# Speech Emotion Recognition with Distilled Prosodic and Linguistic Affect Representations

Scan for more!

**Debaditya Shome**, Ali Etemad

Queen's University, Canada

## Introduction

**Motivation:**

- ✗ Implicitly learning prosodic information from audio is often less than optimal due to the discretization of audio signals during training of leading speech models (e.g., HuBERT)
- ✗ Direct fine-tuning of existing speech models originally trained for ASR doesn't perform well on SER.
- ✗ Direct use of transcripts at *run-time* can lead to low performances due to transcription errors.
- ✗ Using both audio and linguistic information at run-time requires a multimodal system which can increase computational overhead.

**Our contributions:**

- ✓ We introduce EmoDistill, a novel cross-modal Knowledge Distillation (KD) framework for learning unimodal representations from speech that explicitly capture both the linguistic and prosodic aspects of emotions.
- ✓ EmoDistill outperforms previous state-of-the-art methods on IEMOCAP and achieves $77.49$% UA and $78.91$% WA.

## Experiment Details

**Dataset:**

- IEMOCAP benchmark
- 4 emotions (neutral, angry, sad, happy)
- 10-Fold cross-validation
- Subject-independent

**Implementation:**

- Prosodic Teacher: 4-layer ResNet2D trained on eGeMAPs LLDs.
- Linguistic Teacher: BERT-base (pre-trained)
- AdamW, base LR of 1e-4
- 4 $\times$ NVIDIA A100 GPUs, Batch size = 128

## Method

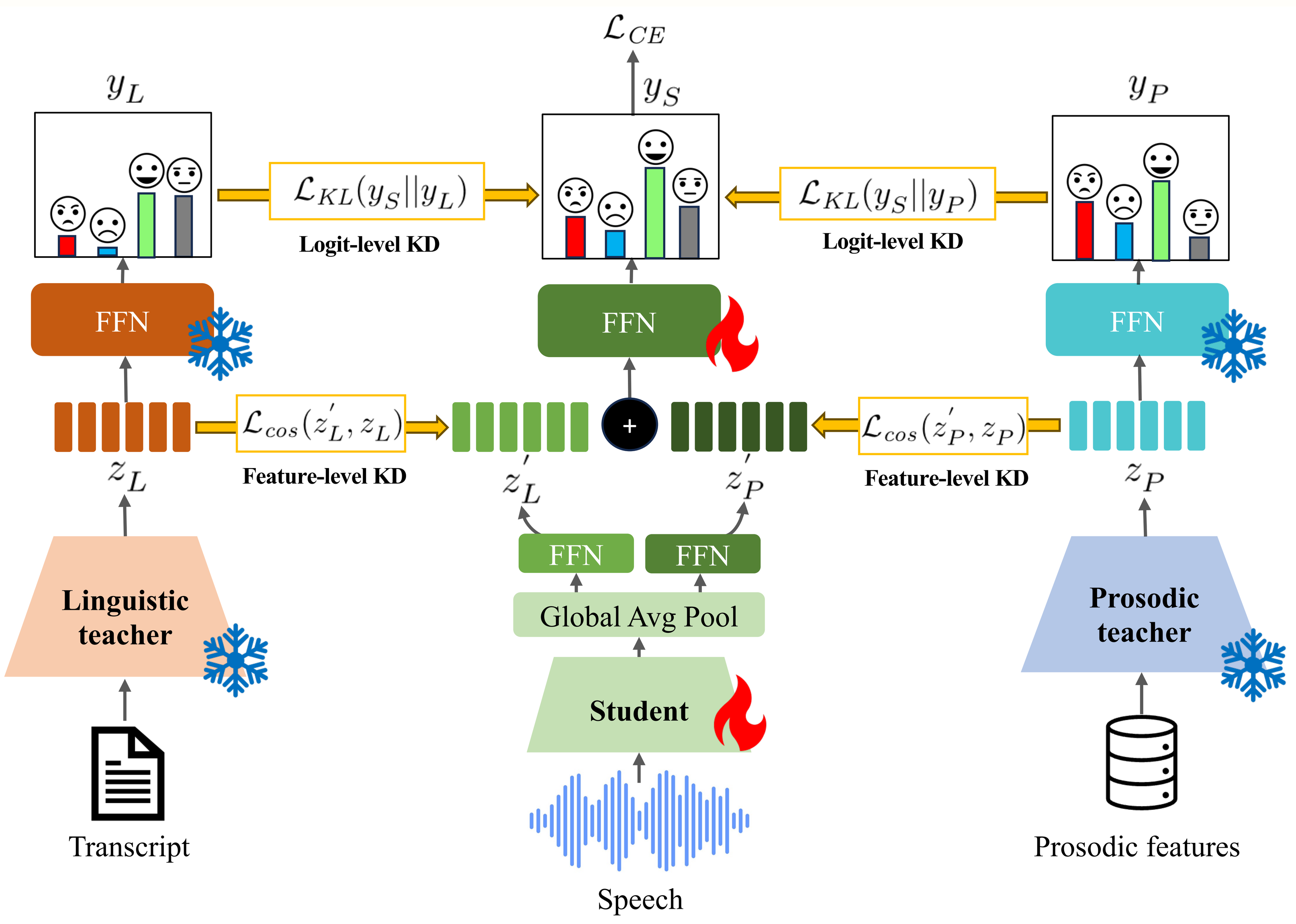


Fig. 1: EmoDistill Framework. Our student network is trained using a distillation of logit-level and embedding-level knowledge from frozen linguistic and prosodic teacher networks, along with standard cross-entropy loss. **During inference, we only use the student network in an unimodal setup, avoiding computational overhead as well as transcription and prosodic feature extraction errors.**

**Logit-level KD.** First, we transfer the logit-level knowledge using traditional KD with temperature-scaled labels [1]. Specifically, we minimize the KL-Divergence $L_{KL}$ between the predicted logit distributions of teacher and student models, where the objective becomes:

$$\mathcal{L}_{logits} = \mathcal{L}_{KL}(y_S||y_L) + \mathcal{L}_{KL}(y_S||y_P). \tag{1}$$

Here, $y_S$ refers to the predictions of the student, while $y_L$ and $y_P$ represent the predictions of Linguistic and Prosodic teacher models, respectively. In all cases, the predicted logits $y$ are obtained using temperature parameter $\tau$ in the output softmax activation function. In practice, we use different values of $\tau$ for KD from $f_T^L$ and $f_T^P$. Let $z_c$ be the output logits for class $c$, among a total of $N$ classes. The temperature-scaled logits $y_c$ are obtained as:

$$y_c = \frac{e^{z_c/\tau}}{\sum_{k=1}^{N} e^{z_c/\tau}}. \tag{2}$$

**Feature-level KD.** Next, we use embedding-level KD to transfer knowledge to the student model from the latent space of Linguistic and Prosodic teacher models. Let $z_L$ and $z_P$ denote the embeddings of Linguistic and Prosodic teachers, while $z_L'$ and $z_P'$ denote the embeddings of the student model from linguistic and prosodic projection layers respectively. We minimize the negative cosine similarity $L_{cos}$ among the teacher and student embeddings as follows:

$$\mathcal{L}_{embeddings} = \mathcal{L}_{cos}(z_L', z_L) + \mathcal{L}_{cos}(z_P', z_P). \tag{3}$$

Given two embeddings $a$ and $b$, $L_{cos}$ can be defined as:

$$\mathcal{L}_{cos}(a, b) = \frac{a}{\|a\|_2} \cdot \frac{b}{\|b\|_2}, \tag{4}$$

where $\|\cdot\|_2$ represents $\ell_2$-norm.

**Loss objective.** $\mathcal{L}_{EmoDistill} = \alpha\mathcal{L}_{logits} + \beta\mathcal{L}_{embeddings} + \gamma\mathcal{L}_{CE}$ (5)

## Performance Evaluation

Tab. 1: SER results on IEMOCAP. **Bold** denotes the best results while <u>underline</u> denotes the second-best.

| Method | Inf. Backbone | Modality | WA | UA |
|---|---|---|---|---|
| Sun *et al.* (2021) | CNN+LSTM | Multimodal | 61.2 | 56.01 |
| Heusser *et al.* (2019) | BiLSTM+XLNet | Multimodal | 71.40 | 68.60 |
| Triantafyllopoulos *et al.* (2023) | MFCNN+BERT | Multimodal | - | 72.60 |
| Ho *et al.* (2020) | RNN+BERT | Multimodal | 73.23 | 74.33 |
| Aftab *et al.* (2022) | FCNN | Unimodal | 70.23 | 70.76 |
| Liu *et al.* (2020) | TFCNN+DenseCap+ELM | Unimodal | 70.34 | 70.78 |
| Cao *et al.* (2021) | LSTM+Attention | Unimodal | 70.50 | 72.50 |
| Lu *et al.* (2020) | RNN-T | Unimodal | 71.72 | 72.56 |
| Wu *et al.* (2021) | CNN-GRU+SeqCap | Unimodal | 72.73 | 59.71 |
| Zou *et al.* (2022) | Wav2Vec2+CNN+LSTM | Unimodal | 71.64 | 72.70 |
| Ye *et al.* (2023) | TIM-Net | Unimodal | 72.50 | 71.65 |
| Ours | HuBERT-base | Unimodal | <u>75.16</u> | <u>76.12</u> |
| Ours | HuBERT-large | Unimodal | **77.49** | **78.91** |

## Ablation study

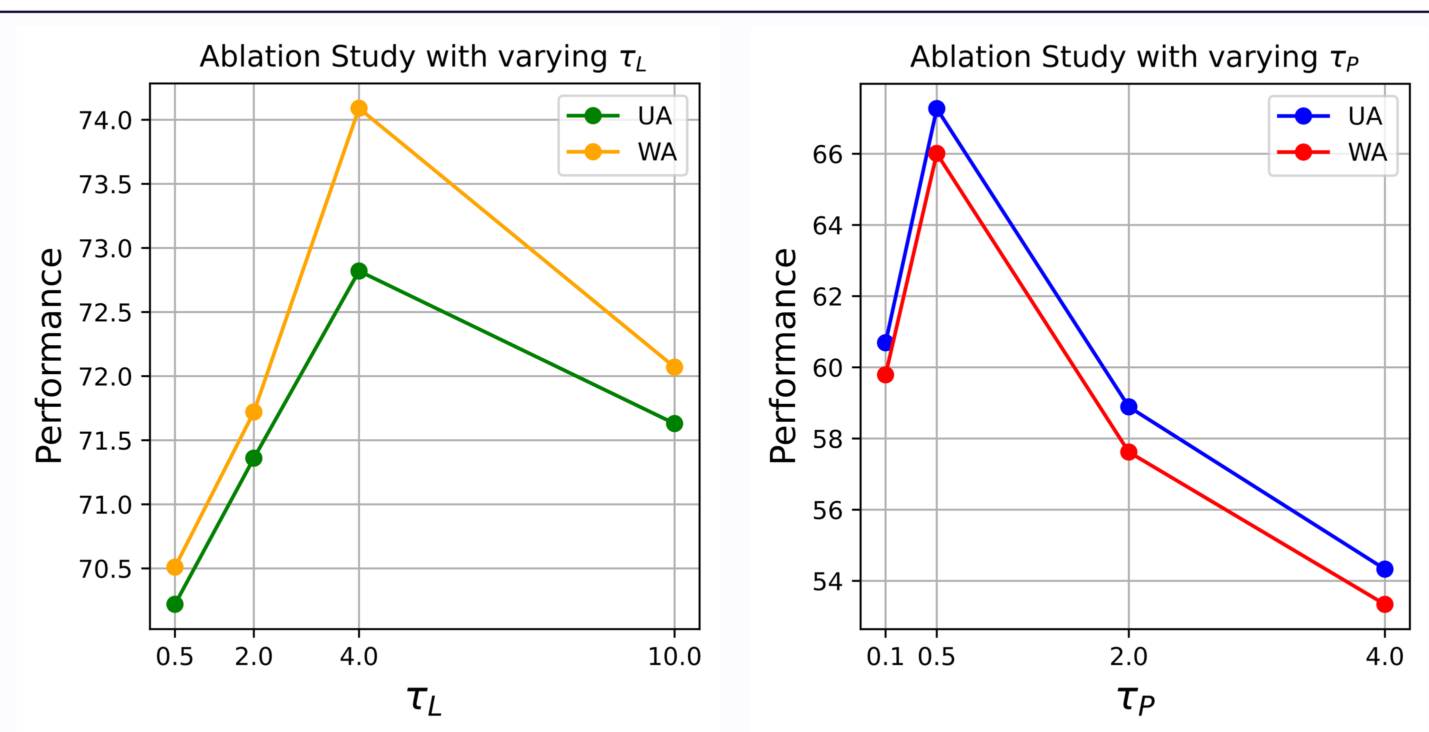


Fig. 2: **Left:** We remove $f_T^L$ and vary $\tau_P$. **Right:** We remove $f_T^P$ and vary $\tau_L$.

| Variants | WA | UA |
|---|---|---|
| Ours | **75.16** | **76.12** |
| w/o $\mathcal{L}_{logits}$ | 73.94 ($\downarrow 1.22$) | 74.02 ($\downarrow 2.10$) |
| w/o $\mathcal{L}_{embedding}$ | 73.88 ($\downarrow 1.28$) | 74.01 ($\downarrow 2.11$) |
| w/o $f_T^P$ | 74.09 ($\downarrow 1.07$) | 72.82 ($\downarrow 3.30$) |
| w/o $f_T^L$ | 66.01 ($\downarrow 9.15$) | 67.27 ($\downarrow 8.85$) |
| w/o $f_T^P$ and $f_T^L$ | 69.92 ($\downarrow 5.24$) | 70.17 ($\downarrow 5.95$) |
| w/o $f_S$ and $f_T^L$ | 49.42 ($\downarrow 25.74$) | 50.08 ($\downarrow 26.04$) |
| w/o $f_S$ and $f_T^P$ | 71.09 ($\downarrow 4.07$) | 71.83 ($\downarrow 4.29$) |

- ✓ Linguistic understanding is crucial for SER.
- ✓ Prosodic understanding is complimentary, but leads to a boost in SER performance.
- ✓ Hard logits are better for KD from the prosodic teacher, as it is a weak teacher.
- ✓ Soft logits are better for KD from the linguistic teacher, as it is a strong teacher.
- ✓ Using Logit and Embedding-level KD together improves performance.

## Acknowledgement

This work was supported by Mitacs, Vector Institute, and Ingenuity Labs Research Institute.