DESIGNING AND EVALUATING SPEECH EMOTION RECOGNITION SYSTEMS: A REALITY CHECK CASE STUDY WITH IEMOCAP

Nikolaos Antoniou [†,∗], *Athanasios Katsamanis* [†,∗], *Theodoros Giannakopoulos* [†], *Shrikanth Narayanan* [†,+]

[†] Behavioral Signal Technologies, Los Angeles, CA, USA
∗ Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
+ SAIL-University of Southern California, Los Angeles, CA, USA
{nikos.antoniou,nassos,thodoris,shri}@behavioralsignals.com

ABSTRACT

There is an imminent need for guidelines and standard test sets to allow direct and fair comparisons of speech emotion recognition (SER). While resources, such as the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) database, have emerged as widely-adopted reference corpora for researchers to develop and test models for SER, published work reveals a wide range of assumptions and variety in its use that challenge reproducibility and generalization. Based on a critical review of the latest advances in SER using IEMOCAP as the use case, our work aims at two contributions: First, using an analysis of the recent literature, including assumptions made and metrics used therein, we provide a set of SER evaluation guidelines. Second, using recent publications with open-sourced implementations, we focus on reproducibility assessment in SER.

***Index Terms*—** Speech Emotion Recognition, emotion evaluation, reproducibility, IEMOCAP

1. INTRODUCTION

Systems with the ability to recognize emotions, simply by processing speech information can be particularly useful for building intelligent machines that can incorporate perceived affective expressions. Inferring expressed human affective state can, for instance, help a voice assistant to adjust its response to the user. Speech Emotion Recognition (SER) falls within a broad class of problems within *computational paralinguistics* [1].

As is typical in such speech machine-learning applications, the cornerstone of progress is the use of annotated datasets. These resources enable us to build and evaluate the performance of systems that learn how to map successfully an input speech sample to the desired output. Collecting appropriate input-label pairs for the SER task is a highly ambiguous procedure. The ambiguity lies on the fact that the quality of labels depends on the annotators' nuanced perception of emotions. Besides that, there is an inherent trade-off between data quality and the desire to facilitate annotation when gathering emotional speech. Spontaneously spoken segments, for example, can be of high (audio) quality but may contain overlapping and or ambiguous affective expressions which can confuse annotators and exacerbates the difficulty of labelling such data [2]. On the other hand, imposing a one-to-one correspondence between a certain speech segment and its emotion may give rise to overly artificial data samples for learning and inference.

Several published studies in the SER domain have used the *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) [3] database. Using a critical review of the recent works using IEMOCAP for SER as use cases, this work aims to investigate questions centered on how the adopted evaluation approaches compare across the studies, and how guidelines for comparable evaluation and reproducible findings can be developed. Our contributions are summarized as follows:

- Using a critical review of past studies, we identify three methodological limitations in the adopted evaluation methods. Our analysis reveals that the lack of consensus about the evaluation protocol leads to results of narrowed range which can not be properly contextualized and generalized across related work in the domain. Based on these insights, we summarize a recommended set of evaluation guidelines (building on practice from several previous works), with the end goal of minimizing potential blind spots related to the overall SER procedure.

- We also report a reproducibility study, assessing whether models which provide opensource code implementations deliver results consistent with the published findings.

2. RELATED WORK

Deep Learning approaches have triggered a paradigm shift including speech emotion recognition (SER) research and development. Systems that solve SER have largely transitioned from classical machine learning models, such as HMMs or SVMs, applied on top of hand engineered features to end-to-end learning of Deep Neural Networks (DNNs).

In designing deep-learning SER systems, the practitioner faces a series of important choices. The first relates to the way that speech information is represented, e.g., using MFCCs, spectrograms or other acoustic features. Then, it is crucial to identify a suitable architecture that will encode this information efficiently, with the most natural candidates being CNNs [4, 5, 6], LSTMs [7, 8], or even a combination [9, 10] of these.

Relatedly, Self-Supervised Learning (SSL) features have received a mounting interest in the last years. There is a flurry of works exploring the prospect of using wav2vec [11, 12] or HuBERT [13] features in emotion recognition [14, 15, 16, 17, 18]. On a similar vein, Conformer Applied to Paralinguistics (CAP) [19] are representations based on the Conformer [20] architecture, trained similarly to the wav2vec-2.0 (w2v2) model, with their downstream performance being evaluated on the Non-Semantic Speech Benchmark (NOSS) [21]. In [22], authors have distilled the efficient CAP representations to more lightweight architectures, aiming to significantly reduce memory and compute overheads, while sacrificing the downstream performance only slightly.

On a different note, it is critical to associate our work with prior publications, such that of Musgrave et al. [23], that aim to increase awareness about methodological flaws found in ML papers. Such works contribute in improving the experimental rigor of the field, and play a key role in realizing the potential limitations related to the reported results of each method.

3. REVISITING IEMOCAP

3.1. Dataset Details

The IEMOCAP [3] database consists of five dyadic interactions sessions, each between a unique male and female speaker per session, amounting to a total of 10 speakers. The conversation of each session is segmented based on speaker turns, and these conversational segments are annotated for perceived expressed emotions. Each segment is labelled by 3 different annotators, where they assigned both a discrete categorical label (e.g. happy, neutral, sad etc.) and a continuous valued one, assessing the valence, dominance and activation dimensions. Additionally, the conversations are a blend of both scripted and improvised speech interactions.

Although the diversity of label information present in this database enables the study of SER from multiple aspects, our focus in this paper is on the most prevalent setting found across literature. In this setting, SER is tackled as a 4-way classification problem, discarding all conversational segments whose discrete labels are not included in the following set $\{$*neutral*, *happy*, *sad*, *angry*, *excited*$\}$. Then, due to their expressive closeness, the happy and excited label classes are merged. This process leads to a dataset of $N = 5531$ samples, with a distribution of $\{1708, 1636, 1103, 1084\}$ examples for the $\{$*neutral*, *happy*, *sad*, *angry*$\}$ classes, respectively. After the filtering procedure, the dataset ends up with audio segments of average duration of ~ 4.5 seconds, leading to 7 hours of speech data, coming from 10 different speakers.

The resulting dataset comes with two major challenges: First, it contains a limited amount of speech, compared to other datasets used in SER e.g., *MSP-Podcast* [24] amounts to 27 hours. Second, the empirical label distribution is imbalanced, with most samples originating from the neutral and happy classes. Both of these challenges affect machine learning methods in various ways; the former means that models of increased capacity suffer from overfitting which should be treated by strong regularization, e.g., with multitask learning [9, 18, 15]. Failing to properly address the latter issue will probably introduce unsolicited biases to the system's outputs. Recent works propose elaborate data augmentation techniques [25, 26] or the use of different losses, e.g., the focal loss in the case of Aftab et al. [4], to combat imbalance.

3.2. Evaluation Protocol

In this section, we highlight the three main assumptions typically made by researchers during the evaluation of their SER models. This leads to a vast amount of distinct evaluation protocols, with final results that cannot be compared easily or fairly with previous methods in the literature. After that, we present a set of a evaluation guidelines, hoping to disambiguate the evaluation process and mitigate the errors made therein.

3.2.1. Speaker Dependent Evaluations

In speech related applications with multiple speakers, it is crucial to guarantee that the speakers on test data are not present in the training set. Otherwise, the system may deceptively appear to be highly performing, whereas in reality it may be the case that it only learns to exploit spurious correlations in an optimal way. For this reason, the evaluation procedure should adhere to the property of Speaker Independence (SI).

An important challenge associated with IEMOCAP is that it lacks an established nominal test set; hence, in each work researchers are left with some freedom on how to carry out evaluation. There are two prevalent ways to perform SI evaluations on IEMOCAP: (1) 5-fold cross validation (or leave-one-session-out), where in each turn four sessions (with eight speakers) are used as training data and one session (with two speakers) as test data and (2) 10-fold cross validation (or leave-one-speaker-out), where one speaker is kept for test and the other nine for training set. Usually, two speakers are preserved for validation set purposes in 5-fold cross-validation (one speaker for the 10-fold case). The per-fold accuracy metrics are aggregated through the Weighted Accuracy (WA) and Unweighted Accuracy (UA) metrics. The WA measures

#	Publication	Code	Acoustic Feats	UA (%)	WA (%)	cross-val	Comments on Eval.
1	Feng et al. [8]	Yes	MFCCs	76.4	75.5	5-fold	-
2	Zhu and Li [5]	Yes	MFCCs	73.90	73.70	5-fold	Excited as Happy
3	Xu et al. [27]	Yes	Spectrograms	67.94	67.28	5-fold	Excited as Happy
4	Gat et al. [15]	No	HuBERT	-	81.0	5-fold	-
5	Wang et al. [14]	Yes	HuBERT	-	79.58	5-fold	-
6			w2v2	-	77.47		-
7	Peng et al. [28]	No	MFCCs	79.1	78.0	10-fold	-
8	Xu et al. [6]	Yes	Spectrograms	77.54	79.34	5-fold	Improv. only, Exc. as Hap.
9	Xu et al. [27]	Yes	Spectrograms	76.36 (63.92)	76.18 (65.90)	5-fold	Improv. (Scripted) only, Exc. as Hap.
10	Zhu and Li [5]	Yes	MFCCs	79.25 (70.39)	81.18 (71.44)	5-fold	Improv. (Scripted) only, Exc. as Hap.
11	Liu and Wang [29]	No	MFCCs	78.30	79.52	5-fold	Improv. only
12	Moine et al. [30]	No	Spectrograms	77.22	-	5-fold	Improv. only

Table 1: Comparison of methods that perform Speaker Dependent (SD) evaluation. Rows 8-12 perform random 5-fold cross val. only on the improvised data.

#	Publication	Code	Acoustic Feats	UA (%)	WA (%)	cross-val
1	Pepino et al. [16]	Yes	w2v2	67.2	-	5-fold
2	Yang et al. [17]	Yes	HuBERT	67.62	-	5-fold
3	Gat et al. [15]	No	HuBERT	-	74.2	5-fold
4	Santoso et al. [7]	No	MFCC+CQT+F0	75.9	76.1	5-fold
5	Li et al. [9]	No	Spectrograms	82.8	81.6	5-fold
6	Zou et al. [10]	Yes	MFCCs,Spec,w2v2	71.05	69.80	5-fold
7				72.70	71.64	10-fold
8	Wang et al. [14]	Yes	HuBERT	-	73.01	10-fold
9			w2v2	-	70.99	
10	Aftab et al. [4]	Yes	MFCCs	70.76	70.23	10-fold
11	Feng et al. [8]	Yes	MFCCs	69.67	68.63	10-fold
12	Cai et al. [18]	Yes	w2v2	-	78.15	10-fold
13	Shor et al. [19]	No	CAP	-	79.2	Session 05
14	Shor and Venugopalan [22]	Yes †	TRILLSSON	-	73.2	Session 05

Table 2: Comparison of methods that perform Speaker Independent (SI) evaluation. The papers of rows 13,14 use only Session05 (i.e. speakers '05M' and '05F') as test set. (†): They only provide the trained weights of the feature extractors.

the percentage of correct predictions, whereas UA averages the recall metric for each class.

Intriguingly, the necessity of performing SI evaluation is yet to become clear in the community, since many recent works randomly create the train-val-test split with overlapping speaker identities. This, in contrast to SI, is called Speaker Dependent (SD) evaluation. Besides the obvious violation of speaker independence, the random split limits the possibility of comparisons only to the very specific baselines used in the respective works, and limits generalization.

3.2.2. *Evaluation on Improvised Interaction Data*

In our previous analysis, we mentioned that IEMOCAP contains speech from both improvised and scripted interactions. Another common evaluation trend is to assess performance on only one of these two components. This frequently happens with the improvised component, e.g., [30, 6]. While this choice may be linked with the unique goals of each paper, it should be noted that this practice may again impose limits on our ability to contextualize results with other published methods. A desirable solution to this challenge would be to perform – and make available – evaluation on both improvised and scripted data, and report results about each data part separately.

3.2.3. *The 'excitement' class case*

Here, we discuss another widely-adopted choice which leads to the existence of a new evaluation branch that limits broad comparisons. There are some early works [31, 32] which only consider the {*neutral*, *sad*, *angry*, *excited*} classes, excluding the *happy* class from their study (instead of merging it with the *excited* class). This, in turn, had prompted some subsequent papers [27, 6, 5] to conduct evaluation following this protocol. Similarly with the two aforementioned evaluation challenges, we believe that researchers should follow the convention of including the merged *happy* and *excited* classes, and then include comparisons with specific baselines that, for example, use only the *excited* class.

3.2.4. Comparing literature results

We collect the reported results of multiple recent SER methods developed on the IEMOCAP database. In Tables 1 and 2, we demonstrate the reported performance metrics for works that use SD and SI evaluations, respectively. We also include additional information, e.g., how the speech signal is represented in each work, whether the respective paper has an official implementation or any other choices followed during evaluation. In the SD case, the results cannot be compared fairly since data are split randomly (except the case that some work evaluates other methods in the exact same partition approach, e.g., [5] outperforms [27, 6]). Overall, it becomes evident that diverging from (and absent) a common evaluation setting can result in an unclear situation and confusion about where each method stands within the published literature, and can impede progress as a community.

3.3. Recommended evaluation guidelines

Our proposal for future work on SER using IEMOCAP is to perform evaluations according to the following protocol to ensure a common minimally-viable and comparable baseline:

- Use *neutral, sad, angry* and *happy+excited* classes, leading to a dataset of $N = 5531$ samples,
- Perform a 10-fold Speaker Independent cross-validation with one speaker as test, eight as training and one as validation set,

After reporting the results following this set up, researchers may deviate as they wish, e.g., with reporting results on the improvised part of IEMOCAP or other label subsets/merges. Numerous methods [10, 14, 4, 18] have been evaluated according to this setting. For the number of folds during cross-validation, we remark that both 5-fold and 10-fold evaluations are equally reasonable choices, with the latter case using more training data per fold, hence leading to slightly higher WA and UA metrics.

4. EXPERIMENTS TO EXAMINE REPRODUCIBILITY

In this section, we conduct a reproducibility study on papers that had open-sourced their implementations. Our goal is to evaluate each method (whose implementation is public) from Tables 1 and 2 according to the evaluation guidelines of 3.3.

First, we remark that, despite uploading their code, the majority of researchers omit to include checkpoints of the trained models. Notably, only two of the examined works [18, 22] share trained weights, which makes the reproduciblity process much easier. However, the authors of [22] only share the weights of the backbone, so one has to extract the train-val-test embeddings and train a linear layer on top of them to reproduce results. Feng et al. [8] have uploaded a codebase that produced errors during execution, so we could not proceed further. Aftab et al. [4] already evaluate using the recommended guidelines, and we confirm that their reported results are reproducible, after retraining their model from scratch. Similarly, while both Zou et al. [10] and Wang et al. [14] do not share checkpoints, training from scratch confirms their reported results. We did not check the reproduciblity of Yang et al. [17] and Pepino et al. [16], since their works are nearly identical with [14].

Next, we turn our attention to the works of Table 1. Zhu and Li [5] propose the GLAM architecture, but their public code does not include checkpoints. Training from scratch, according to their protocol (i.e., drop the *happy* class' samples) yields nearly identical results to the reported ones. However, when we attempt to retrain the model by including the data from the closely related *happy* class, we obtain a significant performance drop. This is evidence that a simple assumption (here, to dismiss *happy* examples) can lead to an incomplete view about a model's effectiveness. The GLAM architecture outperforms both models from [27, 6], which follow the exact same unconventional protocol. The available code of [27, 6] was not executable or maintained.

In Table 3, we summarize the findings from our reproducibility study, where we report the WA metric measured according to the protocol described in 3.3.

#	Publication	Pretrained	Reproducibility	WA (%)
1	Cai et al. [18]	✓	✓	78.15
2	Shor and Ven. [22]	✓	✓	68.05
3	Aftab et al. [4]	✗	✓	71.43
4	Zou et al. [10]	✗	✓	69.2
5	Wang et al. [14]	✗	✓	69.61
6	Zhu and Li [5]	✗	✗	63.82
7	Feng et al. [8]	✗	✗	-
8	Xu et al. [27]	✗	✗	-
9	Xu et al. [6]	✗	✗	-

Table 3: Reproducibility experiments. Pretrained column indicates whether the implementations share trained weights.

5. CONCLUSIONS

In this paper, through an empirical case review and analysis of published work of SER systems based on IEMOCAP, we investigated the choices made in the design and evaluation of SER systems through a lens of ease of comparing performance, generalizabilty and reproducibility. We encourage future works to design and evaluate their methods according to a common protocol (such as that suggested based on well executed studies). The lack of which can hinder our ability to fairly compare and reproduce results of different methods, and make collective advance as a community. In the future, we hope this case study will lead to the establishment of a well-defined set of ”good practices” while designing, building and releasing data and evaluation benchmarks for SER.

6. REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. S. Narayanan, "Paralinguistics in speech and language–state-of-the-art and the challenge," *Computer, Speech, and Language*, vol. 27, no. 1, pp. 4–39, jan 2013.

[2] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, sep 2009.

[3] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[4] A. Aftab *et al.*, "Light-sernet: A lightweight fully convolutional neural network for speech emotion recognition," in *ICASSP 2022*. IEEE, 2022, pp. 6912–6916.

[5] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *ICASSP 2022*. IEEE, 2022, pp. 6437–6441.

[6] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP 2021*. IEEE, 2021, pp. 6319–6323.

[7] J. Santoso *et al.*, "Speech Emotion Recognition Based on Attention Weight Correction Using Word-Level Confidence Measure," in *Proc. Interspeech 2021*, 2021, pp. 1947–1951.

[8] H. Feng, S. Ueno, and T. Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *Proc. Interspeech 2020*, 2020, pp. 501–505.

[9] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proc. Interspeech 2019*, 2019, pp. 2803–2807.

[10] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022*. IEEE, 2022, pp. 7367–7371.

[11] S. Schneider *et al.*, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[12] A. Baevski *et al.*, "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.

[13] 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.

[14] 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.

[15] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *ICASSP 2022*. IEEE, 2022, pp. 7342–7346.

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

[17] S. wen Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[18] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech 2021*, 2021, pp. 4508–4512.

[19] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022*. IEEE, 2022, pp. 3169–3173.

[20] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[21] J. Shor *et al.*, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[22] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," *arXiv preprint arXiv:2203.00236*, 2022.

[23] K. Musgrave, S. J. Belongie, and S. Lim, "A metric learning reality check," in *ECCV 2020*.

[24] 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.

[25] A. Chatziagapi *et al.*, "Data Augmentation Using GANs for Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 171–175.

[26] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021*. IEEE, 2021, pp. 6324–6328.

[27] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 1058–1064.

[28] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *ICASSP 2021*. IEEE, 2021, pp. 3020–3024.

[29] J. Liu and H. Wang, "A Speech Emotion Recognition Framework for Better Discrimination of Confusions," in *Proc. Interspeech 2021*, 2021, pp. 4483–4487.

[30] C. L. Moine, N. Obin, and A. Roebel, "Speaker Attentive Speech Emotion Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2866–2870.

[31] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[32] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-Attention for Speech Emotion Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2578–2582.