Domain mismatch and data augmentation in speech emotion recognition

Dimitra Emmanouilidou, Hannes Gamper, Midia Yousefi

Microsoft, WA, USA

{dimitra.emmanouilidou, hannes.gamper, midiayousefi} @ microsoft.com

Abstract

Large, pretrained model architectures have demonstrated potential in a wide range of audio recognition and classification tasks. These architectures are increasingly being used in Speech Emotion Recognition (SER) as well, an area that continues to grapple with the scarcity of data, and especially of labeled data for training. This study is motivated by the limited research available on the robustness and generalization capabilities of these models for SER and considers applicability beyond a restricted dataset. We invoke the widely adopted network architecture CNN14 and explore its ability to generalize across different datasets. Our analysis demonstrates a potential domain gap between datasets after analyzing the acoustic properties of each one. We bridge this gap with the introduction of acoustic and data variability, by invoking seven suitable augmentation methods. Our approach leads to up to 8% improvement for unseen datasets. However, bridging the acoustic mismatch seems to play a minor role only: an infelicitous finding involving partially scrambled (swapped) annotation labels hints to deeper domain mismatches during multi-dataset learning scenarios. Findings in this work are applicable to any large or pretrained network and contribute to the ongoing research on the robustness and generalization of SER models.

Index Terms: speech emotion recognition, human-computer interaction, computational paralinguistics, data augmentation

1. Introduction

Speech emotion recognition (SER) aims to automatically discern the emotional state of a speaker from their speech signal and is still considered as one of the big challenges in speech processing [1]. Alongside human facial expressions, speech emerges as a highly promising avenue for automated identification of human emotions [2]. Increasing interest in this field supports the pivotal importance of emotion recognition, particularly evident in safety monitoring, human-computer interaction, personal well being, video games [3–5]. Nonetheless, to achieve a robust and universally applicable approach to emotion classification, it is important to consider model efficiency in cross-dataset and cross-domain adaptation. It is beneficial to go even further; conduct comprehensive examinations and provide clarity in experimental findings along with limitations [6, 7]. This is the very motivation of this work, to go beyond reporting top performance on a dataset, while we focus on insightful contributions to ongoing and future advancements in the field.

One of the main obstacles in speech emotion recognition (SER) research is the shortage of data [8]. Common challenges with SER datasets include limited annotated data, variability across languages and dialects, and small-scale datasets [4,9,10]. Moreover, these datasets often lack comprehensive coverage of emotional categories, languages, cultures, and speakers. They vary significantly in several aspects, such as the presence of real or acted emotions, annotation methods and sources, recording conditions, but also when it comes to instructions given to speakers and to annotators [5]. These differences pose a significant challenge for SER model training, resulting in domain mismatches and distribution shifts that ultimately harm model performance and generalization, potentially leading to overfitting across datasets [11]. Data augmentation shows promise in addressing some of these challenges. It involves generating new data samples from the original ones, through transformations like noise addition, masking, and time stretching. Though widely used in computer vision and natural language processing to enhance model performance and generalization, data augmentation in SER has received less attention, often with contradictory evidence. Prior work in this area has showed little or no benefit, especially for larger or more diverse datasets [12–14]. Most existing studies have focused on augmenting individual, single datasets with less attention on general trends that can go beyond choosing a single point on the performance curve.

In this paper, we examine the challenges of recognizing speech emotions across multiple datasets, and provide with a series of ablation studies. Our analysis relates to the use of large pre-trained models in SER, addressing two questions continuing to riddle the community: (1) Is data augmentation an effective technique for bridging the domain mismatch gap in SER ? (2) Does the model learn sufficiently generalizable features per class in cross-dataset learning?

We address (1) by first analysing the acoustic mismatch across datasets. We then bridge this gap by selecting suitable augmentation methods. We evaluate our approach across three SER datasets and demonstrate cross-dataset performance enhancement up to 8%. While significant, this improvement still renders a performance far from the baseline (i.e. train and test on single dataset), which illustrates the challenges of overcoming domain mismatches between various SER datasets. For question (2), we investigate the ability of the model to capture generalizable, class-specific features. We do so by introducing artificial label confusion (scrambling) across the datasets, in joined learning. Surprisingly, we find that model performance is virtually unaffected by conflicting labels across datasets, illustrating a default inability to capture generalizable emotional features, even with pretrained models. Overall, our findings show evidence that domain mismatches run deep in SER datasets, obscuring the capture of universal features and misleading the model into per dataset class memorization. The rest of the paper is organized as follows: section 2 reviews the datasets and evaluation metrics. Sections 3 and 4 describe the model architecture and baselines. Section 5 presents the augmentation methods and experiments, and discusses findings and limitations.

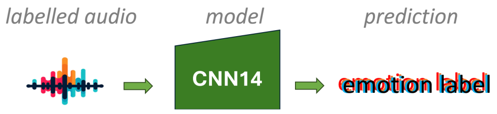


Figure 1: *Depiction of the model, with the output layer predicting the major emotion classes: Happy, Neutral, Sad, Angry.*

2. Datasets and Metrics

2.1. Datasets

MSP-Podcast v1.10 [15]: This large speech emotion dataset comprises of $>100,000$ audio samples (~ 165 hours), with an average duration of 5.7 sec. Samples belong to podcast recordings of >600 speakers. We included all segments with reviewer consensus within the four classes {*Happy*, *Neutral*, *Sad*, *Angry*}, a total of $67,929$ files. We used the standard split for training, *Test1* split for testing and *Dev* for validation.

IEMOCAP [16]: The Interactive Emotional Dyadic Motion Capture dataset is an acted dataset of scripted and improvised dialogues by 10 speakers, with a duration of $\sim$ 12 hours, and an average clip duration of 4.5 sec. We included all samples with reviewer consensus for labels {*Happy*, *Neutral*, *Sad*, *Angry*}. We split the data by speaker (SP), keeping one speaker for testing, one for validation, and the rest for training. We average results over all 10 cross validations for the 10 speakers.

CREMA-D [17]: The Crowd-sourced Emotional Multimodal Actors Dataset contains 7,442 original clips spoken by 91 actors, while reciting 12 unique sentences. This $\sim$ 5-hour dataset contains samples of average duration 2.5 sec. We only considered the four emotional labels as before. We split data randomly at 70-15-15% for training, testing, validating.

2.2. Evaluation Metrics

Model performance was evaluated by: $\mathbf{wACC}$, the percentage of correctly classified samples over all samples (sample-weighted accuracy); $\mathbf{uACC}$, the average of individual class accuracies, not affected by imbalanced classes (unweighted or balanced accuracy); $\mathbf{wF1}$, the weighted F1 score accounting for both precision and recall while considering class imbalances. We used $\mathbf{wF1}$ to compare performances across the various scenarios. All metrics were normalized to [0,1]. For IEMOCAP, metrics were averaged over the 10 per-speaker cross validation folds. For the rest of the datasets, the splits were predetermined (Section 2.1). We further provide confidence intervals for all metrics, by averaging over the last 20% of all training iterations of the model; this corresponds, approximately, to averaging the performance over the last 8 full epochs of training. We ensured that these averages are meaningful, computed after both the loss and performance curves start plateauing.

3. Model Architecture

The overall model architecture used for the speech emotion classification task consists of a pre-trained audio encoder and an additional classification layer on top, Figure 1. The audio encoder $f(a)$, where a_i represents the raw audio, first produces a log Mel Spectrogram from raw audio followed by a learnable embedding function. The audio representation $x_a = \{f(a)\}$ will be of dimension $x_a \in \mathbb{R}^{b \times v}$, where b is the batch size, and v the dimension of the audio representation. x_a is then passed through a linear projection layer with ReLU activation. For this task, the predictions are passed through Softmax activation and the choice of loss is Binary Cross Entropy.

We used CNN14 [18] as the audio encoder, chosen for its wide use in SER. Transformer-based architectures, pre-trained in a self-supervised manner, have shown great promise in many machine learning tasks, including SER and leading to wide adoption of alike architectures [7, 19–26]. The CNN14 encoder was pretrained on large amounts of speech, specifically on AudioSet [27], where already Speech represents about half of the dataset. Data is sampled at 16 KHz, with a 64-bin Log Mel Spectrogram as raw features; hop size of 320, window size 1024 and frequency range between 50 to 14000 Hz. We finetune the CNN14 encoder along with the task linear layers in PyTorch, with a batch size of 128 and learning rate of 10^{-4}, over a maximum of 30 epochs. We fix the batch size and learning rate to this commonly adopted configuration for ease of reproducibility and future comparisons, without any hyper-parameter tuning.

4. Baseline Systems

In this section we showcase baseline model performance when training and testing on the same dataset (i.e. in-domain scenarios), Table 1. Pursuing extensive hyper-parameter tuning experimentation is not within the scope of this work. However, we quickly want to illustrate that even light parameter tuning achieves state-of-the-art equivalent performance. We select one training scenario, training on IEMOCAP, and we perform a light hyper parameter tuning (*H-param tune)*. We only explore 4 discrete values for the learning rate $\in$ [0.0001, 0.01] and 4 values for batch size [16, 64, 128, 256]. Even on this small set of hyper-parameters, we already see a substantial increase in $\mathbf{wF1}$ performance by up to 6%, comparing rows 2 and 1 in Table 1. This evaluation was useful for validating that the network reaches top, per-speaker, performance on IEMOCAP [28], and that any lower baseline performance shown here does not affect generalization of findings. The exercise of surpassing the state-of-the-art is beyond the goal of this analytic study; we continue the rest of this work while keeping all network parametrization fixed, and focus our attention to the overall trends and findings.

5. Generalization to Unseen Datasets

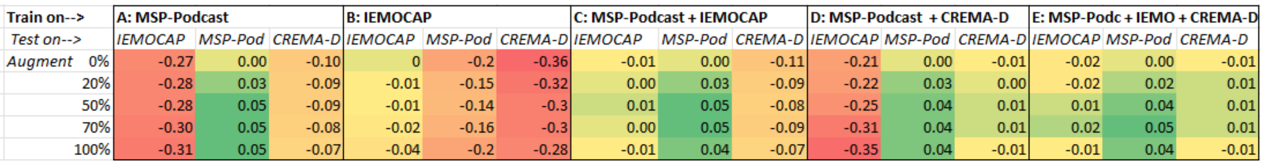
The CNN14 architecture has been widely adopted in classification tasks including speech emotion (SER). In this section we look at the model's classification generalization across the various emotion datasets. The model was pretrained on AudioSet, and its architecture is quite large; we thus hypothesize that it is possible for the model to capture the general, core features corresponding to individual emotion classes, especially as we allowed for brief model finetuning on each training scenario.

Figure 2 depicts results across corpora learning (columns A to E), when testing on each dataset (subcolumns). For now, we focus on row *Augment 0%*. Cell values depict *relative change* in performance $\mathbf{wF1}$, relative to baseline results of Table 1, with green color emphasizing high, and red color low intensity. In Column A, a value of -0.27 in the first cell signifies that training on MSP-Podcast and testing on IEMOCAP brings a 27% drop in performance compared to training and testing on IEMOCAP (baseline $\mathbf{wF1}$=0.63). In Column C, a value of -0.11 in the first row, signifies 11% drop in CREMA-D test performance with joined training on MSP-Podcast and IEMOCAP, compared to training on CREMA-D alone (baseline $\mathbf{wF1}$=0.79).

Looking across the whole first row, *Augment 0%*, in Figure 2, we immediately notice *poor generalization capability of the baseline systems to an unseen dataset*. Let's focus on

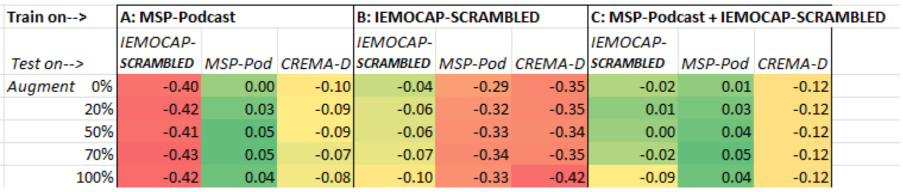
Table 1: *Baseline performance evaluation averaged over the last 20% of model iterations (see Section 2.2). Confidence intervals (std) are shown in parenthesis. Performance at chance level is at 0.25 (4-way classification). IEMOCAP dataset was split by speaker (SP).*

Baseline System 0% Augmentation					**Augmented**			
					20%	**50%**	**70%**	**100%**
Condition	Train/Test	**uACC**	**wACC**	**wF1**	**wF1**			
H-param tune	IEMOCAP	.68 (5e-3)	.67 (5e-3)	.68 (6e-3)				
No tuning	IEMOCAP	.64 (3e-3)	.62 (4e-3)	.63 (4e-3)	.62 (4e-3)	.62 (3e-3)	.61 (4e-3)	.59 (3e-3)
No tuning	CREMA-D	.80 (4e-3)	.59 (2e-3)	.79 (3e-3)	.78 (1e-3)	.78 (1e-3)	78 (1e-3)	76 (1e-3)
No tuning	MSP-Podcast	.54 (4e-3)	.36 (2e-3)	.53 (3e-3)	.56 (2e-3)	.58 (2e-3)	.58 (2e-3)	.58 (2e-3)



Train on-->	A: MSP-Podcast			B: IEMOCAP			C: MSP-Podcast + IEMOCAP			D: MSP-Podcast + CREMA-D			E: MSP-Podc + IEMO + CREMA-D		
Test on-->	*IEMOCAP*	*MSP-Pod*	*CREMA-D*	*IEMOCAP*	*MSP-Pod*	*CREMA-D*	*IEMOCAP*	*MSP-Pod*	*CREMA-D*	*IEMOCAP*	*MSP-Pod*	*CREMA-D*	*IEMOCAP*	*MSP-Pod*	*CREMA-D*
Augment 0%	-0.27	0.00	-0.10	0	-0.2	-0.36	-0.01	0.00	-0.11	-0.21	0.00	-0.01	-0.02	0.00	-0.01
20%	-0.28	0.03	-0.09	-0.01	-0.15	-0.32	0.00	0.03	-0.09	-0.22	0.03	0.00	-0.02	0.02	0.01
50%	-0.28	0.05	-0.09	-0.01	-0.14	-0.3	0.01	0.05	-0.08	-0.25	0.04	0.01	0.01	0.04	0.01
70%	-0.30	0.05	-0.08	-0.02	-0.16	-0.3	0.00	0.05	-0.09	-0.31	0.04	0.01	0.02	0.05	0.01
100%	-0.31	0.05	-0.07	-0.04	-0.2	-0.28	-0.01	0.04	-0.07	-0.35	0.04	-0.01	-0.01	0.04	-0.01

Figure 2: *Change in* **wF1** *performance relative to Table 1 baselines across learning scenarios (columns) and augmentation % (rows).*



Train on-->	A: MSP-Podcast			B: IEMOCAP-SCRAMBLED			C: MSP-Podcast + IEMOCAP-SCRAMBLED		
Test on-->	*IEMOCAP-SCRAMBLED*	*MSP-Pod*	*CREMA-D*	*IEMOCAP-SCRAMBLED*	*MSP-Pod*	*CREMA-D*	*IEMOCAP-SCRAMBLED*	*MSP-Pod*	*CREMA-D*
Augment 0%	-0.40	0.00	-0.10	-0.04	-0.29	-0.35	-0.02	0.01	-0.12
20%	-0.42	0.03	-0.09	-0.06	-0.32	-0.35	0.01	0.03	-0.12
50%	-0.41	0.05	-0.09	-0.06	-0.33	-0.34	0.00	0.04	-0.12
70%	-0.43	0.05	-0.07	-0.07	-0.34	-0.35	-0.02	0.05	-0.12
100%	-0.42	0.04	-0.08	-0.10	-0.33	-0.42	-0.09	0.04	-0.12

Figure 3: *Change in* **wF1** *relative to Table 1 baselines, after scrambling IEMOCAP labels (Neutral* ⟷ *Sad ; Happy* ⟷ *Angry).*

columns A and B, subcolumn IEMOCAP: baseline **wF1** performance on IEMOCAP is 0.63 (Table1); this performance drops to 0.36 (not shown) when training on MSP-Podcast and testing on IEMOCAP, registering a drop value of -0.27 in Figure 2, first cell (0.36−0.63=-0.27). Considering that MSP-Podcast is about 10x larger and contains many and more varied speakers and spoken content, this result may come as a surprise. Could it be that the acoustic conditions are so different among the datasets, that generalization fails? We explore this question in the following sub-sections.

Let us now compare the three training scenarios in columns A, B, C, fixing the test set to (subcolumns) IEMOCAP. Relative **wF1**=-0.27 in column A, when training on MSP-Podcast. Adding IEMOCAP for joined training in column C, increases relative **wF1** to -0.01, which is almost a ”full” performance recovery compared to baseline IEMOCAP. In other words, adding MSP-Podcast in training brings no performance benefit to IEMOCAP. Considering that the size of IEMOCAP in joined training is only a fraction of the total data ($\sim$ 10%), this result may be an indication that the network is unable to find common patterns of the corresponding emotion labels across datasets.

If the acoustic conditions in the two datasets are highly mismatched, the network could be treating audio samples from the two datasets separately, even when training jointly. Given these findings, we hypothesize and investigate whether the observed poor generalization can be a result of the mismatch in acoustic conditions across datasets, or whether there may exist a deeper bias in cross-domain learning in SER.

5.1. Acoustic Conditions and Data Augmentation

While prior studies have integrated data augmentation into network training in SER, findings are generally analyzed on a per dataset basis [22, 29], and have lead to conflicting findings in the community. Here, we look into cross-domain data augmentation, intended to bridge the gap of the acoustic conditions between datasets and promote joined learning. If the acoustic conditions of the individual datasets are highly dissimilar, this fact alone could prevent the model from achieving generalization ability. For example, audio excerpts from IEMOCAP exhibit increased background noise and reverberation, which is not often present in MSP-Podcast. IEMOCAP samples were recorded using shotgun microphones pointed at two actors in a medium-sized room filled with cameras and motion tracking equipment. This recording setup may introduce more background noise and reverberation than one would expect in a typical podcast setting.

We could argue that if we bridge the gap on the acoustic conditions across datasets, this may help the model improve generalization on unseen data. Figure 4, illustrates the domain mismatch between IEMOCAP and MSP-Podcast in terms of the quality and reverberance estimated from the audio samples using blind estimation methods for MOS, T60, C50 [30–32]. Top row, *raw*, shows the statistics of the original audio samples. The estimated parameters clearly depict high dissimilarity among the two datasets. We carefully select a number of augmentation techniques to help mitigate the large difference in noise and acoustic variability, and to bring the acoustic parameters of MSP-Podcast closer to those present in IEMOCAP. After employing the augmentations proposed in the next section, we see in Figure 4 bottom row, a clear shift in the distributions of the acoustic parameters of the MSP-Podcast corpus, bridging the acoustic gap of the two corpora.

5.2. Types of Augmentation

We augment the training samples using a set of methods suited to enhance robustness of speech recognition models, can introduce variations in speaking styles, and introduce variations in

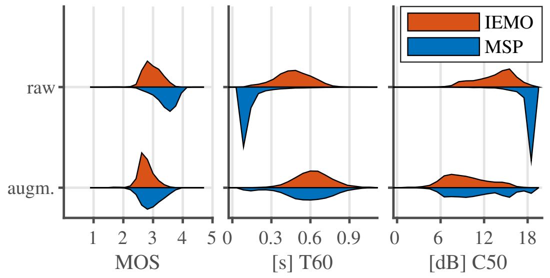


Figure 4: *Distribution of estimated mean opinion score (MOS), reverberation time (T60), and clarity (C50) for IEMOCAP and MSP-Podcast, with and without data augmentation. Each statistical distribution shown is averaged over 10,000 random samples. Notice how the application of the selected augmentations (row augm.) bridges the distribution gap across datasets.*

recordings, acoustic conditions and ambient noise. **Specaugment** [33], allowing for masking of both temporal and spectral axis; **TimeStretching**, allowing for speech rate scaling, where we limit the stretch factor within [0.9, 1.2] to prevent altering of the emotional signature; **TimeShifting**, allowing for temporal variations; **Fading** in no more than 10% of the signal; **Equalization**, implemented via a series of high and low pass 9th order Butterworth filters; **pink noise**, for ambient noise; and **reverberation**, implemented with the pedalboard python package, room size factor up to 0.2, wet level factor up to 0.5, dry level factor of 0.4. These augmentations are applied to each original audio sample, with parameters randomly chosen within the allowed ranges. We partially augment the training sets at increasing rates of **augmentation** $\mathbf{\{0\%, 20\%, 50\%, 70\%, 100\%\}}$, corresponding to the percentage of the samples affected: 0% means no augmentation; 50% means half the samples were augmented in training, and so on. Augmentations were applied on the fly, uniquely for each file in the training batch.

5.3. Findings on Augmentation

Effect on baseline scenarios, Table 1: increasing amounts of augmentation improve $\mathbf{wF1}$ by up to 5% for the large dataset MSP-Podcast, but the effect is not apparent for the smaller or less varied sets of IEMOCAP and CREMA-D. *Effect on unseen data, Figure 2*: the figure summarizes relative $\mathbf{wF1}$ change at various rates of data augmentation during training (rows), and across datasets (columns). See Section 5 on interpreting relative $\mathbf{wF1}$. Testing on unseen CREMA-D benefits from augmentation by as much as 3% column A, and 8% column B: from -0.36 $\mathbf{wF1}$ drop to -0.28 drop. This is also evident in joint training column C. We further notice deterioration on unseen IEMOCAP column A, despite having bridged acoustic differences with augmentation. Additional exploration is needed to interpret this bias, evident in joined training in column D too.

Overall, the addition of augmentation seems to allow for performance improvement within a dataset and on unseen dataset scenarios, but it doesn't seem to drastically improve cross-dataset generalization. The domain mismatches seem more intricate, and may go beyond acoustic conditions. This could be an indication that emotion manifestations in classes with the same-name across datasets may be different or have a different "meaning" depending on labeling setting or the dataset context (podcast vs actors). The ambiguous cross-dataset generalization also indicates that large pretrained networks like CNN14 may be picking up on those differences and separating datasets. Next, we look into this very hypothesis by introducing label confusion. What is the impact in joined training performance when we swap (scramble) the labels of one dataset? If class features are arguably learnt across datasets, does label scrambling cause a drastic performance drop?

5.4. Introducing Label Confusion (Scrambling)

Now that mismatched acoustic variability has been addressed with targeted augmentations in the previous section (Figure 4), the goal here is to understand if learning across datasets produces generalizable features per class. For training, we utilize MSP-Podcast and IEMOCAP, the largest datasets the model can learn from. Then, we introduce confusion during cross-dataset learning, via label scrambling: we swap the labels of *Neutral* with *Sad* ; and the labels of *Happy* with *Angry* , but only for the IEMOCAP dataset. Let us assume that the model is able to capture good or generalizable emotional features. Then the introduction of the SCRAMBLED IEMOCAP labels should confuse the model and cause significantly low performance during joined training. We compare results between Figures 2 and 3.

Looking at the interesting case of sub-Columns C, where we combine MSP-Podcast and IEMOCAP-SCRAMBLED during training, performance on the test set of IEMOCAP-SCRAMBLED seems unaffected. This contradicts the earlier assumption of achieving generalizable features, and illustrates that the network fully separates the learning on the two datasets, treating the data and the classes of the two datasets separately during learning/training. This finding encapsulates the dangers of using large models for datasets with moderate size or variety (acoustical, setup, etc), and of interpreting multi-corpora training generalizations, when the same corpora appear in both train and test sets. The catastrophic finding of the scrambled-labels study highlights that learning mitigations are needed, either for appropriate cross-corpora class-label unification or towards promoting robust per-class features during cross-dataset learning.

Looking at the effect of augmentation in the case of scrambled labels, we see that the performance on IEMOCAP-SCRAMBLED decreases with higher rates of augmentation, across all training scenarios. This is a positive indication that the selected augmentations are meaningful, preventing the classifier from performing well after label scrambling, or in other words, help promote generalizable learning on correct labels.

6. Conclusion

In this work, we explored the generalization ability of a widely used pretrained network architecture in SER. While prior works often focus on achieving high performance per specific dataset, cross-dataset experimentation often remains under-explored. We investigate various factors that can improve the generalization ability of the network, including training on multiple, disjoint datasets and shifting the acoustic parameter distribution of the datasets to bring them (acoustically) closer together via selected augmentations. We conclude that paying close attention to the acoustic conditions of each dataset is an important cue, that can lead to up to 8% of improvement for unseen datasets, and it further allows for better feature generalization (see scrambled experiment). While acoustic condition matching, alone, yields a noticeable performance increase, it is still far from reaching in-domain performance equivalence. Other factors of domain mismatch across datasets still riddle generalization, potentially including differences in instructions given to speakers or to reviewers, and warrant continued exploration.

7. References

[1] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, "Speech emotion recognition using machine learning—a systematic review," *Intelligent systems with applications*, p. 200266, 2023.

[2] J. de Lope and M. Grana, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, 2023.

[3] J. Ye, X.-C. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] A. Al-Talabani, H. Sellahewa, and S. A. Jassim, "Emotion recognition from speech: tools and challenges," in *Mobile Multimedia/Image Processing, Security, and Applications 2015*, vol. 9497. SPIE, 2015, pp. 193–200.

[5] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23 745–23 812, 2021.

[6] W.-S. Chien and C.-C. Lee, "Achieving fair speech emotion recognition via perceptual fairness," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] A. Derington, H. Wierstorf, A. Özkil, F. Eyben, F. Burkhardt, and B. W. Schuller, "Testing speech emotion recognition machine learning models," 2023.

[8] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *2017 International conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2017, pp. 109–114.

[9] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface' 05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, pp. 8–8.

[10] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS One*, vol. 13, no. 5, p. e0196391, 2018.

[11] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: challenges," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 1, pp. 16–28, 2015.

[12] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[13] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5084–5088.

[14] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

[15] 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, 2019.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[19] 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[20] D. Tompkins, D. Emmanouilidou, S. Deshmukh, and B. Elizalde, "Multi-view learning for speech emotion recognition with categorical emotion, categorical sentiment, and dimensional scores," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[24] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

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

[26] H. Dhamyal, B. Elizalde, S. Deshmukh, H. Wang, B. Raj, and R. Singh, "Prompting audios using acoustic properties for emotion representation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 936–11 940.

[27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[28] Y. Wang, C. Lu, H. Lian, Y. Zhao, B. W. Schuller, Y. Zong, and W. Zheng, "Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 646–11 650.

[29] V. Praseetha and P. Joby, "Speech emotion recognition using data augmentation," *International Journal of Speech Technology*, vol. 25, pp. 783–792, 2022.

[30] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 136–140.

[31] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 85–89.

[32] H. Gamper, "Blind c50 estimation from single-channel speech using a convolutional neural network," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.