - $$e_i = \boldsymbol{u}^T \tanh(\boldsymbol{W}\boldsymbol{a}_i + \boldsymbol{b})$$ $$\alpha_i = \frac{\exp(\lambda e_i)}{\sum_{k=1}^{L} \exp(\lambda e_k)}$$ $$\boldsymbol{c} = \sum_{i=1}^{L} \alpha_i \boldsymbol{a}_i$$
  - 架构
    - 使用 FCN（AlexNet/VGG）替代 CNN+LSTM
  - 可解释性
    - 可视化**2D 时频热图**，揭示低中频情感相关区域更显著
  - 贡献
    - 打破了以往只能做“时间轴注意力”（如 Mirsamadi 2017）的限制，首次提出 **“时频二维注意力”机制**
    - 探索了图像领域预训练模型（ImageNet）的迁移学习在 SER 中的可行性，是跨模态迁移在情感识别中的早期实践
- Tarantino 等（Interspeech 2019） https://www.isca-archive.org/interspeech_2019/tarantino19_interspeech.pdf
  - 首次将 Transformer 架构中的自注意力机制（self-attention）引入语音情感识别（SER）任务中

- 贡献
  - 开创性地探索了“全局窗口+自注意力”机制在 SER 中的表现，突破了传统 RNN+attention 框架（如 Mirsamadi 等 2017、Ramet 等 2018）的记忆限制
  - 首次**对比了 classification 与 regression 两种 SER 标签机制（硬标签 vs 软标签）**，强调标签分布一致性对模型性能的重要影响
- 与attention pooling 无关，因此不具体讨论

二、近期注意力池化（2020–2025）

- Leygue 等（Interspeech 2025） https://www.isca-archive.org/interspeech_2025/leygue25_interspeech.pdf
  - 首次将 MQMHA（Multi-Query Multi-Head Attentive Statistics Pooling）机制系统应用于 SER
    - **多查询（Multi-Query）**：多个独立的“查询向量”对应不同类型的情绪注意力通道
  - **多头（Multi-Head）**：每个查询在不同的表示子空间中并行执行注意力权重学习
  - **注意力统计聚合（Attentive Statistics）**：不仅获取 attention 权重加权平均（均值），还聚合 attention 权重下的方差信息，形成稳定的句级向量
  - 架构：Encoder → Pooling → Fusion → Classifier”
    - 音频 + 文本双模态：：将音频（W2V-BERT 2.0）和文本（DeBERTa v3）两端编码器输出**分别经 MQMHA 聚合后拼接**
    - 拼接结果输入一个MLP分类器
  - 可解释性
    - 帧级分布可视化：平均仅 15% 的帧占据了 80% 的 attention 权重，呈现**Pareto 分布**
    - 注意力更集中于 **非语言性标记（如呼吸声、笑声）** 以及 **主重音元音、双元音、语音加强部位**
- 近期特点
  - 单查询到多查询，单头到多头
  - 多模态：语音 + 文本(+视频)
  - 预训练模型做embedding
  - 时域 + 频域同时注意

三、金融应用

- Hajek & Munk（Neural Computing and Applications, 2023） https://link.springer.com/article/10.1007/s00521-023-08470-8
  - 文本情绪FinBERT+语音特征CNN+财务指标 输入 LSTM 用于财务预测
  - 局限
    - 所有语音记录只赋予一个总体情绪，未做**段落级情绪变换分析**

        - 分类边界硬（safe/grey/distress），未来可引入**模糊集表示**
- Yang 等（2025） https://cepr.org/voxeu/columns/what-corporate-earnings-calls-reveal-about-ai-stock-rally
    - **研究联储新闻发布会发言人语气与股价的关系，但属于计量经济学，没有引入SER 也没明确说明情绪标签是怎么来的**