A Three-Stage Framework for Speaker-Invariant Speech Emotion Recognition

Anonymous ACL submission

Abstract

Label scarcity remains a significant challenge in speech emotion recognition (SER), often limiting the effectiveness of training models from scratch. Furthermore, speaker variability in acoustic representations hinders the generalization of emotion recognition systems. Prior research has demonstrated that mitigating speaker-related information can improve performance in SER tasks. In this work, we propose an efficient method to learn speaker-invariant representations by suppressing speaker identity from a pre-trained model (Wav2Vec2.0). Our approach enhances the robustness of emotion classification while addressing the limitations of limited labeled data and inter-speaker variability.

1 Introduction

Speech Emotion Recognition (SER) has long faced challenges due to the scarcity of labeled data. Despite continuous efforts to construct emotion-labeled speech datasets, obtaining accurate emotion annotations remains costly and labor-intensive. Consequently, SER systems often suffer from limited training resources, making it difficult to generalize across speakers and contexts.

Speech signals inherently carry multiple types of information beyond linguistic content, including speaker-specific characteristics, emotional expressions, and social cues. This diversity complicates the extraction of emotion-related features, as the underlying representations are often entangled with speaker identity. Among these confounding factors, *speaker variability*—i.e., inter-speaker differences in emotional expression—has been shown to significantly hinder the generalization performance of SER models(Li et al., 2021).

Recent advances in self-supervised learning (SSL) have led to significant improvements in a wide range of speech processing tasks(Wang and Yang, 2025). SSL-based models such as HuBERT(Hsu et al., 2021) and Wav2Vec 2.0(Ando and Zhang, 2005) are pre-trained on large-scale unlabeled corpora, and are capable of learning rich acoustic and linguistic representations. These models have demonstrated strong transferability and effectiveness, even with limited downstream data.

However, directly applying such pre-trained models to SER remains suboptimal. The learned representations often encode speaker-specific features, resulting in degraded performance on unseen speakers. When there is a distributional shift in speaker identities between the training and testing sets, models tend to overfit to speaker characteristics rather than generalizing to emotion-related cues.

To address this limitation, several studies have employed *Gradient Reversal Layers (GRL)*(Ganin et al., 2016) in conjunction with domain-adversarial training to encourage the learning of speaker-invariant representations. For example, prior work utilized GRL-based architectures involving 1D convolutional layers, recurrent networks, and pooling mechanisms. While effective to some extent, these approaches often lack the representational power and generalization ability provided by modern pre-trained models.

In this paper, we propose a novel SER framework that integrates pre-trained speech models (e.g., Wav2Vec 2.0) with a speaker identification model (ECAPA-TDNN)(Desplanques et al., 2020), to suppress speaker-specific information through domain-adversarial learning. Specifically, we apply a GRL-based training strategy to encourage the pre-trained encoder to produce representations that are informative for emotion classification but invariant to speaker identity.

We conduct extensive experiments on the IEMOCAP benchmark dataset, a widely used SER corpus with limited labeled data. Our results demonstrate that even with small-scale data, leveraging

pre-trained models and speaker-adversarial training yields significant improvements in speaker-independent emotion recognition.

The main contributions of this paper are as follows:

1. Our approach adapts an existing pre-trained speech model to be speaker-invariant by eliminating residual speaker characteristics. 2. We empirically demonstrate the benefits of learning speaker-invariant representations on SER performance, particularly in low-resource settings.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 provides a detailed description of the proposed framework. Section 4 presents the experimental setup and results. Finally, Section 6 concludes the paper.

2 Related works

Self-supervised learning (SSL) has significantly advanced speech processing by enabling the learning of powerful representations from large-scale unlabeled data. Most SSL methods fall into three categories: contrastive learning using InfoNCE loss (van den Oord et al., 2019; Rivière et al., 2020), masked token classification, and reconstruction-based approaches such as predicting future frames or reconstructing masked inputs. Representative models like HuBERT and Wav2Vec 2.0 belong to the second and first categories respectively, and have been shown to effectively capture both acoustic and linguistic information. These models have become essential backbones for various downstream tasks, including speech emotion recognition.

In the context of SER, one of the major challenges is speaker variability, where individual differences in emotional expression reduce model generalizability. Early approaches combined CNN and LSTM architectures (Khan and Kwon, 2021; Mustaqeem and Kwon, 2019), followed by attention-based models that improved performance by focusing on emotionally salient segments. More recently, self-supervised models have shown state-of-the-art performance on SER benchmarks such as IEMOCAP (Tripathi et al., 2019), demonstrating their effectiveness in low-resource emotional modeling.

To further improve generalization, several works have explored feature normalization and domain adversarial learning. Approaches such as confusion loss (Nagrani et al., 2017) and gradient reversal layers (GRL) (Li et al., 2021) aim to reduce reliance on domain-specific or speaker-specific cues. While these methods improve robustness, they often assume that confusing the domain classifier guarantees domain-invariant features—an assumption that does not always hold(Ganin et al., 2016).

Our work is motivated by these findings. We adopt a speaker-adversarial training strategy that builds upon GRL, but further introduces an entropy-based loss to encourage true speaker-invariant representations. Additionally, unlike prior work that applies adversarial training on shallow networks, we apply our approach to pre-trained self-supervised encoders, which significantly boosts the representation power for emotion classification.

3 Method

This section introduces our proposed three-stage training pipeline for learning emotion representations that are robust to speaker-specific variations. The framework consists of three main stages: (1) speaker classifier pretraining, (2) adversarial fine-tuning of a pre-trained speech encoder, and (3) emotion classifier training with the frozen, speaker-invariant encoder.

We denote an input utterance as $x \in \mathbb{R}^T$, where T is the number of waveform samples in a 16kHz mono audio. Each input x is associated with a speaker label $y^{(\text{spk})} \in \{1, \ldots, N\}$ and an emotion label $y^{(\text{emo})} \in \{1, \ldots, C\}$, where N is the number of unique speakers and C is the number of emotion classes.

Given the raw waveform x, the pre-trained encoder f_θ extracts hidden representations:

$$h = f_\theta(x) \in \mathbb{R}^{L \times D},$$

where L is the number of time steps after feature extraction and D is the feature dimension (e.g., $D = 768$ for Wav2Vec2.0 Base). These hidden features are then used in subsequent stages for speaker-adversarial training and emotion classification.

3.1 Stage 1: Speaker Classifier Pretraining

In the first stage, we train a speaker classification model independently to extract speaker-discriminative representations. We use the ECAPA-TDNN model for this purpose, which is widely adopted in speaker verification tasks due to its ability to extract robust speaker embeddings.

Because ECAPA-TDNN expects 80-dimensional mel-spectrogram inputs, we add a projection layer that maps the base model's

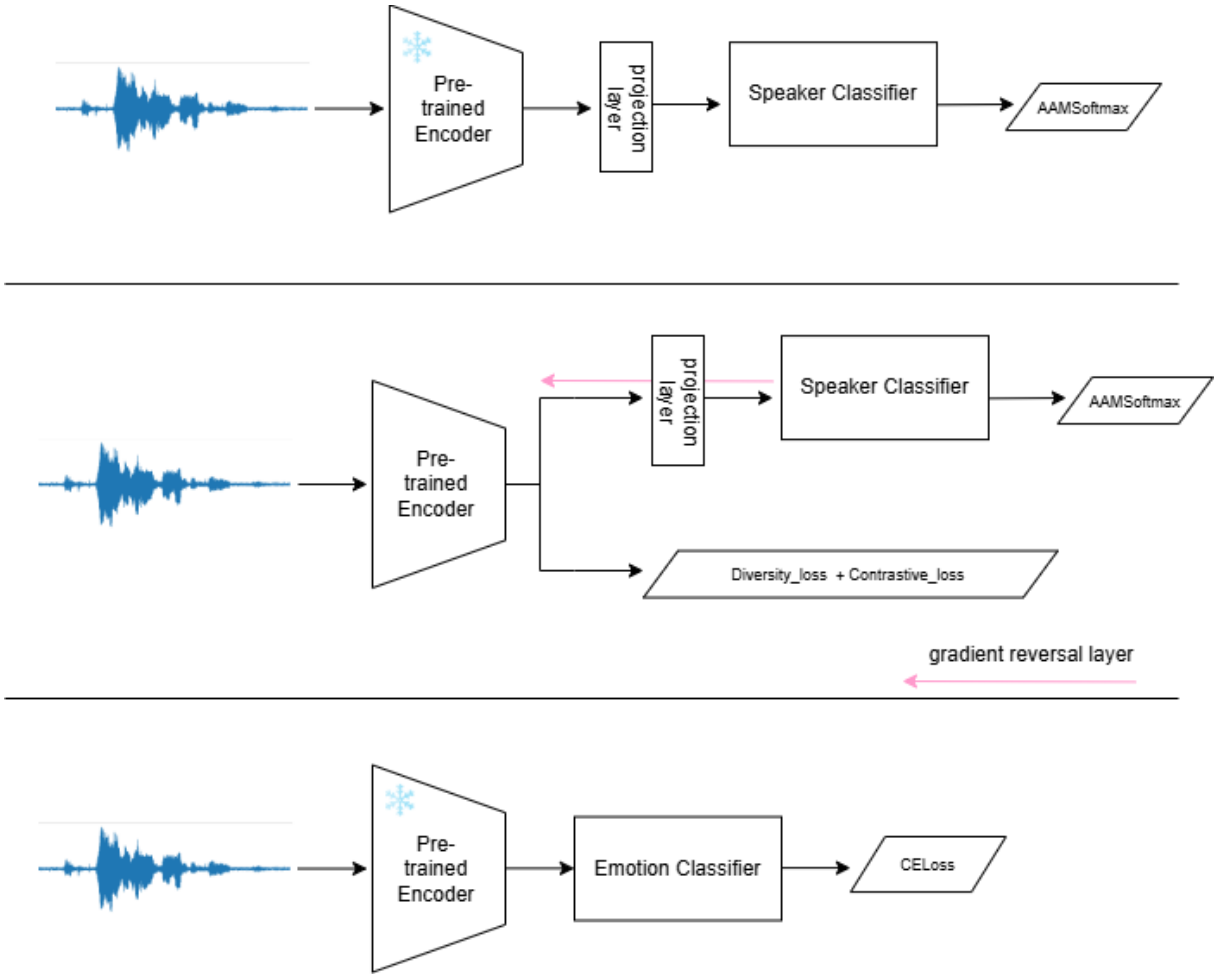

Figure 1: An overview of the proposed three-stage training pipeline for speaker-invariant speech emotion recognition. In **Stage 1**, a speaker classifier is trained on the target dataset to effectively distinguish speaker identity. In **Stage 2**, a pre-trained speech encoder is fine-tuned with only the top two layers unfrozen, using adversarial training to suppress speaker-related information. The pink arrow denotes the Gradient Reversal Layer (GRL), which inverts gradients from the fixed speaker classifier. In **Stage 3**, the output representations from the fine-tuned encoder are used to train an emotion classifier.

768-dimensional output down to 80 dimensions. The model is trained on 16 kHz mono waveforms using the Additive Angular Margin Softmax (AAM-Softmax) loss, and we continue training until it reaches approximately 70% accuracy on the IEMOCAP training split to ensure sufficient speaker classification performance. This trained speaker classifier is then used only in the next stage for adversarial training and is discarded afterward.

3.2 Stage 2: Adversarial Fine-tuning of the Speech Encoder

In this stage, we fine-tune a pre-trained speech encoder—specifically, Wav2Vec 2.0—to learn representations that are invariant to speaker identity while preserving its original self-supervised learning objectives. To this end, we introduce a Gradient Reversal Layer (GRL) between the encoder and the fixed speaker classifier, forming a domain-adversarial learning setup.

The encoder output is passed through two branches: one through the GRL into the speaker classifier to compute an adversarial loss, and the other used internally by the encoder to compute its original self-supervised objectives. Unlike prior work that replaces the encoder's training loss during fine-tuning, we retain both the *contrastive loss* and the *diversity loss* from the original Wav2Vec 2.0 pre-training to maintain representational richness.

The total loss for this stage is given by:

$$\mathcal{L}_{\text{speech}} = \mathcal{L}_{\text{contrastive}} + \lambda_{\text{div}} \cdot \mathcal{L}_{\text{diversity}} + \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}} \quad (1)$$

Here, $\mathcal{L}$contrastive is the InfoNCE-based contrastive loss used in Wav2Vec 2.0, and $\mathcal{L}$diversity encourages codebook usage diversity in quantized representations. The term $\mathcal{L}$adv denotes the adversarial speaker classification loss and is defined as:

$$\mathcal{L}\text{adv} = -\sum_i \text{CE}(C_{\text{spk}}(\text{GRL}(h_i)), y_i^{(\text{spk})}) \quad (2)$$

where h_i is the latent representation of the i-th utterance obtained from the encoder, and C_{spk} is the fixed speaker classifier trained in Stage 1. The coefficients λ_{div} and λ_{adv} control the strength of the diversity and adversarial terms, respectively. After training, the encoder is frozen for the next stage.

3.3 Stage 3: Emotion Classifier Training

In the final stage, we train an emotion classifier on top of the speaker-invariant encoder obtained from Stage 2. The encoder is frozen during this phase, and only the parameters of the classifier are updated.

The emotion classifier is implemented as a lightweight 1D convolutional network designed to process the temporal output representations from the pre-trained speech encoder. The input to the classifier is a sequence of hidden representations with 768 channels, corresponding to the encoder output.

The classifier consists of two convolutional blocks. Each block contains a 1D convolutional layer with kernel size 3 and padding 1, followed by a ReLU activation and a max-pooling layer with kernel size 2 to reduce the temporal resolution by half. The first convolution maps the input from 768 to 128 channels, while the second convolution reduces it further from 128 to 4 channels, corresponding to the number of emotion classes. The resulting output is aggregated and used to predict emotion labels via cross-entropy loss.

$$\mathcal{L}\text{emo} = \sum_i \text{CE}(C\text{emo}(h_i), y_i^{(\text{emo})}) \quad (3)$$

where C_{emo} is the classifier and $y_i^{(\text{emo})}$ is the emotion label for the i-th utterance.

This final step ensures that the learned features are not only invariant to speaker identity but also effective for the downstream task of speech emotion recognition.

4 Experiments

4.1 Dataset

We conduct our experiments using the IEMOCAP dataset, a widely adopted benchmark for speech emotion recognition.

IEMOCAP Dataset We evaluate our proposed method on the IEMOCAP dataset, a widely used benchmark for speech emotion recognition. The IEMOCAP corpus consists of approximately 12 hours of audiovisual recordings collected from 10 professional actors (5 male and 5 female), organized into five sessions. Each session includes a dyadic interaction between one male and one female speaker, engaging in both scripted and improvised dialogues. This structure provides a wide variety of emotionally expressive speech, making it well-suited for emotion recognition research.

In this work, we focus exclusively on the audio modality. We use the categorical emotion annotations provided with the dataset and follow the standard protocol by selecting the four most commonly used emotion classes: *angry*, *happy*, *sad*, and *neutral*. Each dialogue is segmented into utterances, where a single utterance typically corresponds to a single sentence. After filtering for the selected emotion classes, the dataset contains a total of approximately 39,397 labeled utterances. We then split these utterances into training and test sets using an 80:20 ratio to ensure robust evaluation.

To improve model generalization and robustness, we apply speed perturbation as a data augmentation strategy. Specifically, we generate additional training samples by resampling each utterance at speed factors of 0.8, 0.9, 1.1, and 1.2, in addition to the original speed (1.0). This technique effectively increases the size of the training set by a factor of five, while preserving the temporal and spectral structure necessary for emotion recognition (Ko et al., 2015).

4.2 Experimental Setup

Stage 1. The ECAPA-TDNN model was fine-tuned for speaker classification using the Additive Angular Margin Softmax (AAM-Softmax) loss. The model was trained for 50 epochs using a cosine learning rate scheduler. This classifier was subsequently fixed and used in Stage 2 to generate the adversarial signal for suppressing speaker information.

Table 1: Comparison of weighted accuracy (WA) on IEMOCAP using different models and training strategies.

Model	Model Size	Params	Speaker-Invariant	WA (%)
Vanilla (Baseline)	Base	95M	✗	66.3
+ GRL	Base	95M	✓	57.7
+ GRL + Contrastive + Diversity (Ours)	Base	95M	✓	68.4

Table1. Weighted accuracy (WA) comparison on the IEMOCAP dataset (4-class classification). All models use Wav2Vec 2.0 Base as the pre-trained encoder. "GRL" denotes the use of a Gradient Reversal Layer with a fixed speaker classifier. "Contrastive" and "Diversity" refer to the original self-supervised learning objectives of Wav2Vec2.0, which are preserved during fine-tuning to maintain representational capacity. The "Vanilla" model serves as the baseline and is trained with only a cross-entropy loss.

Stage 2. In the second stage, we fine-tuned the Wav2Vec 2.0 Base model using a composite objective comprising the original contrastive loss, diversity loss, and an adversarial speaker classification loss applied via a Gradient Reversal Layer (GRL). The adversarial classifier was the ECAPA-TDNN model trained in Stage 1. The weight for the adversarial loss was set empirically to $\lambda_{\text{adv}} = 0.9$, while the weight for the diversity loss was set to $\lambda_{\text{div}} = 0.01$ which is used as de-fault setting in the wav2vec2.0 config. This stage was trained for 50 epochs with an initial learning rate of 1×10^{-4}, which was halved every 10 epochs. Early stopping is used with patience 50 steps.

Stage 3. The final stage involved training an emotion classifier on top of the frozen encoder obtained from Stage 2. A simple feed-forward network was used for classification, and it was trained for 10 epochs using a fixed learning rate of 1×10^{-4}. Only the classifier parameters were updated during this stage; the encoder remained frozen.

All experiments were implemented using PyTorch and conducted on a single NVIDIA RTX 4090 GPU. We utilized pre-trained models for both the speech encoder and the speaker classifier: the Wav2Vec 2.0 Base model, pre-trained on Librispeech, and ECAPA-TDNN, pre-trained on the VoxCeleb dataset. The Adam optimizer was employed for all training stages. A batch size of 4 was used, and to enable effective training with limited GPU memory, we applied gradient accumulation with 32 steps, resulting in an effective batch size of 128. Early stopping was not used during training.

5 Results

5.1 Main Evaluation on IEMOCAP

Table 1 reports the weighted accuracy (WA) obtained on the IEMOCAP 4-class setup.[1] The vanilla *Wav2Vec 2.0* encoder—kept frozen and followed by a task-specific classifier—yields a baseline of **66.3%** WA. After applying our three-stage pipeline, the *speaker-invariant* encoder achieves **68.4%** WA, an absolute gain of **2.1%** percentage points. Because both systems share the same model size (95 M parameters) and differ only in the fine-tuning strategy, we attribute the gain to explicitly suppressing speaker cues via adversarial learning and the diversity constraint.

5.2 Zero-shot and One-shot Generalisation

Table 1 (Section 7) summarises the encoder's transferability for zero-shot or one-shot condition:

- **Zero-shot** (classifier only): our speaker-invariant encoder attains **56.63%** WA, more than double the vanilla baseline (**25.85%**). This confirms that disentangling speaker information leads to representations that are inherently more emotion-focused.
- **One-shot**: With a single training for the emotion classifier, performance rises to **58.88%**, indicating strong few-shot adaptability.

5.3 Impact of Adversarial Components

An ablation in Table 1 further shows that including both the *contrastive* and *diversity* losses during adversarial fine-tuning contributes an additional about 10 percentage-point gain over using the GRL alone.

Notably, fine-tuning with the GRL alone leads to a drop in performance (57.7% WA), since the

[1] All results are averaged over five speaker-independent folds, following the standard protocol.

encoder is updated only in the opposite direction of the speaker classifier without preserving its original self-supervised objectives. This suggests that retaining the self-supervised objectives helps preserve the encoder's representational richness while still eliminating speaker-specific information.

5.4 Discussion

Overall, the results validate our central hypothesis: *speaker variability is a dominant confounder in SER, and explicitly neutralising it yields more robust, transferable emotion features*. Notably, the improvements are achieved without enlarging the network or relying on extra emotion labels, highlighting the practicality of the proposed pipeline for low-resource scenarios.

6 Conclusion

We introduced a three-stage adversarial training framework for learning speaker-invariant representations in speech emotion recognition (SER). By leveraging a fixed ECAPA-TDNN speaker classifier and applying gradient reversal to a pre-trained Wav2Vec2.0 encoder, our method effectively suppresses speaker-specific information while preserving emotion-discriminative features.

Empirical results on the IEMOCAP benchmark demonstrated that our speaker-invariant representations significantly improve both in-domain performance and zero-shot generalization. Specifically, we observed over 30% improvement in zero-shot weighted accuracy when comparing the speaker-invariant encoder to a vanilla Wav2Vec2.0 encoder with only a trained classifier. These findings confirm that speaker variability is a major limiting factor in SER, and that explicitly removing speaker cues leads to more robust, transferable emotion features.

Importantly, our framework operates without the need for large-scale emotion annotations or task-specific architecture changes. Once the encoder is adversarially fine-tuned, it can be applied to new tasks such as emotion classification, with minimal data and without further encoder updates. This opens a promising direction toward building general-purpose, speaker-agnostic acoustic encoders for a wide range of paralinguistic tasks.

In future work, we plan to evaluate our speaker-invariant encoder on tasks beyond emotion recognition, such as intent detection or conversational analysis, and to investigate the trade-offs between speaker suppression and retention of speaker-dependent emotional nuance. We also aim to extend our approach to multilingual and code-switched speech, where speaker and language cues are often entangled.

Table 2: Comparison of Weighted Accuracy of the Proposed Method under Zero-Shot and One-Shot Settings

Encoder	Tuning Strategy	WA (%)
Zero-shot		
Wav2Vec2.0 (Vanilla)	None	25.85
Wav2Vec2.0 (Speaker-Invariant)	Adversarial (GRL)	56.63
One-shot		
Wav2Vec2.0 (Speaker-Invariant)	Adversarial (GRL)	58.88

Limitations

While our approach demonstrates the effectiveness of speaker-invariant representations for speech emotion recognition, there are several limitations to be acknowledged.

First, our experiments were conducted solely on the IEMOCAP dataset, which is relatively small and limited to acted emotional expressions in English. As such, the generalizability of our method to more diverse, spontaneous, and multilingual emotion corpora remains to be verified.

Second, although we demonstrated that suppressing speaker information improves zero-shot performance, we did not explicitly measure the trade-off between speaker invariance and potential loss of paralinguistic cues (e.g., speaker identity, personality) that may contribute to emotion perception in natural scenarios.

Third, we focused on emotion classification as the target downstream task. It remains an open question whether the learned speaker-invariant encoder also benefits other paralinguistic tasks such as sentiment analysis, intent detection, or speaker trait recognition.

Finally, while our method improves performance without requiring large-scale emotion labels, it still depends on a speaker-labeled dataset to train the adversarial speaker classifier. Exploring unsupervised or weakly supervised alternatives for speaker disentanglement would be a valuable future direction.

7 Ablation

Interestingly, the speaker-invariant encoder significantly outperforms the vanilla Wav2Vec2.0 model in the zero-shot setting, achieving 56.63% weighted accuracy compared to only 25.85%. This substantial improvement suggests that removing speaker-specific information from the representation leads to more generalizable emotion features, even when the encoder is not explicitly fine-tuned for emotion recognition.

Note that in the zero-shot setting the vanilla Wav2Vec2.0 baseline has never seen any IEMOCAP data, whereas our "speaker-invariant" encoder, although not trained for emotion classification, has been adversarially trained on IEMOCAP speakers during stage (2). This exposure to IEMOCAP speaker distributions naturally helps the encoder learn more robust representations, which partly explains the performance gap in the zero-shot evaluation.

Moreover, in the one-shot setting—where the emotion classifier is trained exactly once on the full IEMOCAP training dataset—the speaker-invariant encoder further improves performance to 58.88% WA. This demonstrates its strong adaptability under minimal supervision. These findings highlight the benefit of adversarial speaker disentanglement not only for in-domain emotion classification, but also for low-resource and unseen-domain scenarios.

References

Shahin Amiriparian, Filip Packań, Maurice Gerczuk, and Björn W. Schuller. 2024. Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets. *Preprint*, arXiv:2406.10275.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *Preprint*, arXiv:1904.03240.

Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *Preprint*, arXiv:2009.09796.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA$.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Preprint*, arXiv:1505.07818.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Zijiang Huang, Tejas Mistry, and Kazuhiko Koishida. 2021. Speech emotion recognition using augmentation with speed perturbation. In *Proc. ICASSP*, pages 6344–6348.

Mustaqeem Khan and Soonil Kwon. 2021. 1d-cnn: Speech emotion recognition system using a stacked network with dilated cnn features. *Computers, Materials Continua*, 67:4039–4059.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou. 2021. Speaker-invariant affective representation learning via adversarial training. *Preprint*, arXiv:1911.01533.

n Mustaqeem and Soonil Kwon. 2019. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, interspeech$_2$017.$ISCA$.

Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. 2019. Learning problem-agnostic speech representations from multiple self-supervised tasks. *Preprint*, arXiv:1904.03416.

Thejan Rajapakshe, Rajib Rana, Farina Riaz, Sara Khalifa, and Björn W. Schuller. 2025. Representation learning with parameterised quantum circuits for advancing speech emotion recognition. *Preprint*, arXiv:2501.12050.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. *Preprint*, arXiv:2002.02848.

Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. 2019. Multi-modal emotion recognition on iemocap dataset using deep learning. *Preprint*, arXiv:1804.05788.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ni Wang and Danyu Yang. 2025. Speech emotion recognition using fine-tuned wav2vec2.0 and neural controlled differential equations classifier. *PLOS ONE*, 20(2):1–13.

Puming Zhan and M. Westphal. 1997. Speaker normalization based on frequency warping. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042 vol.2.