Enhancing HuBERT's Speech Emotion Recognition through Data Augmentation and Fine-tuning

Avital Finanser
finanser@post.bgu.ac.il

Sivan Raviv
sivanrav@post.bgu.ac.il

Abstract

Recent advances in pre-trained speech models have significantly improved speech emotion recognition (SER) performance compared to traditional methods. Despite these advances, deploying these large pre-trained architectures in varied domains or under noisy conditions remains challenging, especially when labeled emotion data is limited. This study enhances a pre-trained HuBERT model by exploring four data augmentation techniques: Time Stretch, Additive Noise, SpecAugment, and Neutral CopyPaste. After establishing a baseline on the RAVDESS dataset, we train separate models on each augmentation method, followed by probability-weighted combinations to identify the best-performing approaches. Our results show that SpecAugment delivers the largest gains on clean data, while Time Stretch and Neutral CopyPaste yield domain-specific improvements for noise-perturbed or partial-audio scenarios. Despite combining augmentations, cross-validation on the SAVEE corpus reveals a consistent performance drop, indicating difficulty adapting to unseen speaker demographics and recording conditions. We conduct emotion-specific analyses highlighting complementary benefits from frequency and time-based augmentations, yet note that none fully address cross-corpus variability. These findings demonstrate the potential of targeted augmentation strategies to bolster SER robustness and guide future research toward more adaptive methods for real-world deployments.

Github Repository: https://github.com/AvitalFinanaser/Augmented-HuBERT-for-SER.git.

Index Terms — Speech emotion recognition, self-supervised learning, HuBERT, data augmentation.

1 Introduction

Speech is the most natural way for humans to communicate, conveying much more than just words. It includes elements such as intonation, tone, emotional cues, and the speaker's identity, all vital for advanced-machine interaction systems. In this context, three core speech-related tasks - Spoken Language Understanding (SLU), Speaker Verification (SV), and Speech Emotion Recognition (SER) - play pivotal roles. While SLU focuses on understanding the semantic content of speech and SV on speaker identification, SER aims to detect the emotions conveyed in an audio signal.

SER has attracted increasing attention in affective computing and human-computer interaction, owing to its potential for creating responsive and empathetic systems [1]. Applications range from mental health monitoring to customer service automation, where automatic emotion recognition enhances user experience and service quality. Despite its advantages, SER still encounters significant challenges in diverse acoustic environments, a shortage of labeled data, and the complexities of human emotional expression [2]. Recent advances in self-supervised learning have produced models such as wav2vec2 [3] and HuBERT [4], which, in turn, have shown significant improvements over traditional feature-based approaches in various speech recognition tasks. However, these approaches are still challenged by real-world variability in speech signals, including background noise, pitch fluctuations, and temporal distortions.

An interesting strategy for mitigating these gaps is to employ data augmentation. This involves enhancing training datasets with synthetic variations that maintain semantic content while exposing the model to various environmental conditions [5].

In this project, we examine how four data augmentation methods - Time Stretch, Additive Noise, SpecAugment, and Copy Paste - can be leveraged to enhance the robustness of a pre-trained HuBERT-based SER model that extracts features using WAV2VEC2 from WAV audio. We evaluate various augmentation techniques individually and in probability-weighted combinations to determine which configurations and compositions yield optimal performance gains. We conduct experiments using the RAVDESS dataset [6], [**?**] for training and primary evaluation, while cross-validation occurs with the SAVEE dataset [7] to assess generalization. By analyzing these augmentations, we aim to identify strategies for developing a robust Speech Emotion Recognition (SER) model that effectively retains essential semantic and emotional information in real-world scenarios.

2 Related Work

2.1 Evolution of SER

Traditional Approaches Early approaches to Speech Emotion Recognition (SER) relied on handcrafted features, including Mel-Frequency Cepstral Coefficients (MFCC), pitch contours and prosodic markers [1]. These features were then fed into traditional machine learning models, such as Hidden Markov Models (HMM) and Support Vector Machines (SVM), to classify emotions [8]. Despite moderate success on controlled datasets, these methods often failed to generalize beyond their narrow training conditions, due to emotional speech complexity as well as the limited handcrafted features scope.

Deep Learning-Based SER and Transformer Models Deep neural networks advanced SER by automating feature extraction and capturing longer temporal dependencies. Convolutional and recurrent architectures, particularly hybrid CNN-RNN models, reduced reliance on expert-crafted features and improved performance on benchmark tasks [9]. Attention mechanisms further enhanced these models by highlighting the most salient emotional cues. However, the scarcity of sufficiently large and balanced emotion-labeled datasets has limited their robustness and generalization to real-world conditions.

2.2 HuBERT Model

The Hidden-Unit BERT (HuBERT) model [4] is a self-supervised learning framework that effectively addresses three key challenges in speech representation: multiple sound units per utterance, no predefined lexicon during pre-training, and variable-length units without explicit segmentation.

HuBERT integrates offline clustering with a BERT-inspired masked prediction paradigm, processing audio waveforms and learning to predict cluster assignments for masked regions.

The framework's architecture comprises a convolutional waveform encoder that extracts acoustic features, a transformer-based encoder that processes partially masked features, and a projection layer that maps encoded representations to cluster predictions. HuBERT's iterative refinement mechanism, where improved representations yield enhanced clusters, subsequently leading to better representations, enables exceptional performance in ASR applications even with constrained labeled datasets.

These sophisticated representations capture nuanced speech characteristics beyond mere phonetic content, making them particularly valuable for tasks requiring deep semantic understanding. The model's ability to extract robust features from limited data and its effectiveness in preserving subtle speech attributes offer significant potential for enhancing performance in emotionally laden speech contexts.

2.3 Data Augmentation Strategies for SER

One of the most effective strategies for mitigating data scarcity and enhancing model robustness in Speech Emotion Recognition (SER) is data augmentation. Techniques such as time stretch, additive noise, and SpecAugment introduce variability in both speaking rate and vocal timbre, thereby improving model generalization across diverse recording environments.

While they are often individually applied to SER, research remains limited to combining multiple augmentation techniques. Recent studies indicate selective combinations can outperform indiscriminate application of numerous transformations, as over-augmentation risks distorting emotional cues [10]. Careful calibration is crucial because aggressive modifications may obscure subtle affective markers [11].

Furthermore, many SER datasets' small and imbalanced nature exacerbates augmentation challenges compared to larger ASR corpora [12]. Despite reported gains within single datasets, cross-dataset transferability between resources like RAVDESS and SAVEE remains largely unexplored [6] [7]. Moreover, applying these augmentation strategies to self-supervised models, particularly HuBERT presents an underexplored direction for enhancing emotion recognition robustness. The following section details our methodology for addressing these research gaps.

3 Methodology

Our approach follows a *structured five-stage experimental framework* to enhance SER performance using HuBERT models. We begin by establishing a baseline model, followed by implementing four data augmentation techniques. Subsequently, we train models separately on clean and augmented data for each method. Then, we assess the effectiveness of combining these techniques and conduct cross-validation with an additional dataset while running rigorous evaluations across multiple conditions for each model.

Stage 1: Baseline Model Establishment

Initially, we attempted to adapt a pre-trained HuBERT model for the speech emotion recognition (SER) task by integrating a classifier layer. However, achieving satisfactory performance required a significantly larger dataset, which conflicted with our emphasis on augmentation methods. Consequently, we shifted our approach to utilize a pre-trained HuBERT for SER [13]. This model originally targeted four emotions only; thus, we modified the classification head to support seven distinct emotion categories: neutral, happy, sad, angry, fear, disgust, and surprise. Then, we trained the model on the RAVDNESS dataset (4.1) to optimize its weights across all emotions. Next, we evaluated performance using a classification report that provided accuracy and the f1-weighted metrics to assess relative performance among the emotions. Additionally, we analyzed the confusion matrix to identify the model's strengths and weaknesses in emotion differentiation.

Stage 2: Data Augmentation Methods

We present strategic data augmentation techniques aimed at enhancing the diversity and robustness of our training dataset while maintaining the core emotional content.

Time Stretch. Modifies the speaking rate by stretching or compressing the temporal axis of an audio signal while preserving the original spectral characteristics. This reflects realistic fluctuations in speech tempo caused by factors such as stress, urgency, or fatigue, all of which can influence emotional expression. The stretch parameter (called rate) governs how much the audio is sped up or slowed down. This allows the model to learn from natural variations in speech rate.

Additive Noise. This technique simulates environmental conditions by introducing controlled background noise into the speech waveform, reflecting scenarios where speech occurs amidst ambient sounds, such as traffic or crowd chatter. The primary parameter is the signal-to-noise ratio (SNR), indicating the noise intensity relative to the clean speech signal. Training on moderately noisy samples enhances the model's robustness to real-world acoustic variability while preserving the intelligibility of emotional cues.

SpecAugment. Originally developed for speech recognition, modifies the log-mel spectrogram rather than the raw waveform by three parameters: **Time Warping** (Slightly stretches or compresses time segments), **Frequency Masking** (Hides consecutive frequency bands) and **Time Masking** (Blocks consecutive time frames).

By creating these localized perturbations, the model learns to fill in missing information and adapt to acoustic variations, which are subtle spectral patterns.

Neutral CopyPaste (N-CP). As demonstrated by [14], humans naturally perceive entire utterances as emotional, even when they contain only brief emotional segments. At the same time, computational models lack this intuitive capability due to their reliance on statistical patterns, potentially introducing biases. This method enhances speech emotion recognition by integrating emotional segments into neutral utterances. In our implementation, N-CP divides neutral utterances into three equal parts, replacing only the middle segment with emotional content while preserving neutral portions at the beginning and end. This approach creates balanced training examples that help the model distinguish between neutral and emotional speech samples where emotions appear in mixed contexts, simulating scenarios where brief emotional cues influence overall perception.

Stage 3: Single-Method Model Training

In this phase, we develop and train distinct models for each augmentation technique to isolate their respective impacts and assess their efficacy compared to the baseline model. Each experimental iteration begins with a fresh pre-trained HuBERT-SER model adjusted for 7 emotions, mirroring the foundation established for the baseline model training. The augmented datasets are systematically combined with the clean data in a 1:1 ratio, ensuring that each sample experiences a single augmentation featuring randomly selected parameter values, thereby forming a training set. The training parameters are consistent to facilitate an equitable comparison across all experiments (as detailed in 4.4). Finally, we evaluate each model on the predefined clean and augmented data test sets.

Stage 4: Combined Augmentation Models

Based on the performance analysis of individual augmentations in the previous step, we develop combined models that employ probability-weighted augmentation. We are motivated by combining multiple augmentation methods to construct a more robust model that captures diverse acoustic features. In order to determine the optimal combination - including which methods to combine, the optimal number, and the specific weighting ratios - we compute the average performance of individual models across diverse test sets. Subsequently, we rank these models based on their performance metrics. After initially implementing the top-performing single method, we gradually incorporate additional methods, adding one at a time until all selected augmentations are combined (top 2, 3, and 4). The augmentation methods are applied probabilistically for each training set based on their observed gain ratio relative to the baseline model. Finally, three additional models are trained and evaluated to assess their performance.

Stage 5: Cross-Dataset Validation

The final experimental stage examines the generalization of the best-performing models from previous stages by validating them on an additional test set - SAVEE. This unseen dataset, which features different speakers and recording conditions, enables us to assess whether performance improvements extend beyond the training domain. We evaluate models directly on SAVEE without further fine-tuning, measuring their ability to cope with domain change. This validation provides critical insights into whether augmentation strategies enhance robustness and addresses the real-world applicability of our approach for diverse acoustic environments.

4 Experiments

The experiments utilized PyTorch 1.10 on a system with NVIDIA A100 GPUs, with model implementation via the Transformers library (version 4.18.0) and audio processing through the Librosa library (version 0.9.1). The following sections detail dataset characteristics and the implementation of our methodology.

4.1 Datasets

We employed two datasets commonly used for SER: RAVDESS for training and initial evaluation and SAVEE for cross-dataset validation. Both datasets share the same emotional categories, enabling cross-corpus assessment.

RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) includes 1,440 speech audio files from 24 professional actors (12 female, 12 male), each expressing 8 emotions with varying intensity. To ensure consistent emotion classification across all datasets, we excluded the ”calm” emotion category, leaving seven emotions for analysis.

SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset contains 480 recordings from four male speakers delivering utterances across 7 emotions. The dataset includes more neutral recordings (120) than other emotions (60). We randomly removed 60 neutral recordings to ensure balance, standardizing the dataset across all emotional classes. Despite its male-only composition, SAVEE provides high-quality, phonetically balanced sentences that complement the other datasets.

4.2 Dataset Pre-processing

The pre-processing steps transformed raw audio data into a suitable format for the HuBERT model. First, we performed an 80/20 split on the dataset to create training and test sets stratified by emotion to maintain a balanced class representation. Second, each audio file was preprocessed: loaded, resampled to a uniform sampling rate (16 kHz), and converted into numerical representation using Wav2Vec2FeatureExtractor from the pre-trained model. We used dynamic padding to maintain emotional content, and the extracted features were converted to PyTorch tensors for model compatibility.

4.3 Pretrained Model

We employed a ported version of S3PRL's HuBERT model for the SUPERB Emotion Recognition task from Hugging Face [1]. This pre-trained model was instantiated using the HubertForSequenceClassification architecture. To better capture the nuances of human emotion, we extended the original four-class schema (natural, happy, sad, angry) to include three additional categories - fear, disgust, and surprise - creating a seven-label classification system.

Table 1 summarizes the key components (layer configurations) of our HuBERT-based architecture along with the overall parameter count.

Table 1: Summary of the Hubert Large Model used for Fine-tuning on SER with **7 Emotions**.

Layer	**Input Shape**	**Output Shape**
Input	[1, 16k]	[1, 16k]
Feature Encoder (7 Convs)	[1, 16k]	[1, 512, 49]
– 1st Conv	kernel=10, stride=5	
– Next 4 Convs	kernel=3, stride=2	
– Last 2 Convs	kernel=2, stride=2	
Feature Projection	[1, 512, 49]	[1, 49, 1024]
Positional Conv	[1, 49, 1024]	[1, 49, 1024]
Transformer Encoder (24 layers)	[1, 49, 1024]	[1, 49, 1024]
– Each layer	Self-Attn + FFN + Norm	
– Hidden size	1024	
– FFN dimension	4096	
Projector	[1, 49, 1024]	[1, 49, 256]
Classifier	[1, 49, 256]	[1, 49, 7]
Total Parameters (Trainable)		**315.7M**

4.4 Training Procedures

The pre-trained HuBERT model was fine-tuned for 7 *epochs* with a *batch size* of 8. All model parameters were updated without freezing any layers, allowing for comprehensive adaptation to emotion recognition. The most stable convergence was achieved with AdamW *optimizer* and a *learning rate* of 1e-5. At each iteration, input values and attention masks were forwarded through the model to obtain logits, from which *softmax* probabilities and predictions were derived. The cross-entropy *loss* was calculated on these outputs and backpropagated to update parameters, while accuracy was monitored by comparing predictions to true labels.

Overall, we trained eight distinct models: Baseline (without augmentation), TimeStretch-Aug, Noise-Aug, SpecAugment-Aug, NCP-Aug, Top-2 (two best-performing augmentations), Top-3 (three best-performing augmentations), and All-Aug (all augmentation strategies weighted). Performance evaluations are in Section 5.

4.5 Augmentation Implementation

Our methodology addresses the challenges of limited clean, labeled data by employing parameterized augmentation techniques. For each audio sample, we randomly select parameters from predefined ranges, allowing for diverse but controlled augmentations while maintaining original emotional content. Below are the predefined ranges for each method:

SpecAugment - To disrupt both temporal and spectral features, time masking and frequency masking are applied. The time mask is defined by a parameter selected from 10, 20, 30, and similarly, the frequency mask parameter is chosen from 10, 20, 30.

Time Stretching - The audio signal is either compressed or expanded in time by applying stretch factors selected from the set 0.8, 0.9, 1.0, 1.1, 1.2.

Additive Noise - Gaussian noise is added to the audio with noise levels drawn from 0.001, 0.005, 0.01.

N-CP - Neutral utterances are segmented into three equal parts, replacing only the middle segment with emotional content. All possible neutral–emotional pairs are generated and

[1] https://huggingface.co/superb/hubert-large-superb-er

then sampled to maintain the original emotion distribution, effectively doubling the dataset without repetition (as illustrated in Appendix A).

Combination - We present a dual sampling approach: one for method type selection and another for parameter settings. To achieve this, the probability of weighted gains is calculated based on the normalized average performance of singular methods, specifically utilizing the F1-weighted score.

We generated an augmented dataset by reviewing all samples and tailoring it to the method with parameters sampled at each iteration, maintaining a 1:1 ratio with the original. The training sets included both clean and augmented RAVDNESS data, while the test set used only the augmented version.

4.6 Model Performance Evaluation

We evaluated each model's performance using classification reports, showing weighted F1 and accuracy scores for classifying emotions. Additionally, we identified specific strengths and weaknesses in emotion discrimination through confusion matrix analysis, highlighting which emotions were more easily confused and with whom. To assess the effectiveness of different augmentation strategies, we conducted a comparative analysis among the baseline model, single-method augmentation models, and combined augmentation approaches.

5 Results and Discussion

This section presents the experimental results on augmentation strategies for training and fine-tuning a HuBERT model for SER. We evaluate baseline performance, examine single and combined augmentation methods, conduct emotion-specific analyses, and conclude with cross-validation.

5.1 Baseline Performance

After training the pre-trained HuBERT model on the RAVDESS dataset, we evaluated it on the clean test set, achieving a weighted F1 score of 0.75 and 74% accuracy as our performance baseline (Table 3). The confusion matrix (Table 2) shows the model excels at recognizing the *neutral* (94.74%) and *angry* (89.47%) emotions, but struggles with *sad* (48.72%). Common misclassifications include *sad* as *neutral* (38.46%), *surprise* as *happy* (23.68%), and *fear* as *happy* (20.51%), consistent with established research on acoustic similarities between these emotion pairs. These findings provide benchmarks and identify areas for improvement in subsequent experiments.

5.2 Single Augmentation Method Performance

Single augmentation methods significantly improved emotion recognition. As shown in Table 3, SpecAugment-Aug achieved the highest results, with a 0.91 weighted F1 score on clean test data (16% improvement over baseline). This model also maintained a 0.91 F1 score on SpecAugment-augmented test data. Each augmentation method displayed domain-specific benefits. NCP-Aug demonstrated the most notable improvement on NCP-perturbed test data, increasing from 0.39 to 0.68 (74% relative improvement), indicating enhanced capability to detect emotions from partial audio segments. Despite SpecAugment-Aug's overall success, certain trade-offs were observed. The model still performed poorly on noisy and segmental audio, suggesting that combined augmentation strategies may offer additional benefits.

label	neutral	happy	sad	angry	fear	disgust	surprise
neutral	**94.74**	5.26	0.00	0.00	0.00	0.00	0.00
happy	12.82	**79.49**	0.00	0.00	7.69	0.00	0.00
sad	38.46	7.69	**48.72**	0.00	5.13	0.00	0.00
angry	2.63	2.63	0.00	**89.47**	0.00	0.00	5.26
fear	0.00	20.51	0.00	0.00	**79.49**	0.00	0.00
disgust	2.63	7.89	7.89	7.89	0.00	**63.16**	10.53
surprise	0.00	23.68	0.00	0.00	2.63	0.00	**73.68**

Table 2: Confusion matrix (percentages) for Baseline model on Clean test set (Accuracy: 74.00%). Bold: correct classifications; Underlined: major confusion patterns.

Augmentation Contribution Analysis - Figure 3 (in Appendix B) shows each augmentation technique's contributions to average performance gains. SpecAugment yielded the largest contribution, followed by TimeStretch, CopyPaste, and Noise. These findings indicate that frequency-domain disturbances and temporal modifications are effective improving speech emotion recognition, suggesting emotional content in speech is more resilient to such augmentations than previously assumed.

5.3 Combined Augmentation Performances

Table 4 shows that the Top-1 model (SpecAugment-only) maintains the highest performance on clean and spectrally-augmented data (0.91 F1). However, the Top-2 achieves better results on noise-perturbed (0.65 vs 0.61) and time-stretched data (0.84 vs 0.82).

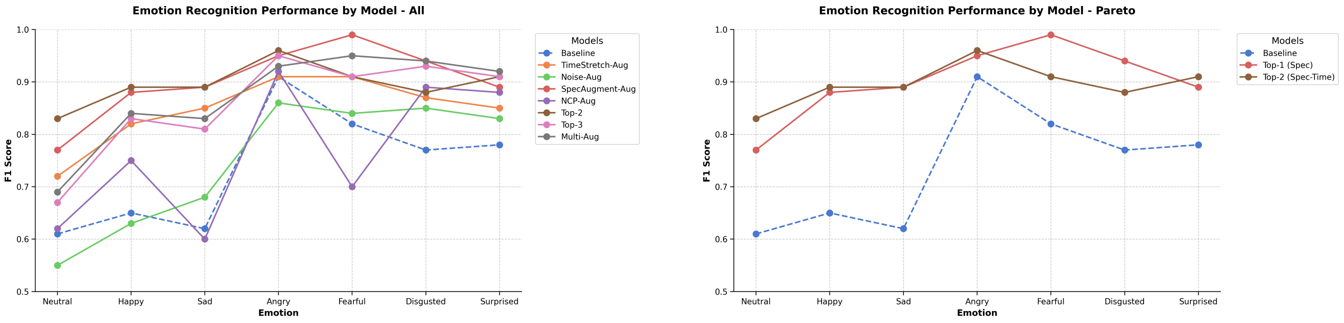
Adding more augmentations, specifically Top-3 and Top-4, did not yield further improvements and, in some cases, even reduced performance, suggesting possible conflicts among the different augmentation methods.

5.4 Emotion-Specific Performance Analysis

Figure 1 shows emotion-specific recognition patterns across models on the clean RAVDNESS dataset. The Pareto-optimal visualization (Fig. 1b) reveals two primary models on the efficiency frontier: Top-1 (SpecAugment), which ex-

Test Data \ Models	Baseline	TimeStretch-Aug	Noise-Aug	SpecAugment-Aug	NCP-Aug
Clean	0.75	0.85	0.76	**0.91***	0.77
TimeStretch	0.55	**0.78***	0.63	**0.78***	0.63
Noise	0.52	0.66	**0.68***	0.61	0.57
SpecAugment	0.72	0.84	0.75	**0.91***	0.77
NCP	0.39	0.46	0.45	0.57	**0.68***

Table 3: Comparison of weighted F1 scores for different single augmentation models across test sets. The highest scores are marked with (*).



(a) Performance comparison across all models

(b) Pareto-optimal models (best in at least one emotion)

Figure 1: Comparison of speech emotion recognition performance; (a) shows all models and (b) highlights Pareto-optimal ones.

Test Data \ Models	Top 1	Top 2	Top 3	Top 4
	(+SpecAug)	(+TimeAug)	(+NoiseAug)	(MultiAug)
Clean	**0.91***	0.90	0.87	0.88
SpecAugment	**0.91***	0.89	0.86	0.87
Time Stretch	**0.78***	**0.78***	0.76	0.76
NCP	**0.57***	0.52	0.49	0.45
Noise	0.61	**0.65***	0.62	0.56

Table 4: Comparison of weighted F1 scores for different augmentation combined models across test sets. The highest scores are marked with (*).

cels in detecting angry and disgusted, and Top-2 (SpecAugment+TimeStretch) which is optimal for neutral, happy, angry and surprised. Both augmented models substantially surpass the baseline for most emotions. The Top-2 model exhibits consistent performance across all emotion categories (with F1 scores above 0.83), suggesting that combining frequency and temporal augmentations confers complementary advantages for emotional speech recognition.

5.5 Cross Dataset Validation

Cross-dataset validation on SAVEE resulted in a noticeable performance decline (to 48% accuracy) across all models (Appendix C), underscoring the challenges of generalizing across different speaker demographics and recording environments. SAVEE's male-only British English corpus presents significant acoustic differences compared to RAVDESS's gender-balanced North American English recordings.

All models struggled to classify fear, sadness, and surprise, though each showed distinct error patterns. The baseline exhibited a strong bias toward ”happy” classification, misclassifying 31% of ”surprise” and 42% of ”fear” samples. The SpecAugment model improved ”angry” detection (62% vs. 42% baseline) and excelled at ”neutral” (95%), but struggled severely with ”fear” (3% recall). In contrast, the SpecAugment+TimeStretch model achieved a more balanced performance with improved ”angry” (70%) and ”happy” (88%) detection. This performance gap indicates that augmentation alone cannot fully address the inherent cross-corpus differences in emotion expression.

6 Conclusions

The results discussed in the previous chapter provide insights into augmentation strategies for speech emotion recognition using HuBERT. Frequency-domain augmentations, particularly SpecAugment, proved most effective in preserving emotional markers while enhancing model robustness, especially for angry and fearful emotions. The SpecAugment+TimeStretch combination demonstrated complementary benefits by targeting both frequency and temporal domains, achieving balanced performance across emotions. However, cross-corpus validation revealed persistent challenges, with performance gaps indicating that augmentation alone cannot bridge fundamental acoustic differences between datasets with distinct speaker demographics. These findings establish a framework for emotion-specific augmentation selection and highlight the need for more sophisticated approaches to cross-corpus generalization.

7 Future Work

Below are several research directions building on our results:

Further Augmentation Exploration. Investigating additional augmentation techniques and implementing adaptive systems that dynamically adjust parameters based on model performance to improve robustness across diverse emotional expressions.

Expanding Dataset Scope. Broadening our evaluation to include diverse datasets, like cross-lingual and domain-specific corpora, can provide insights into the generalization of our approach and its robustness to different speaker characteristics and recording conditions.

Sequential Augmentation Training. Exploring sequential training strategies where augmentations are applied in distinct phases instead of proportional combinations. This approach may enable the model to fully adapt to each type of augmentation before progressing to the next one.

References

[1] 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, pp. 335–359, 2008.

[2] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, “End-to-end speech recognition: A survey,” *arXiv preprint arXiv:2303.03329*, 2023.

[3] 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. 12449–12460, 2020.

[4] 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.

[5] 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,” *arXiv preprint arXiv:1904.08779*, 2019.

[6] 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.

[7] S. Haq, P. J. Jackson, and J. Edge, “Speaker-dependent audio-visual emotion recognition.,” in *Avsp*, vol. 2009, pp. 53–58, 2009.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: audio, visual and spontaneous expressions,” in *Proceedings of the 9th international conference on Multimodal interfaces*, pp. 126–133, 2007.

[9] M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” *arXiv preprint arXiv:1706.00612*, 2017.

[10] B. T. Atmaja and A. Sasou, “Effects of data augmentations on speech emotion recognition,” *Sensors*, vol. 22, no. 16, p. 5941, 2022.

[11] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” *IEEE access*, vol. 9, pp. 47795–47814, 2021.

[12] M. S. Fahad, A. Ranjan, J. Yadav, and A. Deepak, “A survey of speech emotion recognition in natural environment,” *Digital signal processing*, vol. 110, p. 102951, 2021.

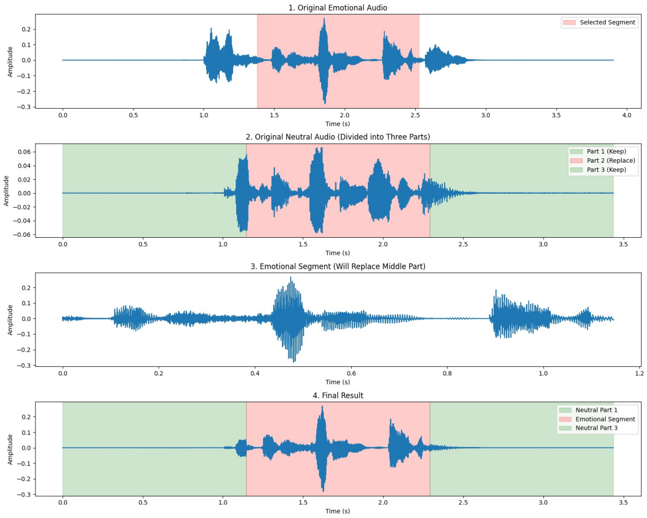
[13] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, *et al.*, “Superb: Speech processing universal performance benchmark,” *arXiv preprint arXiv:2105.01051*, 2021.

[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)*, pp. 6324–6328, IEEE, 2021.

Appendix

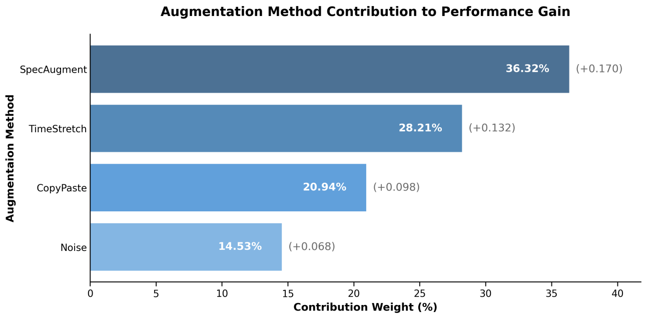
A Augmentation Implementation

Figure 2: Process of Neutral Content Preservation (NCP) augmentation method.



B Augmentation Contribution

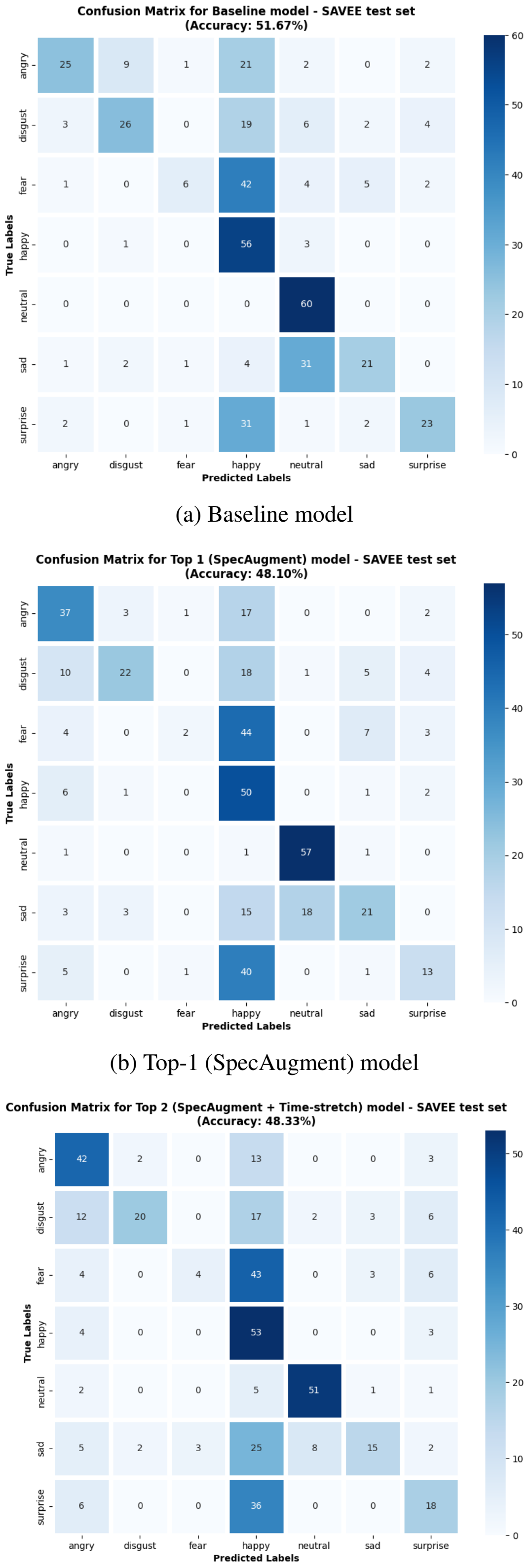
Figure 3: Relative contribution of augmentation methods to emotion recognition performance, showing percentage impact and absolute averaged F1-score improvement.



C Cross Validation with SAVEE

Despite similar overall accuracy, the models show distinct error patterns, with the Top-2 model demonstrating the most balanced emotion recognition profile.

Figure 4: Confusion matrices for cross-validation on SAVEE dataset.



(a) Baseline model

(b) Top-1 (SpecAugment) model

(c) Top-2 (SpecAugment+TimeStretch) model