INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Pre-Finetuning for Few-Shot Emotional Speech Recognition

Maximillian Chen, Zhou Yu

Columbia University

maxchen@cs.columbia.edu, zy2461@columbia.edu

Abstract

Speech models have long been known to overfit individual speakers for many classification tasks. This leads to poor generalization in settings where the speakers are out-of-domain or out-of-distribution, as is common in production environments. We view speaker adaptation as a few-shot learning problem and propose investigating transfer learning approaches inspired by recent success with pre-trained models in natural language tasks. We propose pre-finetuning speech models on difficult tasks to distill knowledge into few-shot downstream classification objectives. We pre-finetune Wav2Vec2.0 on every permutation of four multiclass emotional speech recognition corpora and evaluate our pre-finetuned models through 33,600 few-shot fine-tuning trials on the Emotional Speech Dataset.

Index Terms: emotion recognition, low-resource learning, pre-finetuning, transfer learning

1. Introduction

Speech models tend to generalize poorly to out-of-distribution speakers due to a phenomenon called speaker overfitting [1, 2, 3]. Speaker overfitting can be problematic when deploying systems in production environments. There, speakers typically do not exist in training corpora, and it requires time to amass sufficient amounts of training data. This motivates systems which can adapt to individual speakers "on-the-fly" with little data.

To this end, out-of-domain speaker adaptation can be viewed as a few-shot learning problem [3]. In low-resource data settings, learning can be difficult, particularly with the regularization used in typical speaker-invariant learning approaches [3, 4]. However, in recent years, many approaches to few-shot learning have found success with transfer learning, leveraging pre-trained models which have learned representations from multiple corpora [5, 6, 7]. But, these pretraining corpora are not necessarily relevant to target downstream tasks. This has motivated recent work to propose an additional step between pre-training and downstream fine-tuning called pre-finetuning [5]. While most work has examined pre-finetuning with multi-task learning, [8] found that large-scale multi-task pre-finetuning performance can nearly be matched by only using corpora from tasks that match the downstream task type.

Overall, pre-finetuning has been studied extensively and successfully in natural language processing tasks, but our study is the first to attempt pre-finetuning for any speech or audio processing task. We reason this is due to the comparatively higher number of tasks and datasets with accessible licensing for natural language. Additionally, many language tasks are easily cross-compatible during transfer learning (e.g. every pre-training task for T5 is trained using textual inputs and outputs without needing to adapt the model or loss function [9]). Here, we consider the few-shot[1] emotional speech recognition task. We pre-finetune a separate model using every member of the power set of four large emotional speech corpora according to the workflow in Figure 1.[2] We evaluate our pre-finetuned models across 33,600 controlled few-shot classification experiments. We contribute ablations and analyses into how different experimental conditions affect pre-finetuning efficacy.

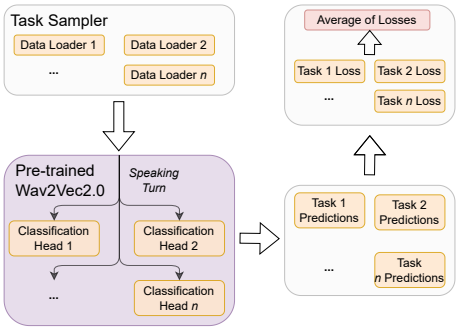

Figure 1: *Workflow of pre-finetuning an emotion recognition model. Wav2Vec2.0 is initialized with a separate linear classification head for each pre-finetuning dataset in order to ensure the correct output space. Pre-finetuning tasks are continuously randomly sampled, and each instance is mapped to the corresponding classification head. Each task's loss is computed separately and averaged during validation.*

2. Related Work

Pre-trained models have been a successful case study of transfer learning for low-resource tasks. When scaled to billions of parameters, pre-trained models have been able to generalize pre-trained knowledge to downstream tasks through few-shot in-context learning, taking the form of both downstream task models (e.g. [7, 10, 11, 12]) and models for synthetic data generation (e.g. [13, 14, 15, 16]) in natural language tasks. However, capabilities such as in-context learning typically only appear through this tradeoff with model size, as they only emerge when using sufficiently large models. To date, pre-trained speech models have rarely been scaled to the same extent. The largest model is BigSSL with eight billion parameters [17], whereas recent work on in-context learning for low-resource language data augmentation has used models with a *minimum* size of six billion [13, 18, 16], and often reaching model sizes as large as 175 billion parameters [18]. Smaller pre-trained models lack

[1] We use as few as two downstream training examples.

[2] https://github.com/maxlchen/Speech-PreFinetuning

10.21437/Interspeech.2023-136

the same expansive pre-trained representations found in large models, making it impractical to perform in-context learning or generate high-quality synthetic data. However, due to their smaller size, it is more feasible to perform fine-tuning to directly transfer knowledge to downstream tasks.

But, fine-tuning still is impractical when there is not a sufficient amount of data. Recent work proposed an additional learning step between pre-training and fine-tuning to better facilitate knowledge distillation: pre-finetuning [5]. They pre-finetuned the base and large variants of BART [19] and RoBERTa [20] using large-scale multitask corpora. Follow-up work found that it was more efficient to do single-task pre-finetuning than to have a diverse set of corpora [8]. With the right corpora selection, pre-finetuning has potential to help with few-shot learning problems [21] such as speaker adaptation, but to date it has yet to be explored in any speech processing task.

Traditional approaches to speaker adaptation focus on invariance using adversarial learning or regularization. [22] proposed a network to jointly learn a speaker classifier and senone discriminator through adversarial multitask learning. Several works have investigated using KL divergence-based regularization (e.g. [23, 24]). Other studies used gradient reversal layers to regularize learning gains from individual speakers [25]. But, these approaches often still require large amounts of training data and are not as applicable to low-resource settings [22, 4].

The overall success with using pre-trained models in few-shot learning and increasing popularity of pre-trained speech models such as Wav2Vec2.0 [26] and HuBERT [27] naturally motivates the exploration of pre-finetuning for few-shot speaker adaptation. The most similar lines of work examine multi-task learning for speech processing [28, 29, 30], which also involves learning representations from multiple data sources. However, the key difference compared to our setting is that classic multitask learning involves training a model to learn a representation shared between a target downstream task and any auxiliary tasks *simultaneously* [31]. This requires sufficient downstream data. Pre-finetuning is a form of multitask learning which instead takes place during an intermediate step dedicated to learning an auxiliary task representation, which in turn can be used as a close initialization for a low-resource downstream task. Our work is the first to examine pre-finetuning speech models. We ground our study in the context of few-shot speaker adaptation for emotional speech recognition, and draw upon findings from [8] in our selection of pre-finetuning corpora.

3. Methodology

3.1. Corpora Selection

In this study, we focus on adapting speakers for emotion recognition as our downstream task. As such, we chose four large, diverse pre-finetuning corpora that each fall within the category of emotional speech recognition. MSP-IMPROV contains 8,438 improvised speaking turns from four emotions (happy, sadness, anger, neutral) [32]. MSP-PODCAST contains 100 hours of speech from 62,140 from speaking turns collected from podcast recordings [33]. Each turn is annotated with one of nine categorical emotion labels (anger, happiness, sadness, disgust, surprised, fear, contempt, neutral, other). The Mandarin Affective Speech (Mandarin AS) corpus contains 25,636 utterances from 68 unique speakers, with annotations according to five emotion labels (anger, elation, neutral, panic, sadness) [34, 35]. The IEMOCAP benchmark contains 12 hours of audio consisting of 10,039 total turns from ten unique speakers, with nine different emotion labels (anger, happiness, excitement, sadness, frustration, fear, surprise, other, neutral). All corpora use English speech other than Mandarin AS, which uses Mandarin.

3.2. Pre-Finetuning in Speech

In this work, we pre-finetune base Wav2Vec2.0 (94.4M parameters), inspired the workflow used by [8] with language models. As depicted in Figure 1, we initialize one linear classification head for each of our pre-finetuning tasks, appropriately setting the output space. For each training step, we load an instance from a randomly sampled pre-finetuning task and map the loaded instance to the corresponding classification head. We compute a scaled loss for each task separately, as in [5, 8]. The scaled loss is given as $\frac{L_i(x_i,y_i,\theta)}{ln\ n(i)}$ for a model parameterized by θ, where $L_i(x_i, y_i, \theta)$ is the loss for training instance i, and $n(i)$ is the size of the output space for the prediction task of instance i [5]. During validation, we average the taskwise losses, which helps prevents the model from overfitting to individual tasks during pre-finetuning. We pre-finetune for up to 200 epochs with early stopping after three epochs without improvement.

We pre-finetuned each model j on one of the combinations of corpora from the power set[3] of the corpora in Section 3.1. That is, each combination is a set of corpora C_j where $0 \leq |C_j| \leq 4$ and $j \in \{1..16\}$. This includes a base Wav2Vec2.0 model without any pre-finetuning as a baseline (i.e., when $|C_j| = 0$). We additionally attempted to use standard emotion recognition baselines [36] such as the ComParE 2016 automatic paralinguistic feature extractor [37] with a dense artificial neural network, but these approaches cannot surpass the performance of constant prediction. This highlights the difficulty of the few-shot version of this task. All experiments are run using individual NVIDIA RTX A6000 GPUs. We used HuggingFace Transformers [38] and PyTorch [39].

3.3. Downstream Finetuning

To model speaker adaptation as few-shot learning, we perform downstream finetuning on emotion recognition for each individual speaker in the Emotional Speech Dataset (ESD; [40]). There are 10 English speakers and 10 Mandarin speakers. Each study participant was a native speaker of their respective language. Each unique lexical utterance is spoken with five different emotions: Happy, Sad, Surprised, Angry, and Neutral. We perform binary emotion classification[4] for each speaker under seven few-shot settings, $k \in \{2, 4, 8, 16, 24, 32, 64\}$, where k is the number of available training examples. For each few-shot setting, we randomly sample $\frac{k}{2}$ positive and $\frac{k}{2}$ negative training instances. We conduct three trials for each downstream condition. Each trial runs for up to 200 epochs, with 30 epochs of early stopping patience.

4. Experimental Results

Accounting for the 16 combinations of pre-finetuned models and all fine-tuning conditions, we evaluated the effects of pre-finetuning across 33,600 downstream model fine-tuning trials.

[3] We use $1 + \sum_{r=1}^{4} \binom{4}{r} = 16$ different models downstream.

[4] The targets are whether or not an instance matches a specific emotion, e.g. Happy/Not Happy, Sad/Not Sad, etc.

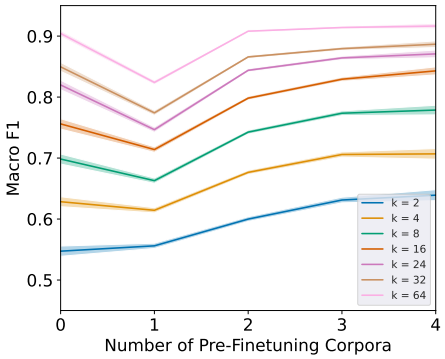

Figure 2: *Comparison of downstream task performance of models pre-finetuned on varying numbers of corpora. Each line depicts change in mean and standard error of F1 Macro.*

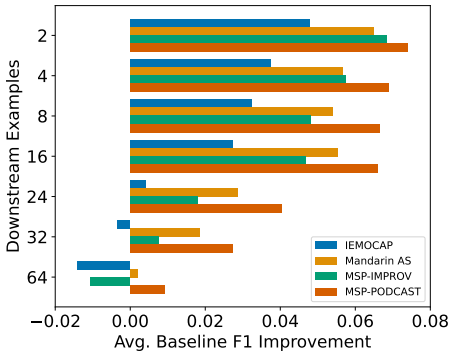

Figure 3: *Average difference in Macro F1 resulting from pre-finetuning on each corpus compared to the Wav2Vec2.0 baseline. Differences shown are aggregations controlling for the number of few-shot examples, each speaker, and each emotion.*

4.1. Effect of Number of Pre-Finetuning Corpora

In Figure 2 we demonstrated the effect of varying the number of corpora (n) used during pre-finetuning in the low-resource setting. Each line represents one few-shot condition (k). Each point represents the average test set F1 score of pre-finetuned models for a particular pre-finetuning corpus size for all downstream classification tasks under few-shot condition k. The shading around each line represents the region of standard error surrounding the mean. We observe a few patterns. In six of the seven few-shot settings, using only one pre-finetuning corpus may hurt performance compared to direct fine-tuning without a pre-finetuning step, which is consistent with the "critical point" for pre-finetuning utility, as discussed in [5]. However, in the most extreme case with $k = 2$, even using just one pre-finetuning corpus still yields performance improvements over the baseline. *After* $n = 1$, we witness continuous improvements in average performance as n increases. However we typically see the largest improvements from $n = 1$ to $n = 2$. We examine possible reasons by ablating the pre-finetuning corpora.

4.2. Ablation on Individual Corpus Contributions

We attempt to quantify how much each individual pre-finetuning corpus contributes to changes in downstream classification performance. We controlled for each speaker, each emotion, and each few-shot condition for fair comparisons. Then, under each of these controlled settings, we compute the average F1 score across all model trials for which pre-finetuning set C_j includes each corpus c, subtracted by the average performance of the no-prefinetuning baseline. For each of these controlled settings, we calculated the change in average F1 scores for each pre-finetuning corpus compared to baseline Wav2Vec2.0.

Figure 3 illustrates these differentials aggregated across speakers and emotions, and stratified by each few-shot setting. We consistently see that models pre-finetuned on combinations of corpora which include MSP-PODCAST result in the most improvements over baseline performance on average. In contrast, IEMOCAP results in the least improvements and hurts performance on average in the $k = 32$ and $k = 64$ few-shot data settings. This applies to MSP-IMPROV in one setting, but MSP-IMPROV is also the second most useful pre-finetuning corpus in the two most extreme few-shot data settings. The fact that certain corpora hurt performance overall may explain why there tends to be a large improvement in performance from $n = 1$ to $n = 2$.

4.3. Ablation on Pre-Finetuning Corpus Inclusion

We further attempted to isolate the contributions of individual corpora by examining the effect of including and excluding individual corpora in the extreme few-shot settings. That is, for each individual pre-finetuning corpus c from the set of all corpora C, we compared models pre-finetuned on $C_j = C \setminus \{c\}$ and $C_j = \{c\}$. As in Section 4.2, we aggregated performances controlled for speakers, emotions, and few-shot settings, to ensure fair direct comparisons.

In Table 1[5], we see patterns mostly consistent with the analysis of corpora contributions in Figure 3. $F1_{in}$ is the aggregated performance of the model pre-finetuned on $C_j = \{c\}$ and $F1_{ex}$ is the performance of the model pre-finetuned on $C_j = C \setminus \{c\}$. MSP-PODCAST is the only corpus which has any positive differentials when compared to pre-finetuning on all other corpora. For all few-shot settings, MSP-PODCAST has the highest differential, followed by Mandarin AS, MSP-IMPROV, and IEMOCAP.

4.4. Scaling Downstream Training Data Sizes

From Figure 2 and Figure 3, it is clear that the most performance improvements arise from the fewest-resource setting ($k = 2$ downstream training examples), but we do see performance improvements in higher resource settings when pre-finetuning on all four corpora. We take a closer look into these performance improvements to further understand the upper bounds and limits of this pre-finetuning approach. In Figure 4, we compare the performance of the full pre-finetuning setting against the no pre-finetuning baseline in all of our data settings, stratified by individual emotions and separated by speakers' native language.

For both the English and Mandarin speakers, we see that using pre-finetuning yields higher performance on average than the baseline for all few-shot data settings for all emotions except for "Surprise." This may be due to the fact that while the other four emotions are relatively common, Surprise is not as well-represented in the pre-finetuning corpora. Surprise only appears in MSP-PODCAST and IEMOCAP. Overall, we see pre-finetuning can boost classification performance substantially, but performance gains generally taper off after $k = 24$.

[5]We see the same patterns at $k \in \{16, 24, 32, 64\}$. We omit these results due to space constraints. $F1_{in}$ for MSP-PODCAST at $k = 64$ is 0.9126 whereas $F1_{ex}$ is 0.9108.

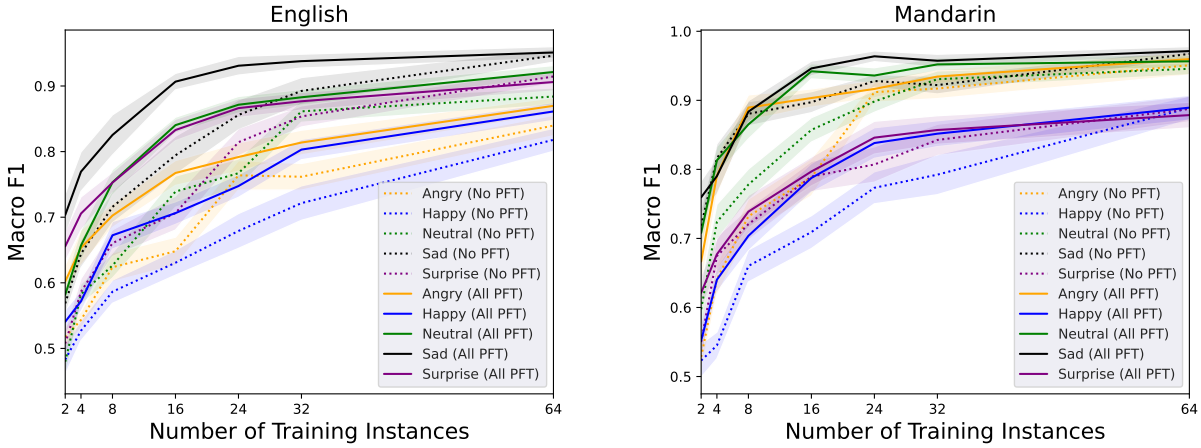

Figure 4: *Effect of number of training examples using during fine-tuning for the baseline model with no pre-finetuning (No PFT), and the model pre-finetuned on all four corpora (All PFT). Results are stratified by emotion. Left: classification results on native English speech. Right: classification results on native Mandarin speech.*

Table 1: *Average Macro F1 from pre-finetuning on each individual corpus ($F1_{in}$) compared to F1 from pre-finetuning on all but that corpus ($F1_{ex}$). F1 is aggregated controlling for the number of few-shot examples, each speaker, and each emotion.*

k	Corpus	$F1_{in}$	$F1_{ex}$	Δ
2	IEMOCAP	0.5048	0.6232	-0.1184
	Mandarin AS	0.5812	0.6445	-0.0633
	MSP-IMPROV	0.5233	0.6299	-0.1066
	MSP-PODCAST	0.6150	0.6272	**-0.0122**
4	IEMOCAP	0.5447	0.7054	-0.1607
	Mandarin AS	0.6495	0.7160	-0.0665
	MSP-IMPROV	0.5623	0.7030	-0.1407
	MSP-PODCAST	0.7010	0.6990	**0.0020**
8	IEMOCAP	0.5816	0.7747	-0.1931
	Mandarin AS	0.7040	0.7862	-0.0822
	MSP-IMPROV	0.6021	0.7783	-0.1762
	MSP-PODCAST	0.7640	0.7549	**0.0091**

5. Discussion

While conventional wisdom suggests that pre-finetuning must be performed on large-scale corpora [5], we show that it is possible to achieve strong results in the few-shot setting with small-scale pre-finetuning. We observe that downstream task performance may improve further as more pre-finetuning corpora are used. In this study we only used four corpora due to licensing and computational constraints, but our findings warrant examining the benefits of adding more corpora. A larger set may delay the onset of the diminishing returns seen in Figure 4.

We also find evidence of a critical point in number of pre-finetuning corpora prior to witnessing performance improvements in speech tasks, as previously shown for text tasks [5]. We hypothesize that this is likely because models such as Wav2Vec2.0 adapt well to downstream fine-tuning due to general representations learned during their masked pre-training process [26], while pre-finetuning on too few corpora may cause such models to lose some of their generality. We also see that MSP-PODCAST appears to contribute the most to improvements in downstream task performance. This may be due to it being the largest corpus, in terms of number of training instances and number of emotions covered, despite averaging each task's loss during pre-finetuning.

Our experimental results reveal that the fewer the number of available downstream training examples, the more valuable pre-finetuned representations are. The $k = 2$ setting is the only context in which there is not a number of pre-finetuning corpora which results in a performance decrease on average compared to Wav2Vec2.0 without pre-finetuning (Figure 2). Moreover, Figure 2 and Figure 3 shows that the performance improvements over the baseline generally seem smaller with a greater number of available training examples, indicating that pre-finetuning helps performance the most in the extreme low-resource settings, which reflects the initial stages of personalized speaker adaptation. This approach to speaker adaptation is also unconstrained by language. Our pre-finetuned models adapt to each speaker regardless of whether they speak English or Mandarin.

In this study, we held the model choice fixed, but our workflow is compatible with any pre-trained model. Base Wav2Vec2.0 is comparable in size to the base variants of RoBERTa and BART, so it is likely that using a larger Wav2Vec2.0 would result in improvements following similar patterns to those language models. While a benefit of pre-finetuning is that we can achieve strong performance using a small model and limited downstream data, we would likely see further performance improvements using larger models.

6. Conclusion

This work is the first to examine pre-finetuning on speech processing tasks. We see large performance improvements in extreme few-shot data settings (e.g. with $k = 2$ training examples). We contribute an in-depth controlled analysis of several experimental factors including ablations of pre-finetuning corpora, motivating applying pre-finetuning to other speech processing tasks with more models and pre-finetuning corpora.

7. Acknowledgements

This work is supported by a DARPA PTG grant. We thank Ta-Chung Chi, Vishakh Padmakumar, and Debasmita Bhattacharya for their helpful feedback and advice.

8. References

[1] J.-w. Jung *et al.*, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," in *Interspeech*, 2018.

[2] G. Pironkov *et al.*, "Speaker-aware long short-term memory multi-task learning for speech recognition," in *EUSIPCO*. IEEE, 2016, pp. 1911–1915.

[3] T. Wang *et al.*, "Spoken content and voice factorization for few-shot speaker adaptation." in *INTERSPEECH*, 2020, pp. 796–800.

[4] J. Zhao *et al.*, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[5] A. Aghajanyan *et al.*, "Muppet: Massive multi-task representations with pre-finetuning," in *EMNLP*, Nov. 2021, pp. 5799–5811.

[6] T. Gao *et al.*, "Making pre-trained language models better few-shot learners," in *ACL-IJCNLP*, 2021, pp. 3816–3830.

[7] T. Brown *et al.*, "Language models are few-shot learners," *NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[8] V. Padmakumar *et al.*, "Exploring the role of task transferability in large-scale multi-task learning," in *NAACL 2022*, Jul. 2022, pp. 2542–2550.

[9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[10] P. Liu *et al.*, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[11] S. Min *et al.*, "Rethinking the role of demonstrations: What makes in-context learning work?" in *EMNLP*, 2022.

[12] J. Wei *et al.*, "Emergent abilities of large language models," *TMLR*, 2022.

[13] M. Chen *et al.*, "Weakly supervised data augmentation through prompting for dialogue understanding," in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

[14] J. Liu *et al.*, "Low-resource ner by data augmentation with prompting," in *IJCAI*, 2022, pp. 4252–4258.

[15] Y. Meng *et al.*, "Generating training data with language models: Towards zero-shot language understanding," in *NeurIPS*, 2022.

[16] M. Chen *et al.*, "PLACES: Prompting language models for social conversation synthesis," in *Findings of EACL*, 2023.

[17] Y. Zhang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1519–1532, 2022.

[18] G. Sahu *et al.*, "Data augmentation for intent classification with off-the-shelf large language models," in *4th Workshop on NLP for Conversational AI*, 2022, pp. 47–57.

[19] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, Jul. 2020, pp. 7871–7880.

[20] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[21] J. Ma *et al.*, "Label semantics for few shot named entity recognition," in *Findings of ACL*, 2022, pp. 1956–1971.

[22] Z. Meng *et al.*, "Speaker-invariant training via adversarial learning," in *ICASSP*. IEEE, 2018, pp. 5969–5973.

[23] M. Kim *et al.*, "Regularized speaker adaptation of kl-hmm for dysarthric speech recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1581–1591, 2017.

[24] C. Liu *et al.*, "Investigations on speaker adaptation of lstm rnn models for speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5020–5024.

[25] Y. Yin *et al.*, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *ICMI*, 2020, pp. 481–490.

[26] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[27] W.-N. Hsu *et al.*, "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.

[28] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 7, pp. 1172–1183, 2015.

[29] X. Cai *et al.*, "Speech emotion recognition with multi-task learning." in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

[30] Q. Meeus *et al.*, "Multitask learning for low resource spoken language understanding," 2022-09-18.

[31] G. Pironkov *et al.*, "Multi-task learning for speech recognition: an overview." in *ESANN*, 2016.

[32] C. Busso *et al.*, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[33] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[34] Y. Yang *et al.*, "Mandarin affective speech," *LDC2007S09*, 2007.

[35] T. Wu *et al.*, "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition," in *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, 2006, pp. 1–5.

[36] J. Zhang *et al.*, "Multimodal deception detection using automatically extracted acoustic, visual, and lexical features." in *Interspeech*. ISCA, 2020.

[37] B. Schuller *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech*, vol. 8. ISCA, 2016, pp. 2001–2005.

[38] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *EMNLP*, 2020, pp. 38–45.

[39] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[40] K. Zhou *et al.*, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP*. IEEE, 2021, pp. 920–924.