Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models

Danilo de Oliveira, Navin Raj Prabhu, Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany

{danilo.oliveira, navin.raj.prabhu, timo.gerkmann}@uni-hamburg.de

Abstract

In large part due to their implicit semantic modeling, self-supervised learning (SSL) methods have significantly increased the performance of valence recognition in speech emotion recognition (SER) systems. Yet, their large size may often hinder practical implementations. In this work, we take HuBERT as an example of an SSL model and analyze the relevance of each of its layers for SER. We show that shallow layers are more important for arousal recognition while deeper layers are more important for valence. This observation motivates the importance of additional textual information for accurate valence recognition, as the distilled framework lacks the depth of its large-scale SSL teacher. Thus, we propose an audio-textual distilled SSL framework that, while having only ~20% of the trainable parameters of a large SSL model, achieves on par performance across the three emotion dimensions (arousal, valence, dominance) on the MSP-Podcast v1.10 dataset.

Index Terms: speech emotion recognition, self-supervised learning, knowledge distillation, paralinguistics, semantics

1. Introduction

Speech signals carry rich information on an individual's emotional states, expressed through both paralinguistic and semantic cues. Backed by the circumplex model [1], research on speech emotion recognition (SER) is typically performed by studying the emotional expressions across multiple dimensions, namely the activation-deactivation dimension (*arousal*), the pleasure-displeasure dimension (*valence*), and the speaker confidence-diffidence dimension (*dominance*). Prior works in literature reveal that certain speech cues carry more information on a particular emotional dimension. For instance, the arousal dimension is well explained by paralinguistic cues [2], and semantic cues are more informative of the valence dimension [3]. This calls for techniques that learn both paralinguistic and semantic cues simultaneously.

The self-supervised learning (SSL) paradigm is a good candidate to fulfill these requirements. Methods from this domain leverage unlabeled data in a pre-training stage as a way of learning robust representations of an input's underlying structure. The pre-training of SSL models is usually succeeded by a fine-tuning stage, in which the model is fed labeled data and trained in a supervised manner for a target downstream task. Following the success of BERT [4] in the natural language processing (NLP) domain, models like wav2vec 2.0 [5] and HuBERT [6] have provided great performance boosts in speech tasks, ranging from automatic speech recognition (ASR) to speaker identification (SI) [7] and SER [8, 9].

While SER systems typically perform well in terms of arousal and dominance, modeling valence from speech signals

is a challenging task [10, 2], resulting in a performance gap between arousal/dominance and valence estimation. Prior works have successfully managed to reduce this gap by fine-tuning wav2vec 2.0 and HuBERT for emotion recognition [9]. The improved valence prediction results come from the fact that these transformer-based SSL models can *implicitly* capture semantic content present in speech, along with the paralinguistics [11]. However, such models still underperform on valence compared with models that *explicitly* include semantic information through BERT-encoded features [12]. These findings suggest that the semantic information is not completely modeled by speech-only SSL techniques.

Despite the reduced performance gap between arousal and valence, an important drawback of SSL models is their size: the smallest versions of BERT, wav2vec 2.0, and HuBERT contain 110, 95, and 90 million parameters, respectively, making their use costly or even prohibitive in some cases. Knowledge distillation [13] methods have found success in addressing this issue, in particular in the NLP domain [14, 15]. The goal of these methods is to compress the knowledge acquired by a large network (the teacher) by transferring it to a smaller one (the student). Applied to HuBERT, this framework results in DistilHuBERT [16], with only 25% of the parameters of its teacher.

An analysis of DistilHuBERT suggests that the distilled network is good at encoding paralinguistic information, given the high importance of its representations in the SI task [16]; this indicates potentially good performance in arousal estimation. Moreover, the inner representations of wav2vec 2.0 have an acoustic-linguistic hierarchy [17], where the information embedded in each layer output mutates with network depth, from an acoustic to a semantic nature. These findings further motivate us to investigate how well the distilled model fares in predicting each of the emotional dimensions, especially valence.

In this paper, we make the following contributions: We show that the shallow layers of SSL models are more important for arousal recognition while the deeper layers are more important for valence recognition. We argue that the students in distilled models like DistilHuBERT lack the required depth to properly model valence. Therefore, we show that adding textual information is particularly helpful for distilled SSL networks like DistilHuBERT, considerably more important than for non-distilled large-scale SSL networks. To the best of our knowledge, we are the first in the literature to use distilled SSL models in arousal, valence and dominance modeling and to analyze the implication of the distillation process (i.e., layer selection and compression) towards bridging the gap between valence and arousal estimation.

2. Proposed methodology

The task of emotion recognition is formulated here as follows: given a single-channel audio input containing a spoken utterance $\mathbf{X}_A \in \mathbb{R}^{1 \times S}$, where S is the number of samples, we want to estimate three emotional expression scalar values: arousal (Y_a), valence (Y_v) and dominance (Y_d). Our SER model f_A should map the input utterance $\mathbf{X}_A$ to an estimate of the three emotion dimensions, simultaneously:

$$\widehat{\mathbf{Y}} = f_A(\mathbf{X}_A), \qquad (1)$$

where $\widehat{\mathbf{Y}} = \mathrm{concat}(\widehat{Y}_a, \widehat{Y}_v, \widehat{Y}_d)$.

In the multi-modal case, we also have text as an additional input. The tokenized text is denoted by $\mathbf{X}_T \in \mathbb{N}^{1 \times N}$, where N is the number of tokens in the utterance. The audio-textual model mapping is therefore

$$\widehat{\mathbf{Y}} = f_{A+T}(\mathbf{X}_A, \mathbf{X}_T). \qquad (2)$$

Figure 1 illustrates the models. They contain essentially audio and text encoders (SSL_A and SSL_T, respectively), pooling operations and a feed-forward regression head (FFN). The pooling block consists of average- and max-pooling across the sequence dimension. The FFN layer contains two fully-connected layers and a hyperbolic tangent activation function. Its final layer has three output features to estimate $\widehat{\mathbf{Y}}$.

2.1. Audio-only SSL framework

The audio-only system is depicted in Figure 1a. Aiming at extracting features from the time-domain input $\mathbf{X}_A$, we employ pre-trained models as SSL_A, namely wav2vec 2.0, HuBERT and DistilHuBERT. They share the same structure: a convolutional encoder followed by transformer blocks. The convolutional encoder creates frames of latent representations that act as tokens for the creation of contextualized features $\mathbf{C}_A \in \mathbb{R}^{d_A \times K}$ by the transformer encoder, where d_A is the hidden dimension size and K is the number of frames, dependent on the configuration of the convolutional blocks. The transformer encoder is similar to its text counterpart, presented in the next section.

Wav2vec 2.0 uses product quantization to generate a finite set of representations and compares them against the context representations using a contrastive loss. HuBERT, on the other hand, is trained in two separate steps: the first is an offline step that creates discrete pseudo-labels by clustering audio-based features, and the second consists in masked prediction of cluster assignments. Finally, DistilHuBERT is a model obtained by distilling HuBERT through a framework of three prediction heads that aim at estimating HuBERT's 4th, 8th and 12th layers. Though the models make use of largely different pre-training techniques, during fine-tuning/inference their architectures differ essentially in the number of layers and hidden size.

In order to use the extracted representations for our emotion recognition task, the pooling block compresses them to an utterance-level representation $\mathbf{P}_A \in \mathbb{R}^{d_A \times 1}$. This is then fed to the FFN head, which outputs the estimates for arousal, valence and dominance.

2.2. Audio-textual SSL framework

For experiments including text inputs, the token IDs from the tokenized text $\mathbf{X}_T$ pass through an embedding layer, and are then fed to a BERT-like encoder. This class of language models is pre-trained by randomly masking tokens and attempting to predict them using the unmasked ones as context. The architecture is
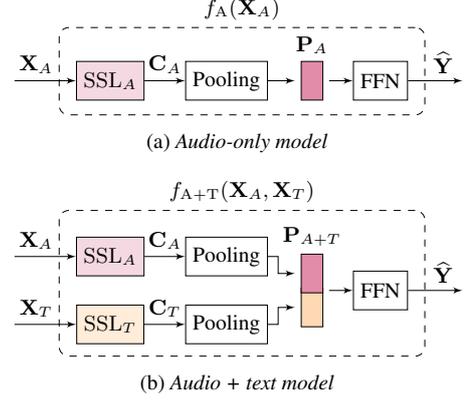

(a) *Audio-only model*

(b) *Audio + text model*

Figure 1: *System architecture, mapping a spoken utterance* $\mathbf{X}_A$ *(and, in the audio-textual case, its corresponding transcript* $\mathbf{X}_T$*) to a set of arousal, valence and dominance estimates* $\widehat{\mathbf{Y}}$*.*

essentially a sequence of bidirectional transformer layers. The resulting vector $\mathbf{C}_T \in \mathbb{R}^{d_T \times N}$, with d_T being the size of the text hidden dimension, is an embedding enriched by contextual information.

In our experiments, we use TinyBERT [15] as SSL_T, with the intent of minimizing overhead. It is a distilled version of BERT, trained by applying distillation in two steps: pre-training and fine-tuning, aiming at reproducing BERT's capabilities of generalization and downstream task performance, respectively. The authors report 96.8% of the performance of BERT base on an NLP task benchmark, even though the model is 7.5x smaller.

Inspired by [18, 19, 9], we use a simple concatenation method for the fusion of modalities, shown in Figure 1b: the pooled features from each encoder are concatenated in the feature dimension, resulting in $\mathbf{P}_{A+T} \in \mathbb{R}^{(d_A + d_T) \times 1}$. We then pass it through a feed-forward block similar to the audio-only case, only with the number of input features adjusted to incorporate the additional text representations.

3. Implementation details

3.1. Dataset

For training, validation and testing, we use the MSP-Podcast dataset [20], version 1.10. It contains approximately 166 hours of audio extracted from podcast data, labeled at utterance level for arousal, valence and dominance on a scale of 1 to 7. V1.10 features human-labeled transcripts, which serve as our oracle text information. The dataset is split into four partitions: `train`, `development`, `test1` and `test2`. Differently from `test1`, the segments in `test2` originate from podcasts not present in any other partitions. We therefore consider it our *unseen scenarios* set, while `test1` is our *seen scenarios* test data.

3.2. Baseline and proposed models

In our experiments, we propose a framework based exclusively on distilled models and compare it against *base* and *large* SSL_A baselines. As the *base* SSL_A model, we use HuBERT base, which has 12 transformer layers and hidden size $d_A = 768$ [6]. As the *large* SSL_A model, we use the *pruned* w2v2-L-robust from [9], which we denote as w2v2-L-robust(p). It uses the first 12 (out of 24) layers of the original model, with a hidden size $d_A = 1024$. As the proposed distilled SSL_A, we

Table 1: *CCC scores on MSP Podcast (v1.10). Seen scenarios is the test set containing segments from podcasts included in the train or development sets, while the unseen scenarios set contains segments from podcasts not present in any other partition.*

Modality	Model	#Params (M)	Seen Scenarios			Unseen Scenarios		
			Arousal	Valence	Dominance	Arousal	Valence	Dominance
Audio	w2v2-L-robust(p)	165	**0.627**	**0.470**	**0.521**	0.462	**0.263**	0.390
	HuBERT base	95	0.608	0.440	0.486	0.388	0.225	0.344
	DistilHuBERT	**24**	0.622	0.328	0.513	**0.470**	0.169	**0.395**
Audio + Text	w2v2-L-robust(p)	180	**0.619**	**0.560**	0.484	0.469	**0.347**	0.353
	HuBERT base	109	0.613	0.532	0.493	0.461	0.315	**0.396**
	DistilHuBERT	**39**	0.614	0.519	**0.509**	**0.475**	0.333	0.392

use the DistilHuBERT model, that has the same hidden size as HuBERT ($d_A = 768$) but contains only 2 layers. For our audio-textual experiments, as the distilled SSL_T, we employ the 4-layer TinyBERT encoder, with $d_T = 312$.

We make use of the pre-trained SSL models available on Huggingface[1]. When fine-tuning, we follow the usual procedure of freezing the weights of the convolutional encoders present in the speech models. Emotional expression target labels are normalized to $[0, 1]$. We use Adam [21] as the optimizer, with a learning rate of 10^{-5} and early stopping to prevent overfitting. We employ dropout of 0.1 throughout the model. The training data are batched with batch size 16, using buckets sorted by length in order to reduce the amount of padding. The networks are fine-tuned on a single NVIDIA A6000 GPU.

3.3. Loss function

To train the proposed architecture, we use the concordance correlation coefficient (CCC) loss [22], a widely used loss function in SER research [23, 2, 3]. The CCC measures the similarity between two variables and varies between -1 and $+1$, where $+1$ denotes perfect similarity and 0 denotes perfect orthogonality between the variables. For Pearson's correlation coefficient ρ, the CCC between $\mathbf{y}$ and its estimate $\widehat{\mathbf{y}}$ is formulated as

$$\mathcal{L}_{\text{CCC}}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\widehat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\widehat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\widehat{\mathbf{y}}})^2}, \qquad (3)$$

where μ, and σ are the mean and standard deviation, respectively, of the corresponding variables. From (3), it can be noted that the CCC takes both the linear correlation and the bias between $\mathbf{y}$ and $\widehat{\mathbf{y}}$ into consideration while quantifying their similarity, hence it is preferred over the Pearson correlation as a loss function for SER. Our training minimizes $1 - \mathcal{L}_{\text{CCC}}(\mathbf{y}, \widehat{\mathbf{y}})$ averaged across the three emotional dimensions.

4. Results and Discussion

4.1. Explicit inclusion of semantic information

We evaluate our audio-only and audio-textual versions of DistilHuBERT along with the HuBERT base and w2v2-L-robust(p) baselines. The results are shown in Table 1. Firstly, from the *audio-only* models, it can be observed that DistilHuBERT manages to match the large SSL baseline w2v2-L-robust(p), for both seen and unseen scenario test data even though it has only ~15% of the total number of parameters. However, the lack of modeling depth and capacity seems to heavily impact valence estimation. The HuBERT base version lags behind in arousal and dominance, but has valence scores closer to those of w2v2-L-robust(p).

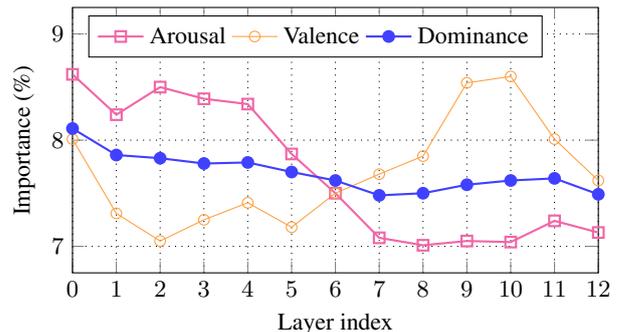
[1] https://huggingface.co

Figure 2: *Normalized importance given to each layer of HuBERT when training in a frozen weights setting. Index 0 corresponds to the output of the convolutional encoder. The subsequent indexes reference each of the transformer encoder's 12 layers.*

Secondly, the *audio-textual* models considerably outperform their audio-only counterparts in terms of valence estimation, even in the case of w2v2-L-robust(p), whose dominance modeling performance decreased, while arousal remained at roughly the same level. The audio-textual versions of the base and distilled HuBERT models either improved or remained at the same level in all cases. Notably, the audio-textual distilled model performs on par with the larger model not only in arousal and dominance but also valence, thus confirming the effectiveness of the explicit inclusion of text, particularly for distilled models.

4.2. Layer importance for the emotional dimensions

As presented in Table 1, the audio-only DistilHuBERT, despite its performance in arousal estimation, performs poorly in terms of valence. To explain this behavior, and also to better understand what kind of information the distilled model learns from the teacher, we proceed by analyzing the relevance of the transformer layers of the pre-trained HuBERT base encoder for arousal, valence and dominance modeling, this time individually. In a similar fashion to [8, 16], we perform a weighted sum of the outputs of the encoder's layers. The weights are normalized to sum to one, and the encoder weights are kept frozen while training the regression head for estimating the target emotional expression. The resulting layer-weighting parameter values found during training are plotted in Figure 2. Index 0 corresponds to the convolutional encoder's outputs, and 1-12 are each of the transformer layers.

Figure 2 reveals that for arousal estimation the first layers up to layer 4 are more important than deeper layers, which hints at the specialization of the first layers on paralinguistic features.

This coincides with the findings of [17] for the similar wav2vec 2.0 model. It is interesting to see that the layer importance follows an opposite trend for valence: Here, lower importance is observed in the first few transformer blocks while deeper layers are more important to model valence. A peak is observed at layers 9 and 10. It raises a concern for the use of DistilHuBERT, since it lacks depth to model the deeper layers. Additionally, its two layers are initialized from the first two of HuBERT. This can explain the distilled model's performance gap between arousal and valence estimations. Furthermore, it also motivates the explicit inclusion of text in SSL models, especially for the distilled models.

4.3. Fine-tuning vs. freezing

Motivated by related works in audio-textual emotion recognition that either fine-tune the SSL encoders' weights [19] or keep them frozen while only training a regression head [18, 9, 12], we run experiments to compare how much each approach helps to improve valence predictions. Figure 3 displays the relative improvement of the valence CCC score over the audio-only case for each of the considered models: large-pruned, base and distilled. "FT" indicates a model with fine-tuned SSL encoders, while "FT⇒FRZ" represents the fusion process used in [9]: a pre-trained speech model is fine-tuned for emotion recognition, then its encoder weights are frozen and used alongside an also frozen text encoder to train a regression head.

We observe in Figure 3 that textual information is particularly helpful for smaller models. This indicates that larger capacity and/or more training data help the model learn some level of semantic information even without explicit textual input. Nevertheless, in all cases the performance is increased by including textual information, confirming the fact that there's still relevant information in the textual input that is not captured by pre-training with audio alone. Furthermore, we observe that fine-tuning brings even larger improvements in all cases; in particular, for the case of the distilled model, valence estimation performance was double that of the audio-only setting in the unseen scenario test set. We therefore hypothesize that fine-tuning the text model helps it focus on modeling the specific semantic information missing from the speech representation.

4.4. Quality of transcripts

When implementing systems that make additional use of text, the audio needs to be transcribed, most practically via an automatic speech recognition (ASR) system. Although practical, ASR models may introduce transcription errors. In order to examine the robustness of the text-informed framework, we run our proposed distilled-only, fine-tuned framework using text transcribed by two ASR systems with different model sizes, and compare it with the model trained on human transcriptions.

The results are presented in Table 2. We can see that the performance is maintained across ASR models, irrespective of their number of parameters. This suggests that the framework is robust with respect to text, as long as the main semantic information is preserved. This claim is further backed by the fact that performance of ASR-generated transcripts is as good as that of human transcripts. It should be noted that the models are run on clean speech; noisy and reverberant conditions are expected to worsen the quality of transcripts and should be addressed by future studies.
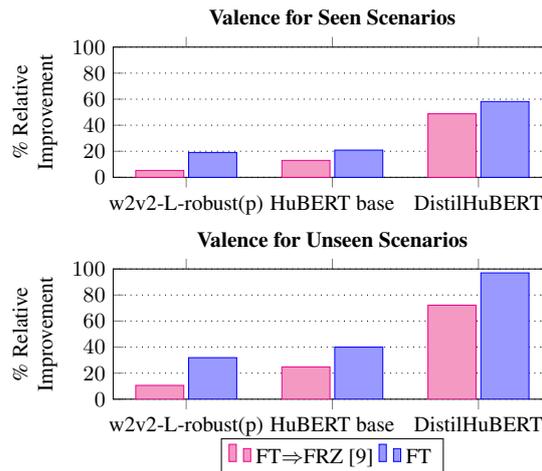

Figure 3: *Relative valence CCC improvement of audio+text models over the audio-only case. "FT⇒FRZ" corresponds to the method used in [9]. "FT" refers to the fine-tuning used in this paper.*

Table 2: *CCC performance of the fine-tuned DistilHuBERT + TinyBERT system with different transcription methods. "A", "V" and "D" denote arousal, valence and dominance, respectively.*

Transcription method	#Params (only ASR)	Seen Scenarios			
		WER	A	V	D
Human	–	–	0.614	0.519	0.509
Whisper base [24]	74M	22.2%	0.620	0.521	0.511
Whisper tiny [24]	39M	24.1%	0.618	0.524	0.510

5. Conclusions

In this study we proposed an audio-textual emotion recognition framework based on distilled models. We highlighted the particular importance of multi-modal audio and text inputs for robust arousal, valence and dominance estimation when using our distilled model. Despite having only ∼20% of the trainable parameters of the largest baseline, the proposed framework's performance is on par with base and large models not only on seen scenarios, but importantly also on unseen scenario data. We investigated the relevance of HuBERT's inner representations to each of the three emotion dimensions and found the initial layers to be more important for arousal modeling, while the deeper layers focus on information instrumental to valence estimation. This analysis further validates the need for text as extra input to distilled networks for improved valence modeling, as these shallow models cannot extract semantic information from speech as easily as their teacher counterparts. Lastly, we confirmed the robustness of our audio-textual network by training it on machine-transcribed audio-text pairs, without loss of performance. Distillation is a promising way to make SSL models more practical, but it is necessary to ensure that both paralinguistic and semantic information is available in order to have robust arousal and valence estimation.

6. References

[1] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[2] N. Raj Prabhu, G. Carbajal, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks," in *Interspeech*, Sep. 2022.

[3] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, "Speech Emotion Recognition Using Semantic Information," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Jun. 2021.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, Jun. 2019, pp. 4171–4186.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Inf. Proc. Systems (NeurIPS)*, vol. 33, 2020, pp. 12 449–12 460.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE Trans. on Audio, Speech, and Lang. Process. (TASLP)*, vol. 29, pp. 3451–3460, 2021.

[7] N. Vaessen and D. A. Van Leeuwen, "Fine-Tuning Wav2Vec2 for Speaker Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 7967–7971.

[8] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Interspeech*, Aug. 2021, pp. 3400–3404.

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," Mar. 2022, arXiv:2203.07378 [cs, eess].

[10] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Trans. on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, 2022.

[11] A. Triantafyllopoulos, J. Wagner, H. Wierstorf, M. Schmitt, U. Reichel, F. Eyben, F. Burkhardt, and B. W. Schuller, "Probing Speech Emotion Recognition Transformers for Linguistic Knowledge," in *Interspeech*, Sep. 2022, pp. 146–150.

[12] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation Learning Through Cross-Modal Conditional Teacher-Student Training For Speech Emotion Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 6442–6446.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," in *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2)- NeurIPS 2019*, Feb. 2020.

[15] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2020, pp. 4163–4174.

[16] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2022, pp. 7087–7091.

[17] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-Wise Analysis of a Self-Supervised Speech Representation Model," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2021, pp. 914–921.

[18] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, May 2020, pp. 6484–6488.

[19] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Interspeech*, Oct. 2020, pp. 3755–3759.

[20] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Int. Conf. on Learning Representations (ICLR)*, 2015.

[22] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[23] B. W. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, arXiv:2212.04356 [cs, eess].