EMOTION-AWARE SUBTITLES: IMPROVING SPEECH EMOTION RECOGNITION ON IEMOCAP WITH WAVLM AND ATTENTIVE STATISTICS POOLING

Yu Liu & Haohan Ying
245-LY
University of Rochester
Rochester, NY 14627, USA
yliu307@u.rochester.edu, hying4@u.rochester.edu

ABSTRACT

Subtitles convey *what* is spoken but not *how*, leaving hearing-impaired viewers without prosodic cues such as tone and emotion. We study speech emotion recognition (SER) as a step toward emotion-aware subtitles that attach compact emotion tags to utterances. Starting from a Wav2Vec2 mean-pooling baseline, we build a stronger end-to-end model by fine-tuning WavLM with an attentive statistics pooling head, class-weighted cross-entropy, and label smoothing. To reduce ambiguity between highly confusable high-arousal negative emotions, we merge *angry* and *frustrated* into a single *high_neg* category and evaluate a 5-way classifier on IEMOCAP (*high_neg, excited, happy, neutral, sad*). Our final model achieves about 0.66 validation accuracy and around 0.57 macro-F1, substantially improving over our initial baseline. Per-class analysis and the confusion matrix show strong performance on *high_neg* and *excited*, while *happy* remains challenging due to limited data and overlap with *excited* and *neutral*.

1 INTRODUCTION

Standard subtitles tell viewers *what* words are spoken but omit crucial information about *how* those words are delivered. Prosody—intonation, loudness, rhythm, and speaking style—often reveals the speaker's emotional state. For hearing-impaired viewers, this information is largely invisible: the line "Stop it." could be playful teasing or an angry command, yet subtitles typically render both as plain text. This missing emotional layer can reduce narrative understanding and make it harder to follow social dynamics in movies, TV shows, or online videos.

Our long-term vision is *emotion-aware subtitles* that automatically add compact tags indicating the speaker's emotion, e.g.,

```
[High_neg] I told you to stop!
[Happy] That's wonderful news!
```

Realizing this vision requires robust speech emotion recognition (SER) from raw audio. SER is difficult because emotions are expressed through subtle acoustic cues, labels are noisy and imbalanced, and natural conversation contains background noise, interruptions, and speaker variability.

In this project, we focus on utterance-level SER on IEMOCAP. We begin with a simple reproduction-style baseline (Wav2Vec2 + mean pooling), then improve it by (i) switching to WavLM as the backbone, (ii) replacing mean pooling with attentive statistics pooling, and (iii) adding training stabilizers such as label smoothing. We also adopt a pragmatic relabeling that merges *angry* and *frustrated* into a single *high_neg* class to reduce systematic confusion.

2 RELATED WORK

Self-supervised speech representations. Wav2Vec2 (Baevski et al., 2020) popularized self-supervised pretraining for speech by learning contextualized acoustic representations from large unlabeled corpora. WavLM (Chen et al., 2022) extends this direction with objectives that improve robustness for general speech processing beyond ASR, making it attractive for paralinguistic tasks such as SER.

Fine-tuning speech SSL models for SER. End-to-end fine-tuning of Wav2Vec2-style models has been shown to outperform hand-crafted features on IEMOCAP and related SER benchmarks (Chen & Rudnicky, 2021). These results motivate using self-supervised backbones as a strong foundation for emotion recognition.

Utterance-level pooling. Since SSL encoders produce frame-level features, SER requires pooling across time. Attention-based pooling mechanisms can improve utterance-level representations by emphasizing emotionally salient frames, outperforming simple mean/max pooling in multiple settings (Li et al., 2018).

Generalization and domain shift. Cross-domain SER remains challenging, especially when transferring between spontaneous conversational speech and acted studio recordings. Prior work explores multilingual and cross-domain settings (Sharma, 2021) as well as multi-stage fine-tuning strategies (Gao et al., 2023) to improve robustness.

Our final system follows these insights but remains intentionally simple: we use a single-stage fine-tuning pipeline and focus on a small set of targeted improvements (WavLM backbone, attentive statistics pooling, and training regularization) suitable for a course project.

3 METHODS

3.1 TASK AND LABEL SPACE

Given a raw speech waveform $\mathbf{x} \in \mathbb{R}^T$, we predict an utterance-level emotion label. Our final model uses a 5-way label space:

$$y \in \{\texttt{high\_neg}, \texttt{excited}, \texttt{happy}, \texttt{neutral}, \texttt{sad}\},$$

where `high_neg` merges *angry* and *frustrated*. We treat SER as single-label multi-class classification and train by minimizing cross-entropy.

3.2 AUDIO PREPROCESSING AND AUGMENTATION

We resample all audio to 16 kHz and cap utterances to $D = 6$ seconds. Long utterances are cropped to a contiguous 6-second segment; during training we use random crops, while for validation we use deterministic cropping for reproducibility. We apply light additive Gaussian noise during training to improve robustness.

3.3 BASELINE: WAV2VEC2 WITH MEAN POOLING

Our initial attempt is a standard reproduction baseline: `facebook/wav2vec2-base` as the encoder, mean pooling over time, and a linear classification head. We train a 6-class classifier (*angry, excited, frustrated, happy, neutral, sad*) with class-weighted cross-entropy to partially mitigate imbalance. This baseline provides a reference point for our later improvements.

3.4 FINAL MODEL: WAVLM WITH ATTENTIVE STATISTICS POOLING

Our final model replaces the backbone and pooling head:

Backbone. We use `microsoft/wavlm-base` (Chen et al., 2022) as the encoder and fine-tune all layers end-to-end.

Emotion (final)	Count	Percentage
high_neg (angry+frustrated)	3767	42.7%
excited	1778	20.2%
neutral	1553	17.6%
sad	1125	12.8%
happy	591	6.7%
Total	8814	100%

Table 1: Class distribution after merging *angry* and *frustrated* into `high_neg`.

Attentive statistics pooling. Let $\mathbf{H} = (h_1, \ldots, h_L)$ be the frame-level hidden states ($h_t \in \mathbb{R}^d$). We compute attention weights α_t over frames and form a weighted mean and (diagonal) standard deviation:

$$\mu = \sum_{t=1}^{L} \alpha_t h_t, \quad \sigma = \sqrt{\sum_{t=1}^{L} \alpha_t (h_t - \mu)^2}.$$

We concatenate $[\mu; \sigma]$ as an utterance embedding and apply a linear classifier. This pooling better captures emotionally salient frames and variability (e.g., arousal), compared to simple averaging.

Loss and regularization. We use class-weighted cross-entropy and apply label smoothing (0.1) to reduce overconfidence under noisy/ambiguous labels.

3.5 Optimization Details

For the final model, we use AdamW with learning rate 3×10^{-5}, batch size 8, and train for 20 epochs with a linear warmup schedule. We clip gradient norm at 1.0. (Our earlier baseline used a smaller batch size and fewer epochs due to GPU memory constraints.)

4 Experiments

4.1 Dataset and Relabeling

We use the IEMOCAP speech modality and keep utterances labeled as *angry, excited, frustrated, happy, neutral, sad*. For our final setting, we merge *angry* and *frustrated* into `high_neg`. Table 1 shows the resulting 5-class distribution (computed from the original counts).

4.2 Train/Validation Split

We perform a stratified split with 10% of utterances held out for validation, preserving class proportions.

4.3 Metrics

We report validation accuracy and macro-F1, as well as per-class F1 and the confusion matrix. Macro-F1 is especially important because minority classes (notably `happy`) should contribute equally to evaluation.

4.4 Training Dynamics of the Final Model

Figure 1 shows that training loss decreases smoothly over 20 epochs. Validation accuracy (Figure 2) improves rapidly in early epochs and stabilizes around the mid-0.6 range, while macro-F1 stabilizes in the high-0.5 range. We select the best checkpoint by validation accuracy (epoch 16 in our run).

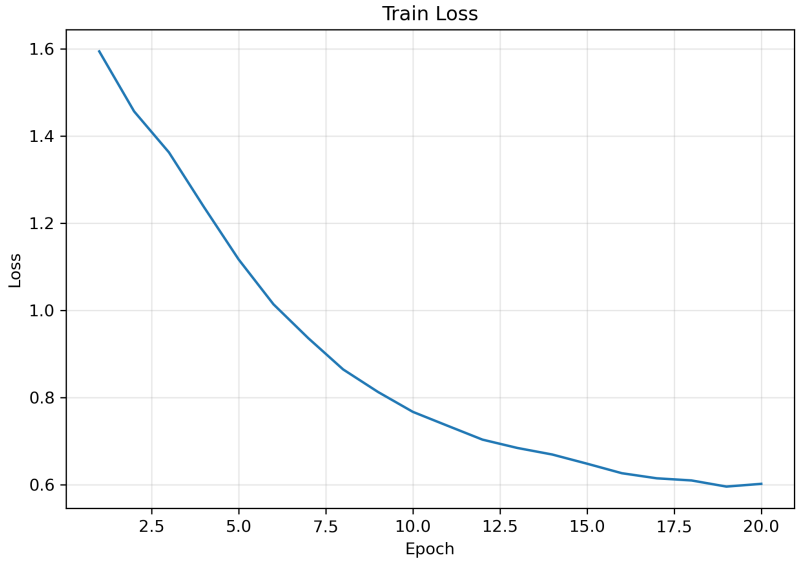

Figure 1: Training loss over epochs (final model).

Component	Baseline (initial)	Final (this work)
Backbone	Wav2Vec2-base	WavLM-base
Label space	6-way (angry, excited, frustrated, happy, neutral, sad)	5-way (high_neg, excited, happy, neutral, sad)
Pooling	Mean pooling	Attentive statistics pooling (μ and σ)
Loss	Class-weighted CE	Class-weighted CE + label smoothing (0.1)
Training	Smaller batch, fewer epochs	Batch 8, 20 epochs, best checkpoint by val_acc

Table 2: Summary of our progression from baseline to final model. Note that absolute metrics are not directly comparable across different label spaces; the final setting reduces systematic confusions by merging *angry* and *frustrated*.

4.5 Per-Class Results and Confusions

Figure 3 shows per-class F1 over training. The model performs strongly on `high_neg` and `excited`, and moderately on `neutral` and `sad`. `happy` remains the hardest class: it is both low-frequency and acoustically close to `excited` (high arousal) and sometimes `neutral` (low arousal).

The confusion matrix (Figure 4) reveals structured errors. Many `neutral` utterances are predicted as `high_neg`, reflecting that natural dialogue often contains subtle irritation or tension that lies near the decision boundary. `happy` is frequently confused with `excited`, consistent with overlapping acoustic cues.

4.6 From Baseline to Final Model: What Improved?

Our project progressed from a straightforward reproduction baseline to a stronger SER pipeline. Table 2 summarizes the main differences.

Qualitatively, these changes produce cleaner separations for high-arousal emotions and improve overall validation performance. The largest gain comes from using a stronger backbone (WavLM) together with a pooling mechanism that can focus on emotionally salient frames and capture temporal

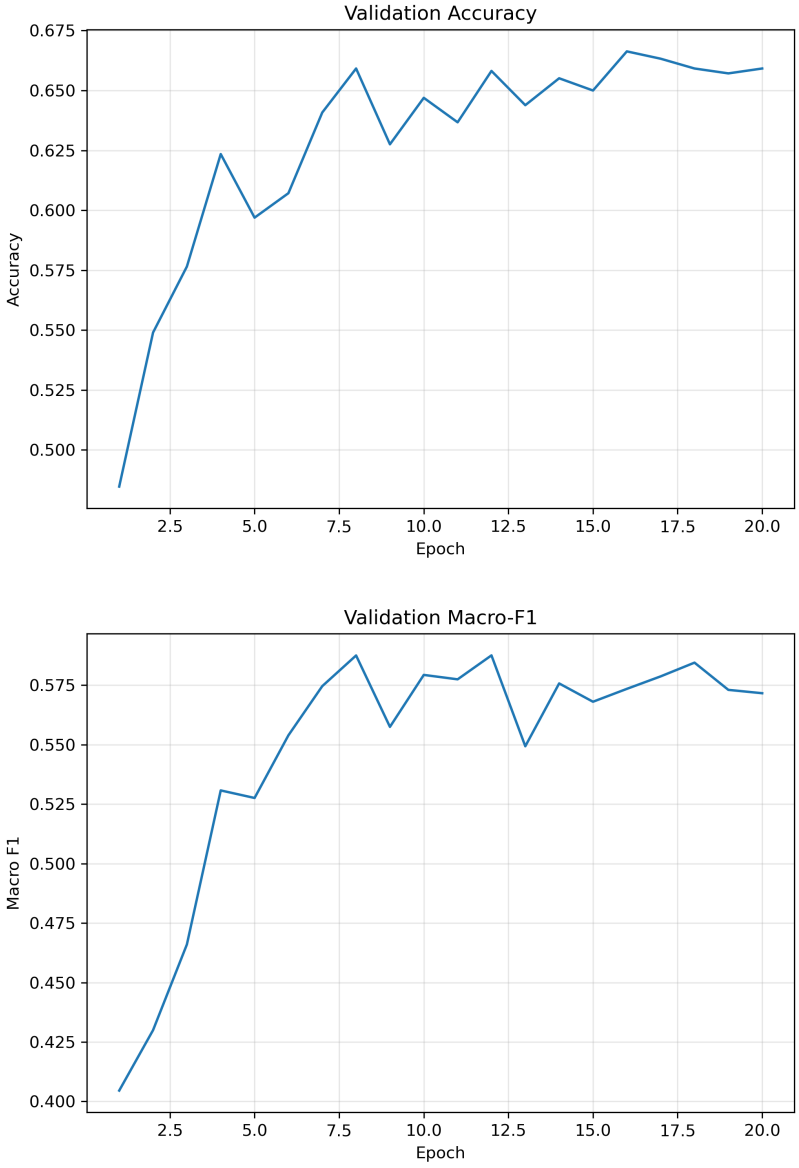

Figure 2: Top: validation accuracy over epochs. Bottom: validation macro-F1 over epochs (final model).

variability. The relabeling to `high_neg` further reduces unavoidable confusion between *angry* and *frustrated*, which are often ambiguous even for humans.

4.7 Discussion

Our final system demonstrates that end-to-end fine-tuning of a self-supervised speech encoder can support emotion-aware subtitles. Compared to our initial Wav2Vec2 mean-pooling baseline, the WavLM + attentive statistics pooling model yields higher validation accuracy and macro-F1, and produces more stable per-class behavior.

Two limitations remain. First, **data imbalance and ambiguity** still affect `happy`, the rarest class; its acoustic realization ranges from low-arousal warmth (close to `neutral`) to high-arousal joy (close to `excited`). Second, **domain shift** is unresolved: IEMOCAP contains spontaneous, noisy

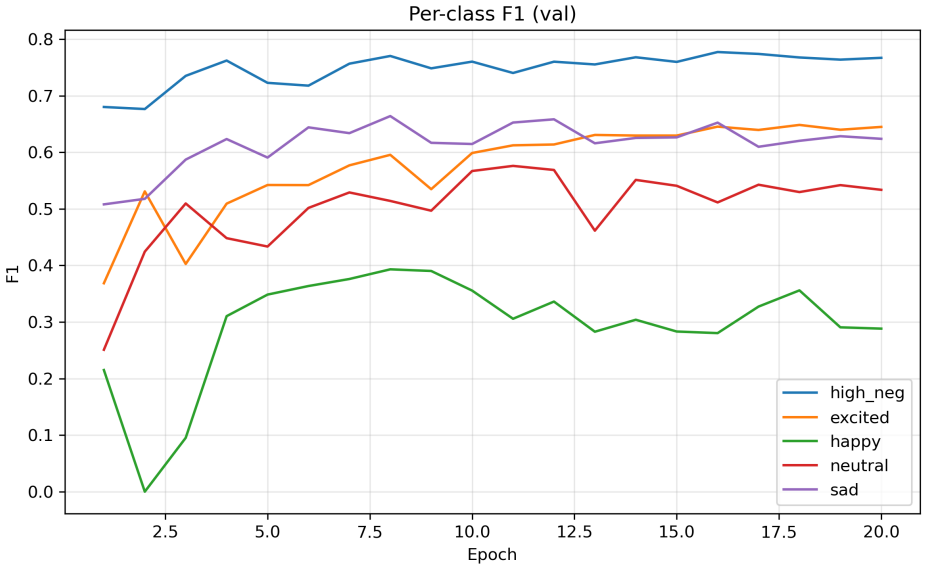

Figure 3: Per-class F1 on the validation set across epochs (final model).

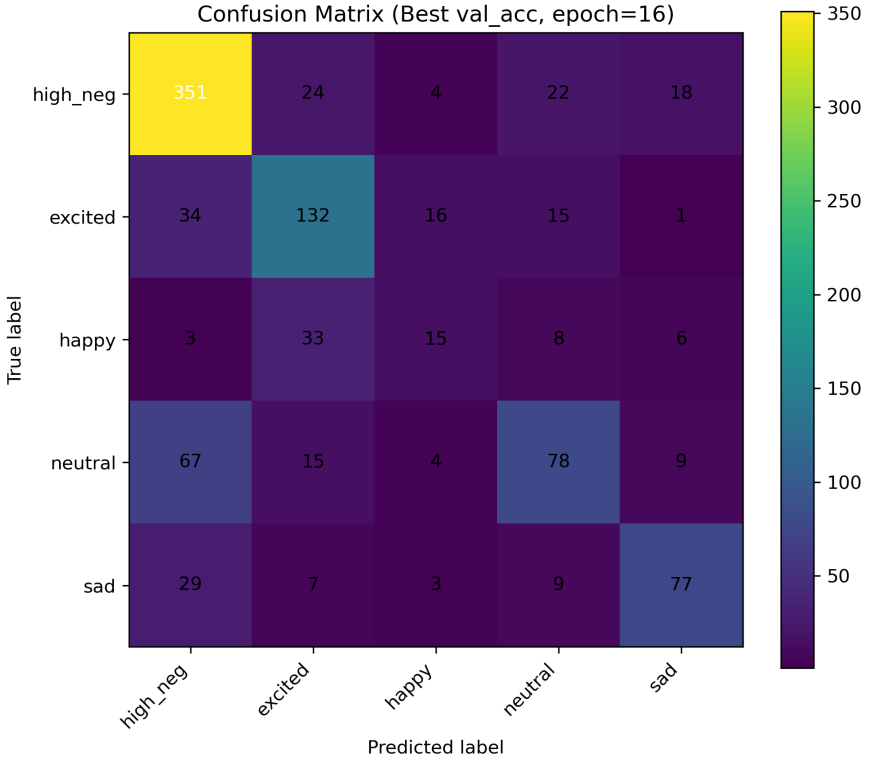

Figure 4: Confusion matrix of the best validation-accuracy checkpoint (final model).

conversational speech, while acted datasets such as TESS are clean and exaggerated. Bridging this gap will likely require multi-domain training, domain-adversarial methods, or additional fine-tuning stages.

For subtitle applications, a practical extension is to output confidence-aware tags (or omit tags when uncertain), rather than forcing a single hard label for every utterance. This may better reflect the inherently graded nature of emotion in real dialogue.

5 CONCLUSION

We studied SER as a foundation for emotion-aware subtitles that convey not only *what* is said but also *how* it is said. Starting from a Wav2Vec2 mean-pooling baseline, we improved performance by adopting WavLM, attentive statistics pooling, and label smoothing, and by merging highly confusable negative high-arousal emotions into a `high_neg` category. Our final model reaches about 0.66 validation accuracy and around 0.57 macro-F1 on a 5-way IEMOCAP setting, with strong recognition of `high_neg` and `excited` but persistent difficulty on `happy`. Future work includes addressing domain shift across datasets and integrating SER with real-time subtitle pipelines.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Szu-Jui Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition. In *Proc. Interspeech*, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, and others. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing (J-STSP)*, 2022.

Yunxia Li, Li Deng, and Dong Yu. Attentive pooling networks for speech emotion recognition. In *Proc. Interspeech*, 2018.

Ankita Sharma. Multilingual speech emotion recognition using wav2vec 2.0. *arXiv preprint*, 2021.

Ming Gao, Wei Wang, and Zhiyong Li. Two-stage fine-tuning of wav2vec 2.0 for robust speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.