INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Episodic Memory For Domain-Adaptable, Robust Speech Emotion Recognition

James Tavernor[1], *Matthew Perez*[1], *Emily Mower Provost*[1]

[1]University of Michigan, United States

tavernor@umich.edu, mkperez@umich.edu, emilykmp@umich.edu

Abstract

Emotion conveys abundant information that can improve the user experience of various automated systems, in addition to communicating information important for managing well-being. Human speech conveys emotion, but speech emotion recognition models do not perform well in unseen environments. This limits the ubiquitous use of speech emotion recognition models. In this paper, we investigate how a model can be adapted to unseen environments without forgetting previously learned environments. We show that memory-based methods maintain performance on previously seen environments while still being able to adapt to new environments. These methods enable continual training of speech emotion recognition models following deployment while retaining previous knowledge, working towards a more general, adaptable, acoustic model.

Index Terms: Speech Emotion Recognition, Episodic Memory, Continual Learning

1. Introduction

Human emotion plays a large role in communication. Consequently, speech emotion recognition (SER) is a critical tool, informing intelligent systems about the feelings, attitudes, and desires of a given user [1, 2]. With the rise of ubiquitous computing, SER can be widely deployed to edge devices such as smartphones, home assistants, and robots. However, to perform as intended, these SER systems must be robust and adaptable to various acoustic environments (i.e., smartphones on the go). This is a challenging task, as previous SER work has shown that there are many contributing factors (i.e., recording condition, label schema, rater bias, etc. [3,4]), which make it difficult for models to generalize well across emotion datasets, let alone to real-world data. Additionally, collecting new data from various end-users to retrain more robust models on a central server is challenging given the sensitive and personal nature of speech and concerns involving centralized storage. With this in mind, we argue that learning methods that can adapt to an end-user's environment is desirable. However, there is an open question of how to prioritize generalizable emotion recognition learning as the model is exposed to multiple acoustic environments. We investigate how continual learning approaches can contribute to SER training to enable generalization and adaptation to new domains without sacrificing performance on previous domains and without requiring access to all previous data or centralized storage of new data.

Consider the following example of a user and their smartphone (SER system). This person will use their phone in various acoustic settings, whether at home, work, or in a new location like a coffee shop. If the user's phone works well at home, but they then spend more time at work, the model should adapt to the work environment to improve in this setting. However, we need to ensure that as we improve on the workplace acoustic environment, the model does not lose its general ability in other acoustic environments. In order to perform as intended, these SER systems must be scalable and adaptable to a variety of acoustic environments (i.e., smartphones on the go) but also preserve performance on previous environments. We represent this problem by learning a model across different SER datasets to approximate varied acoustic environments.

A common approach to adapting models to new acoustic environments is finetuning to data from the new domain. However, finetuning models on new data often results in models forgetting previously seen data [5, 6]. As a result, simply finetuning end-user models with new speech data in a single environment (such as in the workplace) could result in lost performance in other previously seen environments. Research in federated learning also aims to adapt to end-user data on the client device and collate the resulting gradients to produce a more general server model [7]. However, in this work we investigate a different question: can these models be locally tuned for an end-user's environments? Memory-based learning methods, where select samples are stored in memory and re-used when learning a new task, offer a possible solution. They can address finetuning challenges and single device updates [6, 8, 9]. However, we do not know how these methods will work focusing on emotion in varying acoustic environments.

In this work, we present a new framework for learning SER models for real-world systems using memory-based methods for newly encountered domains (i.e., datasets). We investigate the following research questions:

- Can memory-based methods encapsulate varying acoustic environments while preserving properties of emotion?
- Does memory sampling improve these memory-based methods' ability to overcome forgetting?

To this end, we explore storing episodic memory of previously seen samples using frameworks based on memory experience replay (ExpeR) and Gradient Episodic Memory (GEM). We compare these methods against a standard pretrain and finetune approach, and show significantly improved performance on previous and unseen data. We focus on activation as we investigate varying acoustic environments. Activation classifiers often rely upon acoustic input, while valence is more oriented to text [10]. Further, retaining valence performance when training with data that is partially scripted and less natural may not be as valuable in practice. However, we use valence as regularization by predicting activation and valence in a multitask context [11]. We demonstrate the ability of episodic memory methods to learn patterns relevant to newer training sets while retaining information from prior training sets without requiring

10.21437/Interspeech.2023-2111

access to entire datasets or centralized storage of newer data.

2. Related works

The goal of continual learning is for a model to learn over time while retaining the ability to perform well on tasks it learned in the past, avoiding catastrophic forgetting. One method of addressing the forgetting previous tasks is to replay a small subset of experiences from past tasks. This is often referred to as Experience Replay, and it is an effective method for remembering previous tasks [6, 8]. Experience replay methods have been investigated for speech recognition, most relevantly focused on two Dutch language dialects [12]. However, they have not yet been investigated for emotion recognition or varying acoustic environments. Since our approach makes use of a single task in different domains, we hypothesize that replay methods are best suited for this approximating training across multiple domains.

GEM is another framework designed for continual learning [9]. Instead of replaying the previously seen samples, GEM iterates over the memory samples calculating the loss and model weight gradients during each training step. GEM compares the calculated gradients for each previous task to the current batch's gradient. GEM will prevent an update if the gradients conflict and will solve a quadratic programming problem to find an update as close as possible to the current batch's gradient such that it no longer conflicts with past task gradients. The advantage of GEM is that it ensures no overfitting to memory samples while retaining strong memory of the previous tasks. GEM has been used for automatic speech recognition to avoid retraining on complete datasets when new data is introduced to improve total training time [13]. This method looks at growing data of the same domain that supersede one another. It is still unknown if these methods can handle various acoustic domains or preserve emotion, as opposed to speech properties.

3. Experiments

3.1. Data setup

We use three datasets to investigate continual learning through different acoustic environments for SER (Table 1). We focus on adapting models from one dataset environment to the next, investigating the model's performance on previously seen, currently used, and yet unseen environments.

The **IEMOCAP** dataset is a multimodal emotion dataset that consists of five sessions, each involving two actors (one male and one female) who have scripted and improvised conversations [14]. We remove samples where the labels were partially missing or contained values exceeding the defined label range. For IEMOCAP, we use sessions one through three as the training set and designate session four as the validation set and session five as the test set, as in previous work [15].

MSP-Improv is another acted dataset featuring speech from improvised scenarios designed to elicit particular emotions [16]. Participants were recorded with collar-worn microphones. We partition the dataset into a speaker-independent split by treating sessions one through four as the training set, session five as the validation set, and session six as the test set to be consistent with IEMOCAP data splits.

MSP-Podcast is a dataset consisting of labelled speech from podcasts [17]. We use the predefined training split and validation split and report results on test_set_1. The dataset does not contain transcripts, so we use Microsoft Azure automatic speech recognition to generate them.

Table 1: *Table showing the number of utterances in each dataset following data processing.*

	MSP-Podcast	MSP-Improv	IEMOCAP
# Train	40727	4257	4847
# Val.	7012	1156	1723
# Test	13903	1227	1657

We process each dataset in the same manner by removing instances of speech shorter than three seconds or longer than thirty seconds to introduce a length requirement that an end-user application may impose. Using min-max scaling on each dataset, we bring the averaged evaluator labels for activation and valence into the $[-1, 1]$ range.

3.2. Model architecture

We rely on large, pretrained transformers to encode both speech and text. We use Wav2Vec2 [18][1] acoustic features as they better generalize in cross-domain performance over Mel-spectrogram features [10]. We use BERT [19][2] lexical features from the hidden state corresponding to the CLS token [20, 21]. The CLS token has dropout (p=0.2) applied and the result is passed through a linear layer. Due to the computational complexity of GEM, we freeze the Wav2Vec2 and BERT models to reduce the number of trainable parameters and improve training time. While freezing Wav2Vec2 is less effective than training the full model, the frozen embedding still provides useful information [10, 22].

Our model architecture is similar to that used in recent work utilizing Wav2Vec2 with CCC as the loss function [10]. To account for freezing Wav2Vec2, we use four linear layers on top of the Wav2Vec2 features and apply LeakyReLU to allow for additional complexity. We have two prediction heads for activation and valence, each consisting of two linear layers, which we jointly optimize via multi-task learning. The first three layers after Wav2Vec2 features and the first layer in each prediction head apply LeakyReLU and dropout with probability 0.2 and 0.1 respectively. All linear layers apply LayerNorm to their inputs. Additionally, as BERT features degrade activation performance when using Wav2Vec2 features [10], we concatenate BERT features to the acoustic representation only for valence prediction. Doing this allows us to still regularize the acoustic representation by learning valence without degrading activation performance. Figure 1 presents our multimodal model.

The model is trained using early stopping with a patience of 10, a batch size of 32, and optimized with stochastic gradient descent with a learning rate of 1e-3 using Lin's Concordance Correlation Coefficient (CCC) as the model's loss function [10]. Additionally, we apply tanh and bin the continuous activation and valence to low, medium, and high using the ranges $[-1, -\frac{1}{3}]$, $[-\frac{1}{3}, \frac{1}{3}]$, and $[\frac{1}{3}, 1]$ respectively, and report Unweighted Average Recall (UAR) and CCC for metrics.

3.3. Experiment setup

We explore the effect of continual learning as data from a new domain becomes available. We investigate how a general model will change as it is trained on more restrictive environments simulating adaptation to specific acoustic settings of the end-user. As such, we start by training a general model on MSP-Podcast, which is the largest and most naturalistic dataset. The baseline experiment consists of pretraining and finetuning,

[1] https://huggingface.co/facebook/wav2vec2-base

[2] https://huggingface.co/google/bert_uncased_L-4_H-512_A-8

Table 2: *Activation Results. * indicates statistical significant improvement compared to Finetuning. † indicates statistical significant decrease. P, M, and I indicate training on MSP-Podcast, MSP-Improv, and IEMOCAP respectively. Gray cells represent unseen data results. Bold indicates best performance in the current finetuning setting.*

Seen Datasets	Model Type	MSP-Podcast		MSP-Improv		IEMOCAP	
		CCC	UAR	CCC	UAR	CCC	UAR
P	**Initial Model**	**0.578±0.003**	**0.539±0.028**	**0.520±0.018**	**0.533±0.022**	**0.480±0.031**	**0.458±0.033**
P+M	**Finetuning**	0.453±0.038	0.480±0.020	0.566±0.010	**0.570±0.013**	0.329±0.017	0.455±0.010
P+M	**ExpeR**	0.504±0.018	0.494±0.010	**0.573±0.003**	**0.570±0.013**	**0.509±0.034***	**0.531±0.007***
P+M	**GEM**	**0.527±0.032***	**0.513±0.028**	0.571±0.018	0.570±0.016	0.461±0.038*	0.510±0.014*
P+M+I	**Finetuning**	0.292±0.009	0.416±0.005	0.494±0.016	0.541±0.018	**0.698±0.005**	**0.629±0.018**
P+M+I	**ExpeR**	**0.510±0.015***	**0.496±0.029***	0.557±0.024*	**0.562±0.020**	0.691±0.004	0.609±0.005†
P+M+I	**GEM**	0.508±0.031*	0.487±0.026*	**0.561±0.028***	0.560±0.022	0.682±0.006†	0.599±0.013†

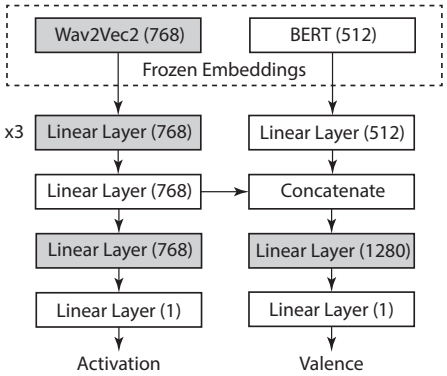

Figure 1: *Diagram of the model architecture. All linear layers have LayerNorm applied to their inputs, and all layers in gray have LeakyReLU and Dropout applied. Dropout is also applied to the BERT CLS token.*

while the memory-based methods will augment the finetuning step of the baseline experiment. The model architecture for each experiment is the same. All results represent the mean and standard deviation over five runs with different seeds.

3.4. Baseline

The model is pretrained and then finetuned on each dataset sequentially. The model is first trained on MSP-Podcast, this pretrained model is then finetuned on MSP-Improv, and finally, we further finetune this model on IEMOCAP. During finetuning, the model has no access to previously seen data and will train until early stopping is triggered.

3.5. Sampling Method

In the memory-based methods, we select samples in the memory using a ringbuffer that tracks the most recently seen training samples (a single utterance from the training dataset) and stores the ringbuffer as memory once training concludes. Since the models train on multiple epochs and shuffle the training data, we will refer to this as **random memory**. We also consider three alternative sampling methods to investigate the impact of samples stored in the memory.

In all sampling cases, we fill the episodic memory with samples that capture a representative distribution of a given dataset with regards to activation and valence. To do this, we divide the dataset into nine buckets containing speech from each combination of activation and valence over low, medium, and high bins. For each bucket, we then choose samples to represent audio length, gender, and speaker identity.

For **audio length sampling**, we follow a sampling method used in previous work where 30% of data come from the first and fourth quartiles of audio length, with the remaining 70% from the second and third [23]. For **gender sampling**, we balance the samples in each bucket between male and female speakers. Likewise, for **speaker sampling**, we equally balance samples in each bucket in regard to the sample's speaker. When the speaker or gender is unknown, the speech is not a candidate to be stored in model memory.

3.6. Gradient Episodic Memory (GEM)

We use GEM to perform cross-corpus continual learning with the aim of minimizing performance degradation as new datasets are added during training. GEM usually imposes a restriction of only one epoch of training on new data. We differ from the original GEM implementation by allowing multiple training epochs to ensure model convergence to the new acoustic domain during training on a given dataset to simulate a deployed model converging to new incoming speech. The number of epochs is determined by early stopping (Section 3.2).

We use an episodic memory size of 261 per dataset since this is divisible by nine, which is helpful in our sampling methods (Section 3.5). Other GEM parameters are unchanged[3].

3.7. Experience Replay (ExpeR)

We augment the baseline finetuning approach with an episodic memory of size 261 to replicate the memory size of GEM. When the model encounters a new dataset, a new episodic memory of 261 samples is created and filled by sampling (Section 3.5). During training, all samples in the episodic memory are added to the current training dataset, and the model proceeds with finetuning similarly to Section 3.4.

4. Results

4.1. Memory-based approaches

We present our results in Table 2 and we find that memory-based methods retain emotion knowledge across varying acoustic environments. We test statistical significance using a two-sided paired t-test with a confidence threshold of 95%.

4.1.1. Baseline

We find that as more domains are learned, the baseline method (finetuning) forgets significantly. Specifically, CCC on MSP-Podcast reduces from 0.578 to 0.453 and 0.292 when further finetuned on MSP-Improv and then IEMOCAP, respectively (Table 2). We find that the baseline also drops in MSP-Improv

[3]https://github.com/facebookresearch/GradientEpisodicMemory

from a CCC of 0.566 to 0.494 once trained on IEMOCAP. Additionally, we find that a model trained on MSP-Podcast (P setting) performed much better on IEMOCAP CCC compared to after finetuning on MSP-Improv (P+M setting). This demonstrates how naive adaptation to new speech can decrease a model's performance on both unseen and previously seen data.

4.1.2. GEM

GEM outperforms the baseline experiment on previously seen datasets at all stages of training for both UAR and CCC (Table 2). In the P+M setting, we find statistically significant improvement on MSP-Podcast CCC compared to the finetuning P+M baseline. Additionally, we find that the unseen IEMOCAP performance significantly improves compared to the baseline. We find that GEM has maintained MSP-Improv performance compared to P+M finetuning despite limitations imposed during learning of MSP-Improv.

In the P+M+I setting, we find statistically significant improvement compared to the P+M+I finetuning baseline for MSP-Podcast performance in both metrics, and on MSP-Improv when considering CCC. GEM struggles the most with learning IEMOCAP in this setting, likely due to GEM having the strictest learning process for new data.

4.1.3. ExpeR

ExpeR follows similar performance to GEM, except for MSP-Podcast performance in the P+M setting. While ExpeR shows improvement, it is not significant over the P+M finetuning baseline. This is likely because GEM has stricter constraints. In the baseline method, finetuning on MSP-Improv was more harmful to MSP-Podcast (Table 2 initial model $\rightarrow$ P+M finetuning: -0.125 CCC) than finetuning on IEMOCAP was harmful to MSP-Improv (Table 2 P+M finetuning $\rightarrow$ P+M+I finetuning: -0.072). This suggests that the MSP-Podcast and MSP-Improv domains may be more in conflict, and GEM's stricter learning may improve retention in such cases.

We find that ExpeR marginally improves on adapting to new data compared to GEM. In the P+M setting, we find that ExpeR not only improves the unseen IEMOCAP performance over GEM, but further improves on the initial general model (P setting), suggesting ExpeR has managed to leverage new information that improves generalizability. The improvement over GEM suggests the learning limitations that improve GEM's retention in this setting may also be preventing GEM from learning a more general model that ExpeR may be learning.

4.2. Sampling approaches

We find that the samples stored in memory significantly impact the model performance and retention. We present results only for ExpeR as the trends in sampling results are similar for GEM. Table 3 shows relative performance improvements against random sampling. We observe significant UAR performance improvement on MSP-Podcast in both P+M and P+M+I settings. The sampling methods also generally show improved performance retention over random memory on MSP-Improv in the P+M+I setting (significant only for speaker sampling).

All sampling methods show significantly improved MSP-Podcast retention in the P+M setting. However, in P+M+I, sampling did not improve CCC performance on MSP-Podcast (Table 3), though there was some improvement on CCC scores for MSP-Improv in this setting. Table 2 shows that the CCC drop from the baseline finetuning experiment is greatest for MSP-Podcast in the P+M setting, so this may suggest that sampling is more useful the more domains conflict with one another.

Table 3: *Activation Results relative to ExpeR performance.* $\uparrow$ *indicates metric increase* $\downarrow$ *indicates decrease. *, †, P, M, and I as in Table 2*

Seen Datasets	Sampling	MSP-Podcast CCC	MSP-Podcast UAR	MSP-Improv CCC	MSP-Improv UAR
P+M	**Length**	0.031↑*	0.059↑*	0.018↓†	0.000↑
P+M	**Gender**	0.035↑*	**0.068↑***	**0.006**↓†	0.003↓
P+M	**Speaker**	**0.043↑***	0.047↑*	0.007↓	**0.005**↑
P+M+I	**Length**	0.031↓	0.049↑*	0.010↑	0.017↑
P+M+I	**Gender**	0.011↓	**0.058↑***	**0.022**↑	0.028↑
P+M+I	**Speaker**	**0.002**↓	0.049↑*	0.019↑	**0.030↑***

We hypothesize that the improvement in UAR compared to CCC results from the fact that the memory sample distribution is equal across activation and valence bins. For CCC, this balancing may be problematic because the memory sample distribution will not match across datasets.

4.3. Discussion

ExpeR is much less computationally expensive than GEM. Since GEM must make predictions and calculate the loss for samples in the episodic memory at each training step. For each training batch, GEM must also run through all memory samples to compare the gradients with the current batch's update gradient (i.e., 522 samples in P+M+I). In comparison, ExpeR will only increase the training set size by the number of samples.

As discussed in Section 4.1.3, ExpeR may be better at learning general models due to its strong performance on unseen data. Combined with the much lower computational cost and comparable performance, in most cases ExpeR would be preferential to GEM for adapting SER models to new data. However, there are two notable exceptions.

Firstly, GEM exhibits much less forgetting when the domains conflict more (Section 4.1.3). Secondly, GEM does not train directly on the memory samples, which is significant for an SER model that continually learns. GEM may be essential in cases where batch updates are much smaller than the sample memory size, which could be useful in end-user devices, which may perform batch updates frequently with limited samples. ExpeR will train on a training set that consists primarily of memory samples introducing a significant risk of overfitting.

5. Conclusion

Episodic memory methods are effective at retaining classification performance on previously seen datasets and acoustic environments. GEM and ExpeR have comparable performance, with ExpeR being much more computationally efficient but GEM offering other benefits such as less forgetting. Additionally, we show that sampling episodic memory can significantly improve categorical classification performance. In future work, we will look at adaptation to emotion labels reported by the speaker and how memory methods can retain SER performance when adapted with unsupervised auxiliary tasks. Furthermore, we will investigate how these methods integrate with privacy-preserving, federated learning.

6. Acknowledgements

This material is based in part upon work supported by the National Science Foundation (NSF IIS-RI 2230172 and IIS-RI 2230172).

7. References

[1] J. J. Gross, H. Uusberg, and A. Uusberg, "Mental illness and well-being: an affect regulation perspective," *World Psychiatry*, vol. 18, no. 2, pp. 130–139, 2019.

[2] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health," *Clinical psychology: Science and practice*, vol. 2, no. 2, pp. 151–164, 1995.

[3] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[4] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Interspeech*, 2017.

[5] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017.

[6] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," *ICML Workshop: Multi-Task and Lifelong Reinforcement Learning*, 2019.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[8] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] D. Lopez-Paz and M. A. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf

[10] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *ArXiv*, vol. abs/2203.07378, 2022.

[11] Z. Zhang, B. Wu, and B. Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6705–6709.

[12] S. V. Eeckt and H. V. hamme, "Continual learning for monolingual end-to-end automatic speech recognition," in *European Signal Processing Conference*, 2021.

[13] M. Yang, I. Lane, and S. Watanabe, "Online Continual Learning of End-to-End Speech Recognition Models," in *Interspeech*, 2022, pp. 2668–2672.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[15] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3502–3506.

[16] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[17] 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, October-December 2019.

[18] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[20] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 3755–3759. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1212

[21] M. Perez, M. Jaiswal, M. Niu, C. Gorrostieta, M. Roddy, K. Taylor, R. Lotfian, J. Kane, and E. M. Provost, "Mind the gap: On the value of silence representations to lexical-based speech emotion recognition," in *Proc. Interspeech 2022*, 2022, pp. 156–160.

[22] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[23] M. Jaiswal and E. M. Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition "in the wild"," *arXiv preprint arXiv:2104.08806*, 2021.