Investigating HuBERT-based Speech Emotion Recognition Generalisation Capability*

Letian Li[1,2], Cornelius Glackin[2], Nigel Cannings[2],
Vito Veneziano[1], Jack Barker[2], Olakunle Oduola[2],
Chris Woodruff[2], Thea Laird[2], James Laird[2], and Yi Sun[1]

[1] School of Physics, Engineering and Computer Science,
University of Hertfordshire, Hatfield, UK
y.2.sun@herts.ac.uk
[2] Intelligent Voice Ltd, London, UK
{letian.li,neil.glackin}@intelligentvoice.com

Abstract. Transformer-based architectures have made significant progress in speech emotion recognition (SER). However, most published SER research trained and tested models on data from the same corpus, resulting in poor generalisation ability to unseen data collected from different corpora. To address this, we applied the HuBERT model to a combined training set consisting of five publicly available datasets (IEMOCAP, RAVDESS, TESS, CREMA-D, and 80% CMU-MOSEI) and conducted cross-corpus testing on the Strong Emotion (StrEmo) Dataset (a natural dataset collected by the authors) and two publicly available datasets (SAVEE and 20% CMU-MOSEI). Our best result achieved an F1 score of 0.78 over the three test sets, with an F1 score of 0.86 for StrEmo specifically. Additionally, we are pleased to release the spreadsheet of key information on the StrEmo dataset as supplementary material to the conference.

Keywords: Speech emotion recognition · HuBERT · StrEmo · Generalisation ability · Cross-corpus testing.

1 Introduction

In recent years, there has been a growing interest in Speech Emotion Recognition (SER) as it plays a crucial role in analyzing human-to-human conversations and enabling effective human-to-machine interactions [25,8]. This technology empowers machines to comprehend and respond to human emotions aptly [26].

There is now a trend of using large foundation models, which are pre-trained on extensive datasets and then fine-tuning them for downstream tasks, which is replacing specialized architectures created for specific tasks [4]. Such large foundation models are recently dominating in many Artificial Intelligence (AI) fields, such as SimCLR [7] in Computer Vision (CV), BERT [9] in Natural Language Processing (NLP), and wav2vec 2.0 [2] in speech processing.

* Supported by Innovate UK.

HuBERT [14] model is one of the Transformer-based architectures pre-trained by self-supervised methods. This BERT-like model follows the architecture of wav2vec 2.0 [2], but either matches or surpasses wav2vec 2.0 performance [14,27].

This study applied the HuBERT model on six public datasets (RAVDESS [19], CREMA-D [6], TESS [23], SAVEE [13], IEMOCAP [5] and CMU-MOSEI [3]) and one natural dataset collected by the authors, named the Strong Emotion (StrEmo) dataset, to investigate the model's generalisation ability in SER. To combine the different corpora, it was necessary to unify the different emotion labels from the aforementioned datasets into three categories: positive, neutral, and negative. Furthermore, we are pleased to release the key information of the StrEmo dataset proposed in this work. In accordance with privacy policies, we will publicly release the information that was used and obtained during the collection of this dataset in the form of a spreadsheet containing YouTube video IDs and segment timings. The StrEmo dataset is available at https://github.com/IntelligentVoice/IV_SER_Data.

2 Related Work

Speech Emotion Recognition (SER) typically involves two phases: feature extraction and feature classification [18]. Researchers in speech processing have developed various features, including continuous features, qualitative features, and spectral features such as Linear Predictor Coefficients (LPC) and Mel-Frequency Cepstral Coefficients (MFCC) [10]. In the second phase, researchers traditionally applied machine learning methods, such as the Maximum Likelihood Principle, Support Vector Machine, and Decision Trees, for feature classification.

Over the past decade, deep learning has emerged as a rapidly advancing field of research, owing to its multi-layered structure that allows for the efficient processing of complex data and the delivery of high-quality results [1,22]. Khalil et al. [17] conducted a review of deep learning approaches in SER, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. They deduced that deep learning algorithms outperform traditional techniques in emotion recognition.

Researchers have applied different deep learning techniques in recent years to improve the accuracy of SER. In 2018, Etienne et al. [11] designed a CNN and LSTM architecture for SER on the IEMOCAP dataset [5] and obtained results of 64.5% for weighted accuracy and 61.7% for unweighted accuracy on four emotions. Georgescu et al. [12] implemented ResNet-18 and PyNADA to work on CREMA-D [6] in 2020. Luna-Jiménez et al.[20] achieved their best results using the CNN-14 of the PANNs framework on RAVDESS dataset [19] in 2021. In 2022, Ye et al. [28] proposed a temporal emotional modeling approach for SER, termed Temporal-aware bI-direction Multi-scale Network (TIM-Net). With the recent rise in popularity of foundation models [27,4], Shreyah et al. [15] compared the performance of an AlexNet-based CNN model with a Transformer-based model, wav2vec 2.0, on speech emotion datasets.

Most of the previous studies in this field either trained and tested the model on the same dataset [11,12,20,15] or did not use a Transformer-based foundation model [11,12,20,28]. This study aims to combine seven datasets [19,5,6,23,13,3] and perform cross-corpus testing using the advanced Transformer-based foundation model, HuBERT [14], to examine its generalisation ability in SER. Additionally, this research introduces a new natural dataset (StrEmo) collected by the authors, which is used to verify the generalisation ability of the model.

3 Description of Datasets

This study utilizes seven datasets, including six publicly available datasets for speech emotion recognition research. These datasets can be categorized as acted datasets (RAVDESS [19], CREMA-D [6], TESS [23], SAVEE [13]), elicited dataset (IEMOCAP [5]), and natural datasets (CMU-MOSEI [3]). In addition, we introduce a new natural dataset (StrEmo) collected from YouTube videos.

3.1 Publicly Available Datasets

Utterances in six publicly available datasets were spoken in English. Each dataset includes various detailed emotion categories. For instance, in the RAVDESS dataset, there are eight emotions: neutral, happiness, sadness, fear, anger, disgust, surprise, and calm. The duration of each audio clip varies, with an average of 2 seconds in the TESS dataset and ranging up to an average of 8 seconds in the CMU-MOSEI dataset. The descriptions of datasets are presented in Table 1.

Table 1. Descriptions of six publicly available datasets.

Dataset	**Number of Utterances**	**Number of Emotions**	**Average Duration (seconds)**	**Category**
CMU-MOSEI	23500	9	8.0	Natural
IEMOCAP	10093	9	4.4	Elicited
RAVDESS	1440	8	3.7	Acted
CREMA-D	7442	6	2.5	Acted
TESS	2800	7	2.0	Acted
SAVEE	480	7	3.9	Acted

3.2 StrEmo

We gathered a new dataset called the Strong Emotion (StrEmo) Dataset to further assess the model's generalisation performance. The data were collected from English YouTube videos using only the audio signal and were labeled by the authors into three emotional categories: 180 happiness, 94 neutral, and 222 anger emotions. The audio clips have an average duration of approximately 5 seconds. We have released a spreadsheet for this dataset that includes information on the emotion label, the YouTube video ID, the time points of clip segments, as well as audio characteristics such as spoken language, speaker gender, ethnicity, accent, and age of the data.

4 Methodology

In this study, we utilize the HuBERT [14] model for SER. To ensure balanced data across emotion labels in the training set, we employ the undersampling technique. Additionally, we use cross-validation to determine the optimal hyperparameters.

4.1 HuBERT Model

The Hidden-Unit BERT (HuBERT) [14] approach is a self-supervised speech representation learning method that utilizes offline clustering to provide aligned target labels for a BERT-like prediction loss. It applies the prediction loss over the masked regions only and relies on the consistency of unsupervised clustering. The model either matches or improves upon wav2vec 2.0 performance on several benchmarks with various fine-tuning subsets. Facebook proposed five types of HuBERT models, and the large version, which was pretrained on Libri-Light [16] and fine-tuned on Librispeech [21], was used in this study.

4.2 Undersampling

Various techniques can be used to address class imbalance issues, including undersampling and oversampling. Undersampling involves reducing the size of the majority class, while oversampling involves increasing the data from the minority class. In this study, after merging seven datasets and splitting them into training and test sets, the training set had an imbalanced distribution of three emotion labels: 11975 (negative), 5198 (neutral), and 4149 (positive). To balance the data volume of each emotion classification, the undersampling method was used, resulting in 5198 samples for negative and neutral emotions, and 4149 samples for positive emotion.

4.3 Cross-Validation

In this study, we performed hyperparameter experiments for HuBERT training using 5-fold cross-validation. The hyperparameters typically include learning rate, batch size, and epoch. However, since we used the early stop approach in the final training, the epoch was not explored as a hyperparameter. Due to GPU memory limitations, the maximum batch size we could use was 2, so batch size was not included in the hyperparameter experiments. Therefore, we only studied the effect of learning rate on the model's performance, testing values of 1×10^{-4}, 1×10^{-5}, and 1×10^{-6}. Since 1×10^{-5} is the default value for many frameworks, we specifically tested this value along with the others.

4.4 Experiment Procedure

The experiment included the following eight steps:

1. Loading and combining the data from all datasets (StrEmo, RAVDESS [19], CREMA-D [6], TESS [23], SAVEE [13], IEMOCAP [5], CMU-MOSEI [3]).
2. Relabeling the data into three emotion categories - positive (includes "Happiness", "Excitement" and "Pleasant Surprise"), negative (includes "Anger", "Sadness", "Fear", "Disgust", "Frustration" and "Disappointment") and neutral (includes "Neutral" and "Calm"). All other emotions (includes "Surprise" and "Other") were discarded because they were not as directly classifiable into these three categories.
   In this step, we implemented additional preprocessing on the CMU-MOSEI dataset to improve data quality. Firstly, we used Whisper [24] to transcribe the audio from the CMU-MOSEI dataset into text. Next, we used the SequenceMatcher in a python module named "difflib" to compare the obtained text with the original text provided by the dataset and calculated a similarity score. We only selected data with a similarity score greater than 0.85. Since the CMU-MOSEI dataset is a multi-emotion dataset, each data point has emotion scores for all emotions. We selected data with either i) exclusively positive or negative emotion scores greater than 2 or ii) all emotion scores equal to 0 (indicating neutral data) for experimentation. This preprocessing allowed us to obtain high-quality data, but the size of the dataset was drastically reduced to 3071. After relabeling the emotions and selecting only positive, negative, and neutral data from all datasets, Table 2 shows the number of audio files used in this study.
3. Splitting all the data into a training set (IEMOCAP, RAVDESS, TESS, CREMA-D and 80% CMU-MOSEI) and a test set (StrEmo, SAVEE and 20% CMU-MOSEI). Since CMU-MOSEI is a relatively complex dataset for SER, we used it in both the training and test sets.
4. Balancing the training set by undersampling the data (Section 4.2).
5. Performing 5-fold cross-validation on the training set to determine the optimal learning rate (Section 4.3).
6. Retraining the model on the entire training set using the best learning rate obtained from step 5.

Table 2. Distribution of emotions across datasets after preprocessing.

Dataset	**Positive**	**Neutral**	**Negative**	**Category**
StrEmo	180	94	222	Natural
CMU-MOSEI	298	2157	616	Natural
IEMOCAP	1636	1708	4078	Elicited
RAVDESS	192	288	768	Acted
CREMA-D	1271	1087	5084	Acted
TESS	800	400	1600	Acted
SAVEE	60	120	240	Acted

Table 3. Cross-validation on different learning rate (LR). Kx denotes the xth fold out of five folds and Avg. the average value over Five folds.

LR	**K1**	**K2**	**K3**	**K4**	**K5**	**Avg.**
1×10^{-4}	0.40	0.94	0.82	0.53	0.99	0.74
1×10^{-5}	0.94	0.97	0.95	0.97	0.99	0.96
1×10^{-6}	0.68	0.76	0.72	0.82	0.88	0.77

7. Testing the fine-tuned model on the unseen test set to evaluate its generalisation performance.
8. Calculating the F1 score as the evaluation metric for the test results.

5 Experiments and results

To investigate the generalisation capability of the HuBERT model in speech emotion recognition, we conducted an experiment using seven datasets. The datasets were divided into a training set (consisting of IEMOCAP, RAVDESS, TESS, CREMA-D, and 80% of CMU-MOSEI) for cross-validation and training purposes, and a test set (consisting of StrEmo, SAVEE, and 20% of CMU-MOSEI) for evaluation. By splitting the dataset in this way, the StrEmo and SAVEE datasets were completely new corpora to the SER model and were used to assess the model's generalisation ability.

5.1 Results of Cross-Validation on Learning Rate

According to Table 3, the cross-validation results demonstrate that a learning rate that is either too large or too small cannot achieve good performance. The optimal learning rate for this training was found to be 1×10^{-5}.

5.2 Results on the Combined Dataset

After we found out the best learning rate from cross-validation, we retrained the HuBERT model on the whole training set with the hyperparameters. Figure 1 is the confusion matrix we computed after we tested the fine-tuned model on the unseen test set. The overall F1 score is 0.73. In addition, F1 scores of unseen

Table 4. Experimental results of unseen test set for model trained with balanced data.

	Precision	**Recall**	**F1 Score**
Negative	0.84	0.56	0.67
Neutral	0.70	0.98	0.82
Positive	0.72	0.60	0.66
Weighted Avg.	0.76	0.74	0.73

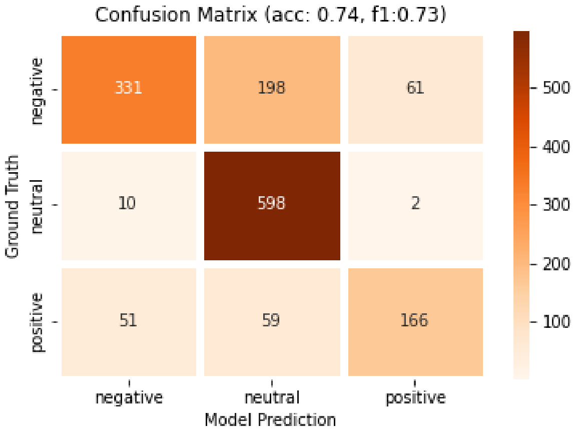

Fig. 1. Confusion matrix of test results for model fine-tuned with balanced data.

test in three emotion categories are shown in Table 4. They are 0.67 (Negative), 0.82 (Neutral) and 0.66 (Positive) respectively.

5.3 Detailed Results on the StrEmo Dataset

In order to verify that the model's generalisation ability has been improved after multi-datasets training, we compared the prediction performance of models with different fine-tuning degrees on StrEmo dataset. From Table 5, it can be seen that without fine-tuning, the original HuBERT model [14] predicted all the data as positive emotion. After fine-tuning only on RAVDESS, the weighted average F1 score improved to 0.8. With the fine-tuning on the balanced training set from the combination of IEMOCAP, RAVDESS, TESS, CREMA-D and 80% CMU-MOSEI, the F1 score slightly increased to 0.82. Finally, we tried the best model fine-tuned with unbalanced training set of the combination of the five datasets, the F1 score continued to improve to 0.86.

Table 5. F1 scores on StrEmo dataset for the models with different fine-tuning degrees. Com-Bal denotes balanced training set from the combination of IEMOCAP, RAVDESS, TESS, CREMA-D and 80% CMU-MOSEI. Com-Unbal denotes unbalanced training set from the combination datasets.

Fine-Tuned	**F1.POSITIVE**	**F1.NEUTRAL**	**F1.NEGATIVE**	**F1**
None	0.53	0	0	0.19
RAVDESS	0.81	0.69	0.84	0.8
Com-Bal	0.79	0.93	0.81	0.82
Com-Unbal	0.83	0.93	0.86	0.86

Table 6. F1 score for each unseen test dataset.

Training Set	CMU-MOSEI	StrEmo	SAVEE	Overall
Balanced	0.69	0.82	0.62	0.73
Unbalanced	0.77	0.86	0.68	0.78

5.4 Discussion

Model Prediction Trend against Emotion According to the F1 scores shown in Table 4, the model performed the best in predicting neutral emotion. However, when examining the confusion matrix (Figure 1), it indicates that the model is more likely to classify negative emotions as neutral.

The Impact of Data Volume From Table 6, it can be observed that training the model with an unbalanced training set yields better performance, as indicated by a higher F1 score of 0.78. We considered that the reason was that we used the undersampling method to balance training set, which led to less training data. A further conclusion is that training with more data can lead to higher performance of the model.

As mentioned in Section 4.2, besides undersampling, oversampling is another method to deal with the imbalanced dataset. Unlike undersampling, where some data is lost, oversampling allows more data to be used. We plan to use data augmentation in future studies to conduct oversampling and to further improve the performance of the model.

Effect of emotional strength on model performance Table 6 also shows the F1 scores obtained on different datasets in testing. We can see that the StrEmo dataset achieved the highest F1 score, which was approximately 20% higher than that of the SAVEE dataset. This can be attributed to the stronger emotional expression in the StrEmo dataset, making it easier for the model to predict. As the SAVEE dataset is an actor simulation dataset, the expression of emotion is not as apparent, resulting in greater difficulty for the model to make accurate predictions.

6 Conclusions

In this study, we evaluated the generalisation ability of the HuBERT model in speech emotion recognition using six publicly available datasets and a natural dataset collected by ourselves. Notably, the test set in this research comprised data from corpora that differed from the training corpora. Our findings demonstrated that the model's performance improved as we fine-tuned it with additional data. The best F1 score achieved on the unseen test set was 0.78, and on our collected dataset StrEmo, it was 0.86. Moreover, we observed that the fine-tuned model was more effective in analyzing stronger emotional data. However, it was noted that the fine-tuned model had a higher tendency to misclassify negative emotion as neutral, indicating a need for further investigation with more data in the future.

7 Acknowledgements

This project has received funding from the Innovate UK grant via the Knowledge Transfer Partnership programme, KTP No.11039.

References

1. Abbaschian, B.J., Sierra-Sosa, D., Elmaghraby, A.: Deep learning techniques for speech emotion recognition, from databases to models. Sensors **21**(4), 1249 (2021)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
3. Bagher Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
5. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
6. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal processing magazine **18**(1), 32–80 (2001)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
10. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern recognition **44**(3), 572–587 (2011)
11. Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., Schmauch, B.: CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. In: Proc. Workshop on Speech, Music and Mind (SMM 2018). pp. 21–25 (2018)

12. Georgescu, M.I., Ionescu, R.T., Ristea, N.C., Sebe, N.: Non-linear neurons with human-like apical dendrite activations. arXiv preprint arXiv:2003.03229 (2020)
13. Haq, S., Jackson, P., Edge, J.: Audio-visual feature selection and reduction for emotion classification. In: Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia (Sept 2008)
14. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
15. Iyer, S., Glackin, C., Cannings, N., Veneziano, V., Sun, Y.: A comparison between convolutional and transformer architectures for speech emotion recognition. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
16. Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al.: Libri-light: A benchmark for asr with limited or no supervision. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7669–7673. IEEE (2020)
17. Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T.: Speech emotion recognition using deep learning techniques: A review. IEEE Access **7**, 117327–117345 (2019)
18. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. International journal of speech technology **15**, 99–117 (2012)
19. Livingstone, S.R., Russo, F.A.: 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 **13**(5), e0196391 (2018)
20. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors **21**(22), 7665 (2021)
21. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
22. Pandey, S.K., Shekhawat, H.S., Prasanna, S.M.: Deep learning techniques for speech emotion recognition: A review. In: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA). pp. 1–6. IEEE (2019)
23. Pichora-Fuller, M.K., Dupuis, K.: Toronto emotional speech set (TESS) (2020)
24. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022)
25. Schuller, B.W.: Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM **61**(5), 90–99 (2018)
26. Trabelsi, A., Frasson, C.: The emotional machine: A machine learning approach to online prediction of user's emotion and intensity. In: 2010 10th IEEE International Conference on Advanced Learning Technologies. pp. 613–617. IEEE (2010)
27. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Eyben, F., Schuller, B.W.: Dawn of the transformer era in speech emotion recognition: closing the valence gap. arXiv preprint arXiv:2203.07378 (2022)
28. Ye, J., Wen, X., Wei, Y., Xu, Y., Liu, K., Shan, H.: Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. arXiv preprint arXiv:2211.08233 (2022)