



# Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition

*Mehedi Hasan Bijoy, Dejan Porjazovski, Tamás Grósz, Mikko Kurimo*

Department of Information and Communications Engineering, Aalto University, Finland

firstname.lastname@aalto.fi

## Abstract

Speech Emotion Recognition (SER) is crucial for improving human-computer interaction. Despite strides in monolingual SER, extending them to build a multilingual system remains challenging. Our goal is to train a single model capable of multilingual SER by distilling knowledge from multiple teacher models. To address this, we introduce a novel language-aware multi-teacher knowledge distillation method to advance SER in English, Finnish, and French. It leverages Wav2Vec2.0 as the foundation of monolingual teacher models and then distills their knowledge into a single multilingual student model. The student model demonstrates state-of-the-art performance, with a weighted recall of 72.9 on the English dataset and an unweighted recall of 63.4 on the Finnish dataset, surpassing fine-tuning and knowledge distillation baselines. Our method excels in improving recall for sad and neutral emotions, although it still faces challenges in recognizing anger and happiness.

**Index Terms:** multi-teacher knowledge distillation, multilingual, speech emotion recognition, large audio-language model

## 1. Introduction

Speech Emotion Recognition (SER) involves identifying and analyzing emotional states in spoken languages. In multilingual settings, it becomes challenging due to the distinct linguistic features of each language. Unlike simple fine-tuning (FT) and knowledge distillation (KD), multi-teacher knowledge distillation (MTKD) specifically addresses this challenge by using multiple teacher models, each specialized in a different language, to enhance the student model's performance. The importance of SER is recognized as it benefits applications in human-computer interaction and speech processing tasks by fostering empathetic and effective interactions. For instance, SER aids in early detection and intervention in mental health issues by monitoring speech [1], even in multilingual communities.

Recent efforts in SER have primarily focused on leveraging pre-trained transformer-based large audio-language models due to their robust speech representation capabilities [2]. Moreover, it is hypothesized that integrating knowledge from multiple teacher models significantly enhances student model performance by utilizing complementary insights. For example, Confidence-Aware MTKD dynamically adapts to each teacher's reliability [3], while Adaptive MTKD employs a meta-weight network to coordinate diverse knowledge from multiple teachers [4]. Their empirical results indicate that these methods outperform FT and KD models, particularly in low-resource settings [5]. Overall, there is a clear shift towards adaptive meta-learning strategies, pre-trained models, and MTKD to achieve high performance and efficiency.

We found that previous studies have not investigated

MTKD for multilingual SER, particularly in training a multilingual student model from monolingual teachers. Therefore, we hypothesize that optimizing MTKD for multilingual SER can enhance emotional knowledge transfer across languages. By leveraging cross-lingual data, we aim to better integrate emotional knowledge in the student model. Our objective is to construct a single multilingual SER model through the distillation of knowledge obtained from multiple monolingual teacher models. We investigate whether MTKD improves cross-linguistic emotional knowledge transfer in multilingual SER. To achieve this, we use cosine similarity scores to select the most suitable teacher for each language. Consequently, this targeted approach should improve cross-linguistic knowledge integration, allowing better generalization across languages. By mitigating score variation through cosine similarity rescaling, our method enhances stability, resulting in improved cross-linguistic generalization and reliable performance, even in challenging multilingual scenarios.

The main contributions of this study are summarized below:

- We propose a language-aware MTKD method using three teacher models to enhance SER in English, Finnish, and French. It includes both monolingual and multilingual configurations, setting a novel benchmark for multilingual SER.
- Our proposed method achieves superior performance across datasets by surpassing different training paradigms, such as standard FT and conventional KD baselines, demonstrating generalizability and robustness.

## 2. Related Works

Numerous approaches have been developed for the SER task, which can broadly be categorized into four groups: simple FT, conventional KD, MTKD, and multimodal fusion. As for simple FT, pre-trained models like wav2vec2.0 [6], HuBERT [7], and WavLM [8] have significantly advanced SER by addressing the issue of data scarcity [9]. Leveraging transfer learning and FT, as demonstrated in the works of [10] and [11], markedly improved the performance of emotion recognition from speech. Sharma et al. [12] expanded on this by developing a multilingual learning system, while another study introduced the P-TAPT method to better align pre-training with target tasks [11]. Unlike the previous studies, Chakhtouna et al. [13] proposed a model combining HuBERTX-large with an SVM to emphasize the importance of advanced feature extraction. However, while the effectiveness of fine-tuning is supported by the empirical outcomes of [10] and [11], limitations such as potential overfitting and model convergence issues are observed in [13].

When it comes to conventional KD, studies have expanded its scope for SER by transferring knowledge from a larger teacher model to a smaller student model to optimize perfor-

mance [14]. For example, Hao et al. [15] introduced the OFA-KD framework for heterogeneous architectures, and Zhao et al. [16] proposed a hierarchical network with decoupled KD. Similarly, another work focused on KD-based model adaptation for emotional speech [17]. These studies generally aim to enhance the model performance. However, challenges such as the effectiveness of dropout [18], resource demands [19], and the need for improvements in multi-modal KD [16] are still persistent.

With respect to MTKD, several methods have been proposed, including a confidence-aware MTKD framework to address low-performing teachers [3], switched-training for different resource settings [20], and adaptive MTKD with meta-learning [4]. In contrast, Yang et al. [21] and Ren et al. [22] concentrated on efficiency and resource optimization through two-stage MTKD and self-distillation, respectively. Additionally, a few studies explored multimodal and ensemble methods for enhancing performance [5, 23]. Despite these advancements, challenges such as high computational demands [4, 5], large dataset requirements [21], and increased complexity [20, 22] persist, highlighting ongoing limitations in the field.

### 3. Methodology

Our proposed method involves training a student model  $S$  guided by three teacher models  $T_1$ ,  $T_2$ , and  $T_3$ . The training process uses the raw audio waveform input  $\mathbf{X}$ , which is a discrete-time sequence of  $N$  successive samples  $\{x[1], x[2], \dots, x[N]\}$  such that  $x_t \in \mathbb{R}$ . Three teacher models,  $T_1$ ,  $T_2$ , and  $T_3$ , and the student model  $S$  each process  $\mathbf{X}$  and produce logits for  $K$  classes. The Kullback-Leibler (KL) divergence between the student logits and each teacher’s logits is computed as  $\mathcal{L}_{KL}^i(\sigma(l_{T_i}) \parallel \sigma(l_S))$ , where  $\sigma$  is the softmax function. The total KL divergence loss is  $\mathcal{L}_{KL} = \sum_{i=1}^3 c_i \times \mathcal{L}_{KL}^i$ , with  $c_i$  being constants derived from cosine similarities. The cross-entropy (CE) loss between the student logits and true labels  $\mathbf{y}$  is  $\mathcal{L}_{CE}(\sigma(l_S), \mathbf{y})$ . The total loss is calculated as:  $\mathcal{L} = (1 - \lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{KL}$ , where  $\lambda$  balances the contributions of CE and KL divergence losses, enabling the model to learn accurate classification while aligning its predicted distribution with the targeted teacher’s distribution.

**Motivations:** Our proposed method advances SER by addressing cross-language variability, a key challenge in multilingual SER tasks that hinders generalization. We introduce a novel language-aware MTKD method, where multiple teacher models, each specialized in a different language, distill language-specific emotional cues into a student model. Using cosine similarity scores, the student model captures nuanced emotional patterns for each language while learning cross-linguistic representations. This approach improves generalizability and multilingual performance, setting a new benchmark compared to prior methods that lack language-specific optimization. Figure 1 presents an overview of our proposed MTKD method.

#### 3.1. Proposed Language-Aware MTKD Method

The proposed method utilizes three pre-trained teacher models,  $\mathbf{T}_i(\cdot)$ , where  $i \in \{1, 2, 3\}$ , each specializing in English, Finnish, and French, respectively, alongside a student model  $\mathbf{S}(\cdot)$  that can be monolingual or multilingual. The raw audio waveform  $\mathbf{X}$  is fed into each model, producing logits  $l_{T_1}$ ,  $l_{T_2}$ ,  $l_{T_3}$ , and  $l_S$ , where each set of logits corresponds to  $K$  number of classes. Then, cosine similarities between the student’s and each teacher’s logits are calculated to measure alignment:

$$cs_i = \frac{\sum_{j=1}^K l_s[j] l_{T_i}[j]}{\sqrt{\sum_{j=1}^K (l_s[j])^2} \sqrt{\sum_{j=1}^K (l_{T_i}[j])^2}} \quad \text{for } i \in \{1, 2, 3\} \quad (1)$$

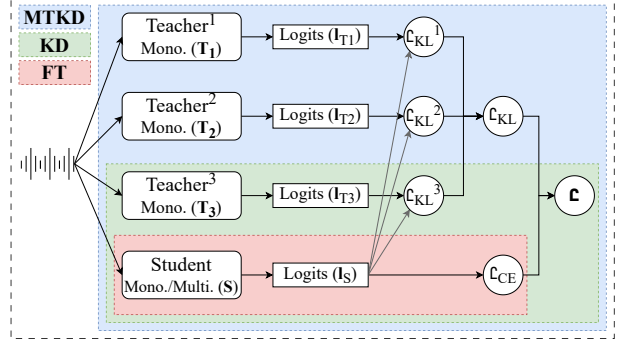


Figure 1: Proposed language-aware MTKD method.

To prioritize the most relevant teacher, these cosine similarity scores are scaled by a small temperature value  $\tau < 1.0$  and transformed via a softmax function, enhancing small differences and focusing on the most relevant teacher:  $cs'_i = \exp(cs_i/\tau) / \sum_{k=1}^3 \exp(cs_k/\tau)$ . Scaling the cosine similarity by dividing by a small  $\tau$  sharpens the differences between similarity scores, as  $\exp(cs_i/\tau)$  amplifies higher scores more strongly, making the most relevant teacher’s contribution more dominant. The student and teacher logits are then converted into probabilities using the softmax function with a higher temperature  $\tau > 1.0$ , such that:

$$P_x[j] = \frac{\exp\left(\frac{l_x[j]}{\tau}\right)}{\sum_{k=1}^K \exp\left(\frac{l_x[k]}{\tau}\right)}; \text{ for } x \in \{S, T_i\} \quad (2)$$

Dividing the logits by a high temperature value  $\tau$  smooths the probability distribution, lowering the model’s confidence in specific classes and fostering exploration by reducing the differences between logits of those  $K$  classes. Next, the KL divergence loss is computed to ensure the student model mimics the teachers’ probability distributions:

$$\mathcal{L}_{KL}^i = \sum_{j=1}^K P_S[j] \log\left(\frac{P_S[j]}{P_{T_i}[j]}\right) \quad \text{for } i \in \{1, 2, 3\} \quad (3)$$

Moreover, we aim to ensure that the student model prioritizes learning from the most relevant teacher. To do so, we combine the distribution errors from the three different teachers by weighting them with the enhanced cosine similarity scores of each teacher such that  $\mathcal{L}_{KL} = \sum_{i=1}^3 cs'_i \times \mathcal{L}_{KL}^i$ . Next, the CE loss is calculated to measure how well the student predicts the classes by comparing the student’s predicted probabilities with corresponding class labels (eq. 4).

$$\mathcal{L}_{CE} = - \sum_{j=1}^K y_j \log\left(\frac{\exp(l_S[j])}{\sum_{k=1}^K \exp(l_S[k])}\right) \quad (4)$$

Finally, to enable language-aware distillation, we combine the classification error  $\mathcal{L}_{CE}$  with the distribution error  $\mathcal{L}_{KL}$  to calculate  $\mathcal{L}$ . This ensures that the model’s prediction logit distribution is similar to the appropriate teacher and that it correctly classifies the actual labels, ensuring that the student model not only aligns its output distribution with the most relevant teacher, but also accurately classifies the actual labels.

## 4. Experimental Setup

**Evaluation Metrics:** To accommodate the varying evaluation metrics used in different studies, we report UR [22], WR [12], unweighted accuracy (UA) [11], and weighted accuracy (WA)

Table 1: *The composition of combined multilingual data.*

Dataset	Language	Splits	Train	Test
			# Samples	# Samples
IEMOCAP	English	5	≈4508	≈1241
FESC	Finnish	9	≈2798	≈461
CaFE	French	1	420	84

[24] to ensure comprehensive results and facilitate cross-study comparisons. UR and UA ignore the class distribution, while WR and WA adjust for it.

**Datasets:** We utilize three datasets: IEMOCAP [25] for English, FESC [26] for Finnish, and CaFE [27] for French. We focus on four common emotion classes, including angry, happy, neutral, and sad, which are available in all three of these datasets. This eliminates class inconsistency issue in the multilingual setup. The statistics for the considered portion of these datasets can be found in Table 1.

**Baselines:** FT-Mono. fine-tunes Wav2Vec2.0-base<sup>1</sup> on a monolingual dataset, while FT-Multi. fine-tunes same Wav2Vec2.0-base<sup>1</sup> model on a multilingual dataset. Additionally, KD-Mono. is a knowledge distillation approach where a monolingual student model (Wav2Vec2.0-base<sup>1</sup>) learns from a multilingual teacher model, which is fine-tuned FT-Multi.

**Cross Validation:** In the multilingual SER task, we use predefined splits from the English and Finnish datasets and a single split from the French dataset to train and evaluate our language-aware MTKD model on non-overlapping sets.

**Experiments:** For our experiments, we use three teacher models specialized in English, Finnish, and French SER tasks, which are fine-tuned Wav2Vec2-base<sup>1</sup> models, trained for 20 epochs with a learning rate of  $3e^{-5}$  and a batch size of 32. In our language-aware MTKD method, processed audio is used as an input to each model to produce soft outputs. We distill knowledge from the teacher models to the student model by calculating the KL divergence loss, comparing the student’s predictions with each teacher’s predictions. The losses are combined using amplified cosine similarity scores and a softmax function to prioritize the correct language. Outputs are further smoothed using a temperature value of 5. The final student model is optimized by combining cross-entropy loss and KL divergence loss, weighted at 75% and 25% respectively, ensuring effective learning of both general and language-specific SER tasks. The code for reproducing both the proposed method and the baselines is publicly available at <https://github.com/aalto-speech/mtkd4ser>.

## 5. Results

### 5.1. Quantitative Analysis

#### 5.1.1. Performance on IEMOCAP

Table 2 presents the empirical results of various training approaches, including our proposed MTKD method, on the IEMOCAP dataset. This comparison highlights the effectiveness and generalizability of three training paradigms: fine-tuning, knowledge distillation, and multi-teacher knowledge distillation. The results demonstrate that the MTKD method with a monolingual student and monolingual teachers (*MTKD-Mono.*) achieves the highest mean UR of 72.5, while the MTKD method with a multilingual student and monolingual teachers

(*MTKD-Multi.*) achieves the highest mean WR of 72.9. While other methods exhibit variability in performance and achieve superiority only in specific splits, our proposed *MTKD-Multi.* method consistently outperforms them across most splits, underscoring its robustness.

#### 5.1.2. Comparison with Baselines

Table 3 presents a detailed comparison of our proposed MTKD method against several baselines for SER in English, Finnish, and French. The performance is reported as the mean of five splits for the IEMOCAP dataset and as the mean of the best and worst splits for the FESC dataset. A more comprehensive analysis was conducted on the IEMOCAP dataset, given its status as a widely used benchmark, whereas the investigation of the FESC dataset was limited to the best and worst splits to optimize computational resources and time. Empirical results indicate that while FT and KD yield promising outcomes, MTKD demonstrates superior generalization and robustness, achieving the highest average scores with a substantial amount of data. In the monolingual setup, MTKD with a monolingual student (*MTKD-Mono.*) achieves the highest UR and WR in the IEMOCAP and FESC datasets. Likewise, in the multilingual setup, MTKD with a multilingual student (*MTKD-Multi.*) attains the highest UR and WR scores.

The *MTKD-Multi.* method demonstrates consistent performance across evaluations. The impact of dataset size on its effectiveness is evident in English SER with the IEMOCAP dataset. Given its relatively large training set (5.7 hours), IEMOCAP exhibits stable performance between the best and average results (Table 2). Furthermore, while *MTKD-Mono.* achieves a marginal improvement over baseline methods in the monolingual setup, *MTKD-Multi.* outperforms the baseline by a significant margin, highlighting its superior ability to leverage multilingual training for enhanced performance. However, MTKD-Mono. achieves the highest UR score across compared monolingual and multilingual methods. In contrast, Finnish SER using the FESC dataset, which contains 3.33 hours of training data, shows a performance drop. For French SER, with only 0.5 hours of training data, *KD-Mono.* demonstrates superior performance compared to other methods. The low performance of *MTKD-Multi.* in French SER is expected, as French data is a minority group in each training batch. As a result, the teacher selected based on the cosine similarity scores is often not the one specialized in French SER, leading to less effective learning for French compared to English and Finnish SER.

#### 5.1.3. Comparison with Existing State-of-the-Art (SOTA)

Table 4 presents a comparative analysis of various SOTA methods for English SER using the IEMOCAP dataset. Our proposed monolingual and multilingual MTKD methods outperform existing SOTA approaches in terms of WA and UR, respectively. However, the slightly higher UA score reported in [16] is anticipated, given that their approach utilizes 20% of the entire dataset, including duplicated data.

### 5.2. Qualitative Analysis

Figure 2 delineates the average inter-class performance of various SER methods. The comparison includes monolingual (English) and multilingual methods, highlighting the impact of different training approaches. It clarifies that integrating multilingual data enhances overall system effectiveness.

Our proposed method, MTKD-Multi., outperforms other

<sup>1</sup><https://huggingface.co/facebook/wav2vec2-base>

Table 2: Performance of our proposed MTKD methods alongside other baselines on the IEMOCAP dataset, where 'Mono.' refers to monolingual, and 'Multi.' stands for multilingual configurations. The upper (numerator) and lower (denominator) bounds of the confidence interval are reported for both UR and WR.

Split	FT-Mono.		FT-Multi.		KD-Mono.		MTKD-Mono.		MTKD-Multi.											
	UR	WR	UR	WR	UR	WR	UR	WR	UR	WR										
Split 1	69.8	$\frac{74.5}{64.4}$	65.8	$\frac{70.6}{65.8}$	70.6	$\frac{74.1}{66.7}$	70.7	$\frac{74.0}{67.2}$	72.5	$\frac{77.4}{66.9}$	70.3	$\frac{75.1}{64.9}$	72.4	$\frac{76.8}{67.1}$	68.5	$\frac{73.1}{63.2}$	<b>72.5</b>	$\frac{76.0}{68.7}$	<b>73.4</b>	$\frac{76.6}{69.9}$
Split 2	<b>79.3</b>	$\frac{83.9}{73.3}$	<b>78.1</b>	$\frac{82.6}{72.8}$	70.9	$\frac{74.4}{67.3}$	74.9	$\frac{77.9}{71.6}$	76.7	$\frac{81.5}{71.0}$	74.8	$\frac{79.5}{69.3}$	77.8	$\frac{82.4}{72.1}$	76.0	$\frac{80.5}{70.6}$	71.2	$\frac{74.8}{67.3}$	71.0	$\frac{74.4}{67.3}$
Split 3	66.2	$\frac{71.3}{60.7}$	65.9	$\frac{71.0}{60.4}$	67.7	$\frac{71.2}{64.0}$	69.5	$\frac{72.8}{66.0}$	69.5	$\frac{74.5}{64.0}$	69.1	$\frac{74.1}{63.6}$	68.9	$\frac{73.9}{63.5}$	69.0	$\frac{74.0}{63.6}$	<b>71.4</b>	$\frac{74.9}{67.8}$	<b>71.5</b>	$\frac{74.8}{67.9}$
Split 4	68.5	$\frac{73.9}{62.7}$	70.2	$\frac{75.1}{64.8}$	64.2	$\frac{68.0}{60.3}$	69.7	$\frac{73.0}{66.1}$	70.2	$\frac{75.3}{64.3}$	70.1	$\frac{75.0}{64.7}$	<b>72.3</b>	$\frac{77.6}{66.3}$	72.8	$\frac{77.8}{67.2}$	71.6	$\frac{75.3}{67.7}$	<b>73.8</b>	$\frac{77.0}{70.3}$
Split 5	72.2	$\frac{76.9}{66.8}$	70.5	$\frac{75.1}{65.4}$	67.9	$\frac{71.3}{64.2}$	70.2	$\frac{73.4}{66.8}$	72.2	$\frac{77.0}{66.7}$	71.2	$\frac{75.8}{66.0}$	71.0	$\frac{75.9}{65.5}$	69.7	$\frac{74.4}{64.5}$	<b>73.0</b>	$\frac{76.3}{69.3}$	<b>74.7</b>	$\frac{77.8}{71.3}$
Mean	71.2	$\frac{76.1}{65.6}$	70.1	$\frac{74.9}{65.8}$	68.2	$\frac{71.8}{64.5}$	71.0	$\frac{74.2}{67.6}$	72.2	$\frac{77.1}{66.6}$	71.1	$\frac{76.1}{65.8}$	<b>72.5</b>	$\frac{77.3}{66.9}$	71.2	$\frac{76.0}{65.8}$	71.9	$\frac{75.5}{68.2}$	<b>72.9</b>	$\frac{76.1}{69.3}$

Table 3: Comparison of our proposed methods with baselines for SER in English, Finnish, and French.

	IEMOCAP		FESC		CaFE	
	UR	WR	UR	WR	UR	WR
FT-Mono.	71.2	70.1	59.5	62.0	78.1	78.6
KD-Mono.	72.2	71.1	62.7	67.8	<b>82.3</b>	<b>79.8</b>
<b>MTKD-Mono.</b>	<b>72.5</b>	<b>71.2</b>	<b>62.9</b>	<b>68.1</b>	78.1	75.0
FT-Multi.	68.2	71.0	62.7	64.3	<b>77.1</b>	<b>79.8</b>
<b>MTKD-Multi.</b>	<b>71.9</b>	<b>72.9</b>	<b>63.4</b>	<b>66.1</b>	73.4	72.3

Table 4: Performance comparison of our MTKD method against SOTA methods on the IEMOCAP dataset. Paradigm CKD denotes Cubic Knowledge Distillation. \* indicates redundancy.

Method	Paradigm	WA	UA	UR
Wav2vec2-PT [9]	FT	—	—	67.2
DKDFMH [16]	KD	—	<b>77.1*</b>	—
P-TAPT [11]	FT	—	74.4	—
Vesper-4 [24]	KD	68.4	69.3	—
CubicKD [28]	CKD	63.3	—	—
<b>MTKD-Mono.</b>	MTKD	<b>76.0</b>	76.2	<b>72.5</b>
<b>MTKD-Multi.</b>	MTKD	74.8	74.9	71.9

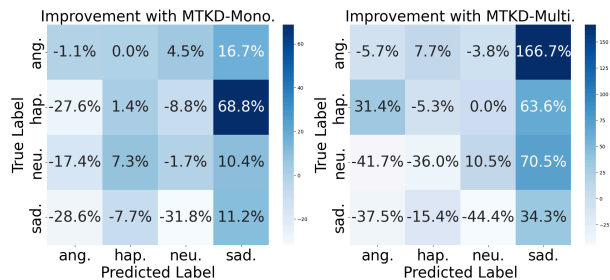


Figure 2: Performance improvement of MTKD-Mono. over FT-Mono. on monolingual set (Left) and of MTKD-Multi. over FT-Multi. on multilingual set (Right), respectively.

approaches in emotion recognition by reducing misclassification numbers. Compared to the monolingual fine-tuning method (FT-Mono.), which struggles with identifying neutral emotions, and the monolingual knowledge-distillation method (KD-Mono.), which still show confusion between happiness and neutral, our method (MTKD-Mono.) demonstrates superior accuracy and generalization, especially for sadness. The multilin-

gual fine-tuning method (FT-Multi.) improves performance but has issues between neutral and sadness. However, our MTKD-Multi. method enhances overall system effectiveness by improving performance for sadness and neutral emotions.

### 5.3. Error Analysis

The confusion matrices in Figure 2 reveals that the MTKD-Mono. method has difficulty distinguishing between Happiness and Neutral classes, while the MTKD-Multi. method struggles with Anger class, likely due to being minority classes. Despite these challenges, MTKD-Mono. performs better in reducing misclassifications between Anger and Sadness, and MTKD-Multi. excels in reducing errors between Neutral and Sadness. In terms of recall, MTKD-Mono. improves scores for Happiness and Anger, whereas MTKD-Multi. boosts scores for Neutral and Sadness. Overall, MTKD-Multi. shows superior performance by reducing misclassification rates and improving recall, as highlighted by its higher average UR and WR scores across languages and enhanced performance in English SER. This indicates that leveraging multilingual data and multi-teacher knowledge distillation provides a significant advantage over traditional monolingual methods.

## 6. Conclusion

This study demonstrates the effectiveness of the MTKD approach in markedly enhancing the performance of multilingual SER systems. By leveraging multiple monolingual teacher models, our method facilitates the transfer of cross-linguistic emotional knowledge, outperforming standard FT and conventional KD methods across various languages and datasets. This highlights the importance of optimizing emotional knowledge transfer for improved multilingual emotional understanding. The quantitative analysis of the IEMOCAP dataset shows that MTKD-Mono achieves the highest mean UR of 72.5, while MTKD-Multi attains the highest mean WR of 72.9. Furthermore, in the FESC dataset, MTKD-Mono achieves the highest WR of 68.1, and MTKD-Multi attains the highest UR of 63.4 among all compared methods. These results validate that training on out-of-distribution data enhances generalization, even with limited in-domain data. However, our method currently relies on homogeneous teachers. Moreover, the associated computational demands and challenges in selecting teacher models underscore areas that require further improvement. In future studies, we will explore heterogeneous teachers, alternative similarity metrics, and broader linguistic and cultural contexts to develop a more robust and versatile SER system.

## 7. Acknowledgements

The computational resources were provided by Aalto SciencelT. The authors are grateful for the Academy of Finland project funding number 345790 in ICT 2023 programme's project "Understanding speech and scene with ears and eyes" and the Business Finland project LAREINA under Grant 7817/31/2022.

## 8. References

- [1] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Speech emotion recognition using supervised deep recurrent system for mental health monitoring," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. IEEE, 2022, pp. 1–6.
- [2] T. Grósz, A. Virkkunen, D. Porjazovski, and M. Kurimo, "Discovering relevant sub-spaces of bert, wav2vec 2.0, electra and vit embeddings for humor and mimicked emotion recognition with integrated gradients," in *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, 2023, pp. 27–34.
- [3] H. Zhang, D. Chen, and C. Wang, "Confidence-aware multi-teacher knowledge distillation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4498–4502.
- [4] —, "Adaptive multi-teacher knowledge distillation with meta-learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1943–1948.
- [5] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, "Multi-label emotion analysis in conversation via multimodal knowledge distillation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6090–6100.
- [6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [7] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [9] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.
- [10] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7026–7029.
- [11] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [12] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.
- [13] A. Chakhtouna, S. SEKKATE, and A. Abdellah, "Unveiling embedded features in wav2vec2 and hubert models for speech emotion recognition," *Procedia Computer Science*, vol. 232, pp. 2560–2569, 2024.
- [14] L. Gao, K. Xu, H. Wang, and Y. Peng, "Multi-representation knowledge distillation for audio classification," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 5089–5112, 2022.
- [15] Z. Hao, J. Guo, K. Han, Y. Tang, H. Hu, Y. Wang, and C. Xu, "One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [16] Z. Zhao, H. Wang, H. Wang, and B. Schuller, "Hierarchical network with decoupled knowledge distillation for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [17] H.-I. Yun and J.-S. Park, "End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation," *Multimedia Tools and Applications*, vol. 82, no. 15, pp. 22 759–22 776, 2023.
- [18] K. Sridhar and C. Busso, "Ensemble of students taught by probabilistic teachers to improve speech emotion recognition," in *INTERSPEECH*, 2020, pp. 516–520.
- [19] R. Takashima, S. Li, and H. Kawai, "An investigation of a knowledge distillation method for ctc acoustic models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5809–5813.
- [20] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Interspeech*, 2017, pp. 3697–3701.
- [21] Z. Yang, L. Shou, M. Gong, W. Lin, and D. Jiang, "Model compression with two-stage multi-teacher knowledge distillation for web question answering system," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 690–698.
- [22] Z. Ren, T. T. Nguyen, Y. Chang, and B. W. Schuller, "Fast yet effective speech emotion recognition with self-distillation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [23] K.-P. Huang, T.-H. Feng, Y.-K. Fu, T.-Y. Hsu, P.-C. Yen, W.-C. Tseng, K.-W. Chang, and H.-Y. Lee, "Ensemble knowledge distillation of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [24] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2024.
- [25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [26] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.
- [27] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proceedings of the 9th ACM multimedia systems conference*, 2018, pp. 399–402.
- [28] Z. Lou, S. Otake, Z. Li, R. Kawakami, and N. Inoue, "Cubic knowledge distillation for speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5705–5709.