数据集

1. SE-cap https://arxiv.org/abs/2312.10381

数据集：内部的EMOSpeech中文数据集，https://github.com/thuhcsi/SECap

仓库里有600条 wav 和对应的描述文本 ✓

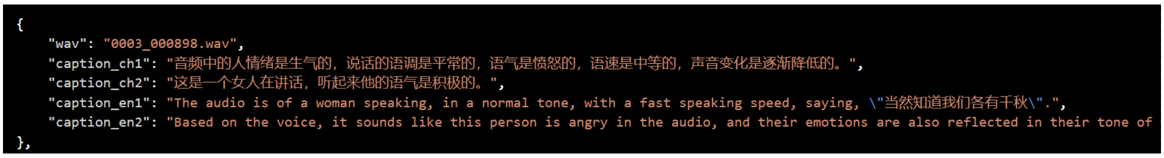
2. AlignCap: Aligning Speech Emotion Captioning to Human Preferences

https://aclanthology.org/2024.emnlp-main.224.pdf

数据集：

a. MER2023 (A large-scale video emotion reason dataset) → MER23SEC
b. NNIME (A Chinese interactive multimodal emotion corpus)：用来评估模型在其它数据集上的可迁移性的。
c. EMOSEC (他们自己提出的41 hours of Chinese-English Speech Emotion Captioning datasets) https://zenodo.org/records/10948423 ✓

(在zenodo上只放了情绪的描述文本merged_file.json, 但它是基于 ESD 和 IEMOCAP构建的，ESD和IEMPCAP都能单独找到)



```
{
    "wav": "0003_000898.wav",
    "caption_ch1": "音频中的人情绪是生气的，说话的语调是平常的，语气是愤怒的，语速是中等的，声音变化是逐渐降低的。",
    "caption_ch2": "这是一个女人在讲话，听起来他的语气是积极的。",
    "caption_en1": "The audio is of a woman speaking, in a normal tone, with a fast speaking speed, saying, \"当然知道我们各有千秋\".",
    "caption_en2": "Based on the voice, it sounds like this person is angry in the audio, and their emotions are also reflected in their tone of
},
```

d. ESD: **Publicly Available Emotional Speech Dataset (ESD) for Speech Synthesis and Voice Conversion** ✓

https://github.com/HLTSingapore/Emotional-Speech-Data

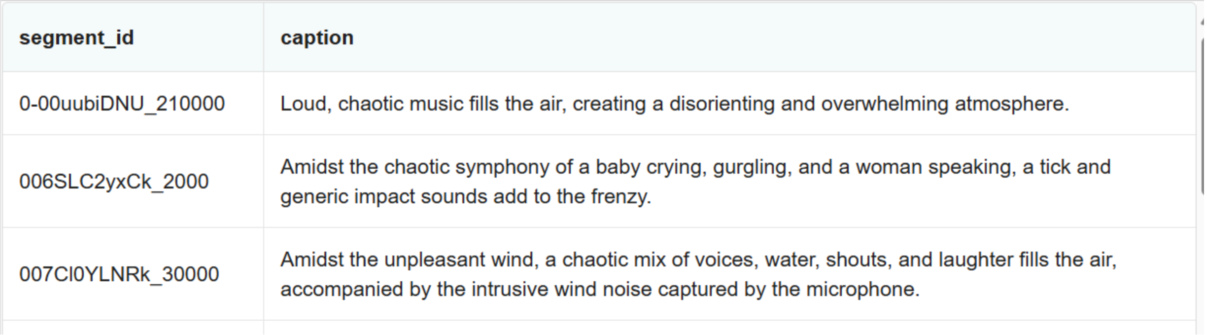
e. IEMOCAP: https://sail.usc.edu/iemocap/index.html

IEMOCAP还在申请中……

3. EmotionCaps: Enhancing Audio Captioning Through Emotion-Augmented Data Generation

https://zenodo.org/records/13755932 （情感增强的音频字幕）

但这个数据集面向环境声为主，并非专门的“语音”集合，但是可以作为让 caption 更贴合情绪语气的过渡数据 ✓

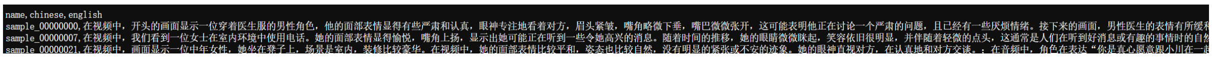


segment_id	caption
0-00uubiDNU_210000	Loud, chaotic music fills the air, creating a disorienting and overwhelming atmosphere.
006SLC2yxCk_2000	Amidst the chaotic symphony of a baby crying, gurgling, and a woman speaking, a tick and generic impact sounds add to the frenzy.
007Cl0YLNRk_30000	Amidst the unpleasant wind, a chaotic mix of voices, water, shouts, and laughter fills the air, accompanied by the intrusive wind noise captured by the microphone.

4. MER2024

https://huggingface.co/datasets/MERChallenge/MER2024

5,030 条带标注的视频切片（含音频），都有对应字幕和离散情感标签。但只有其中 332 条有与情感相关的描述/理由 ✓



```
name,chinese,english
sample_00000000,在视频中，开头的画面显示一位穿着医生服的男性角色，他的面部表情显得有些严肃和认真，眼神专注地看着对方，眉头紧皱，嘴角略微下垂，嘴巴微微张开，这可能表明他正在讨论一个严肃的问题，且已经有一些厌烦情绪。接下来的画面，男性医生的表情有所缓
sample_00000007,在视频中，我们看到一位女士在室内环境中使用电话。她的面部表情显得愉悦，嘴角上扬，显示出她可能正在听到一些令她高兴的消息。随着时间的推移，她的眼睛微微眯起，笑容依旧很明显，并伴随着轻微的点头，这通常是人们在听到好消息或有趣的事情时的自
sample_00000021,在视频中，画面显示一位中年女性，她坐在凳子上，场景是室内，装修比较豪华。在视频中，她的面部表情比较平和，姿态也比较自然，没有明显的紧张或不安的迹象。她的眼神直视对方，在认真地和对方交谈。；在音频中，角色在表达“你是真心愿意跟小川在一
```