Article

Building a Gender-Bias-Resistant Super Corpus as a Deep Learning Baseline for Speech Emotion Recognition

Babak Abbaschian * and Adel Elmaghraby

Posted Date: 30 December 2024

doi: 10.20944/preprints202412.2422.v1

Keywords: speech emotion recognition; deep learning; LSTM; CNN; gender bias; fairness; speech emotion database; transfer learning; transformers

Article

Building a Gender-Bias-Resistant Super Corpus as a Deep Learning Baseline for Speech Emotion Recognition

Babak Abbaschian * and Adel Elmaghraby

University of Louisville

* Correspondence: b0joze01@louisville.edu

Abstract: The focus on speech emotion recognition (SER) has dramatically increased in recent years, driven by the need for automatic speech recognition-based systems and intelligent assistants to enhance user experience by incorporating emotional content. While deep learning techniques have significantly advanced SER systems, their robustness concerning speaker gender and out-of-distribution data has not been thoroughly examined. Furthermore, standards for SER remain rooted in landmark papers from the 2000s, even though modern deep learning architectures can achieve comparable or superior results to the state-of-the-art of that era. In this research, we address these challenges by creating a new super corpus from existing databases, providing a larger pool of samples. We benchmark this dataset using various deep learning architectures, setting a new baseline for the task. Additionally, our experiments reveal that models trained on this super corpus demonstrate superior generalization and accuracy and exhibit lower gender bias compared to models trained on individual databases. We further show that traditional preprocessing techniques, such as denoising and normalization, are insufficient to address inherent biases in the data. However, our data augmentation approach effectively shifts these biases, improving model fairness across gender groups and emotions and, in some cases, fully debiasing the models.

Keywords: speech emotion recognition; deep learning; LSTM; CNN; gender bias; fairness; speech emotion database; transfer learning; transformers

1. Introduction

With the introduction of mainstream deep learning methods, the doors to solving more complex digital signal processing problems, such as Automatic Speech Recognition (ASR) and Speech emotion recognition (SER), have opened. Reviewing research published in the past decade shows a positive trajectory [1] of efficiency and accuracy each year.

A group of the significant drivers of SER is the big tech companies trying to create a robust solution for this problem, with Amazon Alexa, Google Assistant, Apple Siri, and Microsoft Cortana [2]. There has been an influx of research on SER, from proposals to create better training sets to methods that increase accuracy and to the ones that make the systems more robust and reliable in real situations.

The rationale behind these endeavors stems from the anticipation that advancements in Speech Emotion Recognition (SER) will significantly enhance the landscape of human-computer interaction. The pivotal role of SER lies in its potential to facilitate the development of systems capable of comprehending human commands with greater acuity and responding adeptly across diverse scenarios. Noteworthy instances of application include the optimization of interactions between smart speakers, virtual assistants, and end users. In particular, SER proves instrumental in refining the quality of exchanges within speech-to-text applications, addressing the challenges of informal language structures that deviate from conventional grammar and syntax. This is exemplified when written sentences may not accurately convey content due to the absence of appropriate intonation, such as in the case of polar (Yes/No) questions, as illustrated by the query "You pet the cat?" [3]. SER also has benefits in psychotherapy, customer service, remote learning, self-driving vehicles, and more [1].

As mentioned above, older research mainly focused on techniques used in automatic speech recognition (ASR) and machine learning [4] [3]. These methods' results were limited, e.g., 74.2% accuracy [4], and 77% [3]. However, the newer publications focusing on deep learning are generating better results, with reporting 95% [5], 93% [6], 95.5% [7], and 97% accuracy [8].

In addition to the methods, many databases have been introduced for speech emotion recognition. General categories of emotional speech databases include natural, semi-natural, and simulated databases [1].

All considered, the reliability of SER models faces challenges in terms of fairness, robustness, and accuracy when confronted with environmental noise, out-of-distribution test settings, and gender bias. Environmental noise, such as background sounds, can introduce variability and hinder the model's performance. Out-of-distribution test settings, where the data differs significantly from the training set, threaten the generalization capabilities of SER models. Additionally, gender bias can manifest in disparities in the recognition accuracy across different gender groups. Recognizing these vulnerabilities, we are exploring normalization techniques, noise reduction strategies, and the utilization of a comprehensive super corpus to enhance the fairness, robustness, and overall accuracy of SER models in diverse and challenging scenarios.

1.1. Contribution

Despite the notable achievements of deep neural networks (DNNs), their efficacy is contingent upon the characteristics of the training data, and concerns persist regarding their generalization capabilities. While recent attention has been directed toward assessing the fairness and robustness of DNNs in computer vision and natural language processing, to the best of our knowledge, we are the first to explore the robustness of SER models to speaker's gender and out-of-distribution test samples. Specifically, our contribution includes:

1. Building and evaluating a series of modern deep learning-based architectures, establishing a new baseline that outperforms older benchmarks across accuracy, F1 score, precision, and recall metrics while maintaining balanced performance across datasets and preprocessing strategies.
2. Introducing a super corpus that augments the sample pool and diversifies the datasets, enabling broader applicability and robustness.
3. Conducting a detailed examination of model robustness concerning speaker gender and cross-corpora scenarios.
4. Finally, we will provide a comprehensive discussion on how our proposed super corpus, coupled with various preprocessing strategies, contributes to improving generalization, mitigating gender bias, and enhancing the robustness of the SER models to the out-of-distribution data samples.

The resulting models and datasets serve as a foundational baseline for future research in this domain.

1.2. Organization

The rest of the paper is organized as follows: Section 2 revisits related works in cross-corpus SER and fairness discussions in speech processing. Section 3 sets up the problem and formalizes the approach for crafting super corpora. Section 4 discusses the datasets and the examination metrics and models. We will present and discuss the results in Section 5; the last section concludes the paper.

2. Related Works

The known problem of limited data available to train SER models has driven many to try creative methods such as augmenting data or cross-corpus training. In this chapter, we will review some of the works on mitigating data limitations and demographic bias of SER systems. Additionally, one way of mitigating the lack of inductive biases, i.e., architectural choices and robust priors in deep learning, is expanding the training data [9] [10]. However, as suggested by [11], various inductive can be equal, more or less data, and in scenarios involving extensive datasets, the benefits conferred by inductive biases may diminish, implying that the evaluation of the advantage derived from inductive biases and their implementation becomes particularly interesting in transfer settings where limited examples for the new distribution is available.

In one of the early efforts at data augmentation, Zhang et al. [12] researched the suitability of unsupervised learning in cross-corpus settings using ABC [13], AVIC [14], DES [15], eNTERFACE [16], SAL [17], and VAM [18] datasets. SAL and VAM are based on arousal/valence; the rest are categorical emotions. Therefore, they mapped the categorical databases to Arousal and Valence descriptors to unify the data. They employed the openEAR toolkit [19] to extract features and retain 39 functional of the 56 acoustic Low-Level Descriptors (LLDs). As an extra step in their preprocessing, they normalized all databases to zero means. And finally, to evaluate their experiment, they followed a cross-corpus leave-one-out strategy.

To evaluate their methods, they have three experiments planned. The first experiment was creating an agglomeration of 3 databases and testing on one database. The second experiment was creating a supervised training database based on 3 of the databases and then unsupervised training the model with two other databases and testing the remaining database. The last experiment involved training on five databases and testing on the remaining. They conclude that adding unlabeled samples to an agglomerated multi-corpus training set improves the accuracy of the model; however, the improvement on average is half of the results if they had added the same amount of labeled data.

In 2019, Milner et al. [19] investigated cross-corpora SER by incorporating a bidirectional LSTM with an attention mechanism. Their research investigates information transfer from acted databases to natural databases. Moreover, they have also looked into domain adversarial training (DAT) and out-of-domain (OOD) models and considered adapting them.

Their network is a triple attention network consisting of a BLSTM, the attention architecture, and the emotion classifier at the end. For domain adversarial training, they also consider adding a domain identifier to the training set that teaches the model how it is doing with each dataset. In their work, they use two acted datasets, eNTERFACE and RAVDESS, one elicited dataset, IEMOCAP, and one natural dataset, MOSEI.

This study concludes that when testing cross-corpus, the matched results outperform mismatched, and the model trained on simulated datasets generally achieves the best mismatched performance. They also discuss that the model trained on multi-domain is performing better than all of the other mismatched models due to more generalization resulting from having a larger dataset. They continue to show in their results that adding the domain information does not help the multi-domain model to generalize better, but training with more other domains helps improve the mismatched results.

In 2021, Wisha et al. [20], worked on a cross-corpus, cross-language ensemble method to detect emotions from 4 languages. They have used Savee (English) [21], URDU (Urdu) [22], EMO-DB (German) [23], and EMOVO (Italian) [24] in their research. However, their study only investigates binary valence. To train their classifier, they used Spectral features such as 20 MFCC coefficients and prosodic features defined in eGeMAPS [25]. The classifiers that create the ensemble are SVM with a Pearson VII function-based Universal Kernel, a random forest with ten trees, and a C4.5 algorithm decision tree.

As a result of their study, they claim that their methods have an increase of 13% for URDU, 8% for EMO-DB, 11% for EMOVO, and 5% for Savee in their in-corpus tests. They also report that in their cross-corpus tests, they achieved an improvement of 2% training on Urdu and testing on German data, 15% when testing on Italian, and when testing on Urdu while trained on German, 7%, with training with Italian, 3% and lastly training with English, they have gained a 5% improvement in their accuracy.

In 2021, Braunschweiler et al. [26] investigated cross-corpus data augmentation's impact on model accuracy. In their research, they incorporate a network with 6 layers of CNN, a Bidirectional LSTM model with 2 layers of 512 nodes, and 4 fully connected layers fed to an attention mechanism. The databases used in this research are IEMOCAP [27], RAVDESS [28], CMU-MOSEI [29], and three in-house single speaker corpora, named TF1, TF2, and TM1; F and M stand for female and male, respectively. The classes they chose to recognize using their model were angry, happy, sad, and neutral. To increase variability in their samples and improve their model generalization, they also applied variable speed, volume, and multiple frequency distortions such as bass, treble, overdrive, and tempo changes to the samples.

They discuss that their investigation shows that in situations where the model was trained with one database and tested with another, they have an accuracy decline of 10-40%. They further discuss that their results were improved when the model was trained with more than one corpora and tested on one of the corpora in the training set, except for their single-speaker datasets. The last result that they report is a 4% gain in accuracy with additional data augmentation.

Later, in 2022, Latif et al. [30] introduced an adversarial dual discriminator (ADDi) network trained on cross-language and cross-corpora domains. They claim their model improves the performance over the state-of-the-art models. Their model contains an encoder, a generator, and a dual discriminator. In their model, they are mapping the data to a domain-invariant latent representation. The generator uses the result of the encoder to generate target or source domain samples and the two adversarial discriminators that, in combination with the generator, tune the domain invariant representation to minimize the loss function. The generator and the encoder act as decoders to construct the input samples.

In their self-supervised training process, they introduce synthetic data generation as a pretext task that helps to improve domain generalization. As a byproduct, synthetic emotional data is produced that can augment the SER training set and help with more generalization.

They further discuss that introducing the ADDi network improves cross-corpus and cross-language SER without using target data labels. They also add that their model significantly improves by feeding partial target labels. They also claim that with the help of the self-supervised pretext task, they can achieve the same performance by training their ADDi network with 15-20% less training data.

Regardless of the accuracy of deep learning models, their robustness to changes in data distribution and making fair decisions is still open research. In 2020, Meyer et al. [31] published Artie Bias Corpus as the first English dataset for speech recognition applications with demographic tags, age, gender, and accent curated from *Mozilla Common Voice corpus (Needs citation).* Additionally, they published open-source software for their dataset to detect demographic biases in ASR systems.

Similarly, in 2021, Feng et al. [32] quantified the bias in the state-of-the-art Dutch Automatic speech recognition (ASR) system against gender and age. Their work reported the bias regarding word error rates (WER). They concluded that the ASR system studied had a higher WER for male Dutch speakers than for female speakers.

Lastly, in 2022, a team of researchers from Meta [33] released a manually transcripted dataset containing 846 hours of corpus for fairness assessment of ASR and facial recognition systems across different ages, genders, and skin tones. According to their results, several ASR systems lack fairness across gender and skin tone and have higher word error rates for specific demographics.

3. Problem

The current challenge in the field pertains to the resilience of deep learning models against out-of-distribution data instances and demographic biases. This matter persists as an unresolved concern. Our approach involves the utilization of pre-existing open-source datasets to enhance the generalizability of established Speech Emotion Recognition (SER) methodologies. Furthermore, we endeavor to alleviate biases directed towards particular speaker genders whenever feasible. We

doi:10.20944/preprints202412.2422.v1

assert that the augmentation of datasets within the realm of deep learning models for SER holds substantial potential, mainly when such augmentation is cost-effective and maintains the explicability of the end-to-end process. Within our dataset repository, we permit the augmentation of each dataset with all others. Notably, we introduce a specific case wherein solely simulated datasets undergo augmentation. This is motivated by their shared attributes, such as a controlled noisy environment and a predefined set of speakers. Additionally, our experimental framework examines the impacts of noise reduction and normalization.

4. Experimental Setup

In this section, we introduce our approach to generating super corpora. Additionally, we explain how we built our baseline setup based on a wide range of available architectures for deep learning-based SER.

4.1. Super Corpora

As previously stated, the performance of Speech Emotion Recognition (SER) systems experiences degradation when confronted with out-of-distribution samples. Furthermore, as demonstrated in Section 5, the performance of these systems varies inconsistently across different speakers' genders. In our proposed solution, we endeavor to address these issues by mitigating the impact of data distribution disparities. This is achieved through the augmentation and amalgamation of diverse datasets, yielding a comprehensive dataset called a "super corpus."

In Speech Emotion Recognition (SER), datasets are classified into three categories: Natural, Semi-Natural, and Simulated. Natural datasets are derived from authentic speech instances from diverse contexts such as news, online talk shows, and customer service call recordings. Labeling such datasets is inherently challenging due to the ambiguity of speaker intentions and the potential for varied listener interpretations of emotions. Consequently, the labeling process necessitates a sizable cohort of annotators and a structured voting system to determine emotional labels. Another inherent challenge with natural datasets lies in the dynamic nature of emotions within spontaneous speech. For instance, in a customer service call, emotions can transition rapidly from a neutral state to frustration or anger within a span of seconds. This fluidity poses difficulties in precisely labeling utterances or even entire sentences, constituting a complex and subjective task [1]. Examples of natural datasets include Vera Am Mittag (VAM) [18], and FAU Aibo [34].

Semi-natural datasets are created based on predefined scenarios and plots, and then one or more voice actors execute them. The emotional expressions within this dataset category are not strictly organic and may sometimes be exaggerated. Nonetheless, the advantage lies in achieving heightened control over the dataset, as the intended emotions are known, rendering the labeling process more dependable. However, challenges persist within this dataset paradigm, particularly concerning the dynamic nature of emotions and the intricate task of labeling utterances [1]. IEMOCAP [27], Belfast [35], and NIMITEK [36] are examples of this type of dataset.

The third dataset category, simulated, is constructed from a set of emotionally neutral sentences enunciated by voice actors who infuse various emotions into their delivery. The employment of emotionally neutral sentences imbued with diverse emotional expressions serves dual purposes. Firstly, it prevents the acquisition of emotionally biased sentences being learned by machine learning models, thereby mitigating the risk of triggering responses based solely on, for instance, the identification of emotion-related keywords such as "angry" within a speech signal. Secondly, the repetition of identical sentences articulated with different emotions ensures that the classifier model remains impervious to the semantic content of the sentences, thereby facilitating the isolation of the shared direct current (dc) component in the convoluted signal field [1]. EMO-DB (German) [23], DES (Danish) [15], RAVDESS [28], TESS [37], and CREMA-D [38] are examples of this type of dataset.

To systematically select integrated datasets, we establish our criteria set as follows: the primary criterion for dataset selection is language. Given the variation in how emotions are expressed through speech across different languages, we exclusively consider datasets in the English language.

Subsequently, for result comparison, we opt for datasets associated with multiple models substantiated by published papers. Furthermore, given our focus on addressing the challenge of exposure to out-of-distribution data samples, we prioritize datasets characterized by a substantial volume of samples and a diverse spectrum of emotional expressions, encompassing not only positive or negative sentiments. Plus, we will utilize simulated-only and semi-natural datasets to address labeling challenges, variations in utterance size, and linguistic nuances related to emotional content. Therefore, from the list of open-sourced datasets in English, we chose three simulated English datasets of RAVDESS, TESS, and CREMA-D, plus the widely used semi-natural dataset of IEMOCAP. Ultimately, we abstain from employing augmentation processes rooted in deep learning to maintain interpretability within our methodology. Instead, we exclusively leverage the extant data samples to augment each dataset. This will facilitate subsequent extensions of our robustness evaluations to encompass more intricate scenarios, notably including assessments of robustness against adversarial attacks, which we will address in future works.

From the selected databases, which feature a variety of emotions, we found that four emotions, Happiness, Anger, Sadness, and Neutral, are present in all of them. We decided to use these samples for our project. In Table 1, we present a summary of each dataset. Since we aim to investigate the performance of models w.r.t. bias against speaker's gender, the presented statistics are divided based on two groups of speakers, Male and Female.

Table 1. Number of samples per emotion (Happy, Neutral, Sad, Angry) generated by Female and Male speakers for each dataset.

Dataset	**Settings**	**Female Speaker**				**Male Speaker**			
		Happy	**Neutral**	**Sad**	**Angry**	**Happy**	**Neutral**	**Sad**	**Angry**
RAVDESS	Simulated	96	48	96	96	96	48	96	96
TESS	Simulated	400	400	400	400	0	0	0	0
CREMA-D	Simulated	600	512	600	600	671	575	671	671
IEMOCAP	Semi-Sim.	291	911	509	505	304	797	575	598

4.1.1. Building the Super Corpora

To construct our experimental corpus, we utilized a selection of databases, each comprising PCM (Pulse Code Modulation) encoded WAV files. However, the encoding formats, sample rates, and other audio characteristics varied significantly across the databases. For instance, the CREMA-D database had a sample rate of 16 kHz, while others used 48 kHz. Additionally, two databases were recorded in stereo, while the remaining two were mono. We resampled the audio files to 16-bit, 16 kHz, and mono signals to ensure uniformity across all datasets.

Upon reviewing file sizes and utterance durations, we observed that several utterances in the IEMOCAP database were shorter than one second. These short samples presented challenges even for human listeners in reliably identifying emotions. Consequently, we excluded all samples with less than one-second durations from our dataset.

Further analysis of the files revealed significant statistical differences across the datasets, including variations in average mean, DC component (Mean Signal Offset), peak amplitude, amplitude range, Signal-to-Noise Ratio (SNR), voicing characteristics such as Zero-Crossing Rate (ZCR), and prosodic features. Notably, voicing and prosodic features are closely linked to the emotional content of speech signals.

In contrast, features like the DC component, Root Mean Square (RMS) energy, and SNR primarily create energy-related signal denormalization effects. For example, normalizing the audio could reduce the energy in high-frequency samples (e.g., Angry or Happy emotions) or mitigate noise interference in low-energy samples (e.g., Sad emotions). Moreover, the same normalization will reduce the relative prominence of high-energy components (e.g., transients or spikes) across all frequencies. As a result, their perceptual prominence will change. Similarly, spectral noise reduction will attenuate energy at the signal's lower and higher frequency margins as an unwanted effect.

One of the key features in gender identification is the signal's high- and low-frequency energy content [39]. Consequently, these preprocessing methods are likely to reduce gender-specific characteristics in speech. Based on this observation, we hypothesized that normalization and spectral noise reduction would remove more gender-specific discriminative content than emotional information from the speech signals.

We applied preprocessing methods, including RMS Normalization and Spectral Noise Reduction, to test this hypothesis and evaluate their effects on classification performance and bias. We implemented four preprocessing schemes and applied them to all datasets:

1. No preprocessing (Raw),
2. RMS Normalized,
3. Noise Reduced,
4. RMS Normalized and Noise Reduced.

In the next step, we created a series of Mel-Frequency Cepstral Coefficient (MFCC) representations for all datasets while retaining the original PCM (WAV file) versions. For MFCC generation, we experimented with different window sizes in both time and frequency domains. Time-wise, we used shorter window sizes, such as 25 ms, and longer durations approaching one second. Frequency-wise, we generated representations with 13 coefficients (including delta and delta-delta) and 64 coefficients (also with delta and delta-delta).

Additionally, we experimented with various database combinations. These included:

1. Merging all samples from all databases into a single bucket (All),
2. Using only simulated database samples (Simulated Only),
3. Keeping each database separate.

This setup resulted in six database combinations. By applying all preprocessing schemes across these combinations, we built a total of 72 databases for analysis. The following figure illustrates all the database combinations used in this investigation.

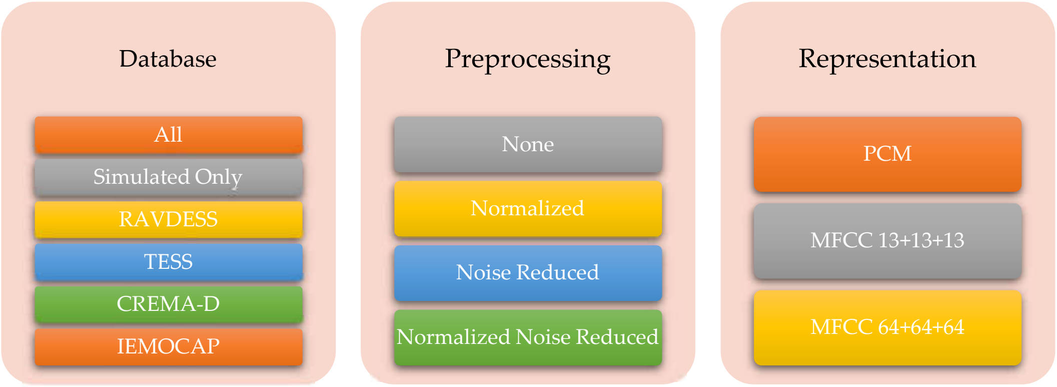

Figure 1. Each combination of databases had gone through all of the preprocessing options, creating four variations. Then, each of the variations was represented by the representation schemas.

4.2. Deep Learning-Based SER

Creating a versatile baseline is essential to covering a wide range of diverse approaches to SER. Therefore, we have chosen several deep learning methods, ANN-based, CNN-based, and LSTM-based, to cover various methods applicable to MFCC preprocessed speech datasets. In Table 2, we present a summary of the networks we have examined, and we will explain the result in Section 5.

The architecture names in the first column of Table 2 follow a systematic notation to represent the layers used in each model. For instance, "3XCNN1D 1XLSTM 2XDENSE" refers to a model that begins with three one-dimensional convolutional neural network (CNN) layers, followed by a single Long-Short-Term Memory (LSTM) layer and concludes with two fully connected dense layers. As clarified in the literature [1], each layer type serves a specific purpose in processing the speech data, which we briefly revisit in the following.

doi:10.20944/preprints202412.2422.v1

Table 2. Top architecture's generalization and predictive performances across each individual and augmented dataset. Four preprocessing approaches have been employed per dataset.

ARCHITECTURE	DATASET	PREPROCESSING	ACCURACY	F1 SCORE	PRECISION	RECALL
3XCNN1D 1XLSTM 2XDENSE	All	Raw MFCC	8.51E-01	8.56E-01	8.61E-01	8.51E-01
3XCNN1D 1XLSTM 2XDENSE	All	Normalized	8.55E-01	8.57E-01	8.61E-01	8.55E-01
3XCNN1D 1XLSTM 2XDENSE	All	Denoised	8.35E-01	8.43E-01	8.53E-01	8.35E-01
3XCNN1D 1XLSTM 2XDENSE	All	Denoise Normalized	8.37E-01	8.42E-01	8.47E-01	8.37E-01
3XCNN1D 1XLSTM 2DENSE	Simulated only	Raw MFCC	8.88E-01	8.91E-01	8.95E-01	8.88E-01
3XCNN1D 1XLSTM 2XDENSE	Simulated only	Normalized	8.87E-01	8.89E-01	8.92E-01	8.87E-01
3XCNN1D 1XLSTM 2XDENSE	Simulated only	Denoised	8.92E-01	8.94E-01	8.97E-01	8.92E-01
3XCNN1D 2XLSTM 2XDENSE	Simulated only	Denoise Normalized	8.77E-01	8.79E-01	8.82E-01	8.77E-01
3XCNN1D 1XLSTM 2XDENSE	IEMOCAP	Raw MFCC	8.23E-01	8.28E-01	8.34E-01	8.23E-01
3XCNN1D 1XLSTM 2XDENSE	IEMOCAP	Normalized	8.36E-01	8.41E-01	8.47E-01	8.36E-01
3XCNN1D 1XLSTM 2XDENSE	IEMOCAP	Denoised	8.16E-01	8.23E-01	8.30E-01	8.16E-01
3XCNN1D 1XLSTM 2XDENSE	IEMOCAP	Denoise Normalized	8.16E-01	8.20E-01	8.27E-01	8.16E-01
3XTDCNN1D 2XLSTM 2XDENSE	CREMAD	Raw MFCC	8.31E-01	8.33E-01	8.39E-01	8.31E-01
3XCNN1D 1XLSTM 2XDENSE	CREMAD	Normalized	8.65E-01	8.68E-01	8.75E-01	8.65E-01
3XCNN1D 1XLSTM 2XDENSE	CREMAD	Denoised	8.65E-01	8.67E-01	8.72E-01	8.65E-01
3XCNN1D 1XLSTM 2XDENSE	CREMAD	Denoise Normalized	8.40E-01	8.43E-01	8.48E-01	8.40E-01
3XTDCNN1D 2XLSTM 2XDENSE	RAVDESS	Raw MFCC	8.72E-01	8.73E-01	8.75E-01	8.72E-01
3XTDCNN1D 2XLSTM 2XDENSE	RAVDESS	Normalized	8.24E-01	8.28E-01	8.34E-01	8.24E-01
2XTDCNN1D 2XLSTM 2XDENSE	RAVDESS	Denoised	8.80E-01	8.82E-01	8.86E-01	8.80E-01
3XCNN1D 1XLSTM 2XDENSE	RAVDESS	Denoise Normalized	8.47E-01	8.50E-01	8.60E-01	8.47E-01
3XTDCNN1D 2XLSTM 2XDENSE	TESS	Raw MFCC	9.97E-01	9.97E-01	9.97E-01	9.97E-01
3XTDCNN1D 2XLSTM 2XDENSE	TESS	Normalized	9.96E-01	9.96E-01	9.96E-01	9.96E-01
DSCNN2D	TESS	Denoised	9.93E-01	9.93E-01	9.94E-01	9.93E-01
2XTDCNN1D 2XLSTM 2XDENSE	TESS	Denoise Normalized	9.97E-01	9.97E-01	9.97E-01	9.97E-01

4.2.1. 1D Convolutional Layers (CNN1D)

In SER, CNN1D captures spatial patterns across features, making them ideal for analyzing temporal sequences like MFCCs, which encapsulate the frequency-time patterns in speech signals.

4.2.2. 1D Temporal Convolutional Networks (TDCNN1D)

Architecturally, TDCNN1D is similar to 1D CNNs but tailored to capture longer dependencies using dilated convolutions. These enable the network to look further back in the sequence without substantially increasing the computational cost.

4.2.3. Long short-term memory (LSTM)

LSTM Layers are recurrent neural networks (RNNs) designed to capture long-term dependencies and temporal relationships, which are critical for speech-related tasks, where context over time can influence emotion detection.

4.2.4. Dense

Finally, Dense (Fully Connected) Layers integrate the features learned by previous layers, enabling the model to make final classifications or predictions.

Each model configuration in Table 2 varies based on the dataset used (e.g., IEMOCAP, CREMAD) and preprocessing techniques applied to the MFCCs, such as denoising or normalization. These variations help us understand the model's robustness and adaptability across different datasets and preprocessing choices, providing a comprehensive foundation for Speech Emotion Recognition (SER) research.

4.3. Downstream Bias

Female speakers convey their emotions more expressively than Male speakers, which could result in inconsistent performance if the classifier is used in real-world applications. The empirical true positive rate (TPR) estimates the probability that the classifier accurately identifies a person's emotion from their speech. Following previous research [40] and [41], we measure downstream bias by examining the empirical TPR gap between speeches for each gender set. First, define

$$TPR_{y,g} = P\left[\hat{Y} = y \middle| G = g, Y = y\right]$$

where g is a set of genders, and y is an emotion. Y and $\hat{Y}$ are the true and predicted emotions, respectively. Then, TPR bias (TPB) is

$$TPB_y = \frac{TPR_{y,Female}}{TPR_{y,Male}}$$

where if a classifier predicts "Angry" for a Male speaker much more often than for a Female speaker, the TPR ratio for the "Angry" class is low.

5. Results and Discussions

This section presents the results of employing our augmentation process to craft super corpora on the deep learning-based SER models. We show both predictive performance-related experiments on all datasets and gender-specific generalization experiments.

When developing SER systems, the final model is trained and tested on a specific dataset. In this situation, the generalization and predictive performance of the developed model are constrained to the dataset. However, based on our experiments in this section, we demonstrate that the models' performances are inconsistent in inner-corpora and cross-corpora settings. Additionally, even the best models, i.e., those with high accuracy and F1 scores, are biased, and their performance is inconsistent across different speaker genders. Ultimately, our proposed augmentation approach effectively improves the cross-corpora performance, also known as exposure to out-of-distribution generalization, and mitigates gender bias.

We have described the build process of our Super Corpora experiment in Section 4.1.1. As a result, we had 72 databases and over 13 models with which to run the experiment. After some experimentation and comparing the results, as one of our objectives was adding a limited computation overhead, we continued working with the MFCC window of (120, 39), about 600ms audio, and 13+13+13 coefficients with a 25% overlapping window. We dropped using the 64+64+64 MFCCs, as their performance gain was limited compared to complexity overhead. Also, in our initial experiments, we explored the utilization of spectrograms, as suggested by Wani et al. [42]. However, spectrogram-based experiments are not computationally efficient and suffer from implicit biases.

We conducted extensive experimental evaluations across a range of network architectures and varying model sizes applied to each dataset. However, in this section, we only report the performance of top networks on each dataset and finally extend our augmentation experiment on them. It is noteworthy that DSCNN2D, as SER state-of-the-art architecture, has not consistently outperformed other architectures in our experiment.

Additionally, the gender-bias hypothesis in the form of downstream bias is measured for each dataset separately and combined. This allows us to investigate whether the bias is influenced by the dataset specification or implicitly by the network architecture. The bias hypothesis here is measured using the TPB metric (described in Section 4.3) as well as the differences in predictive metrics (accuracy, F1 score, precision, recall, and Confusion matrix) between different groups of speakers and emotions. Moreover, we investigate how models trained with a particular dataset using different preprocessing approaches, normalization, and noise reduction are vulnerable when exposed to out-of-distribution data samples. Finally, we present how our proposed super-corpus improves each deficiency and the emotional confusion of these models for female and male speakers.

Table 2 presents the overall accuracy and F1 score, precision, and recall of each top architecture across individual datasets and the proposed super corpus in four preprocessing scenarios. As mentioned in Section 4.1.1, in our experiments, we refer to the super corpora that have been

doi:10.20944/preprints202412.2422.v1

augmented by using all datasets as "All," and as the name implies, "Simulated only" refers to the super corpora that have incorporated RAVDESS, TESS, and CREMA-D in augmentation process.

5.1. Deep Learning Architectures and Their Performance

As mentioned in the previous section, IEMOCAP is a semi-simulated and relatively large dataset. As can be found from the results, SER models can benefit from our data augmentation approach to improve their generalization and demonstrate better predictive performance at no computation cost since the batch size and training epochs have been fixed for all of the reported results. Notably, having the best performance over the TESS dataset can be explained by its unique design, smaller spoken content variation, and only two female speakers uttering all the samples. Apart from the TESS dataset, the "simulated only" augmented dataset effectively outperforms all others, regardless of the preprocessing schemes.

At the beginning of this study, one objective was to establish a modern baseline for the SER models, ensuring future comparisons are not limited to state-of-the-art models from two decades ago. To this end, Table 3 summarizes the previously published state-of-the-art results across the model architecture families we implemented and tested and compares our models and their corresponding best-performing counterparts. Our models demonstrate superior predictive performance compared to results reported in prior research. Moreover, the models presented in this work are assessed for generalization and fairness. In contrast, many prior works, including those referenced in Table 3, primarily report accuracy alone, offering a limited perspective on system performance.

Table 3. Provides a brief comparison of other published algorithms using deep learning and similar databases.

AUTHOR	ARCHITECTURE	TESS	RAVDESS	CREMA-D	IEMOCAP
THIS WORK	3xTDCNN1D, 2xLSTM, 2xDense	99.70%			
THIS WORK	2xTDCNN1D, 2xLSTM, 2xDense	99.70%			
THIS WORK	3xTDCNN1D, 2xLSTM, 2xDense	99.60%			
THIS WORK	DSCNN2D	99.30%			
DAREKARA & DHANDE [43]	1xANN, 1xPSO-FF		88.70%		
THIS WORK	2xTDCNN1D, 2xLSTM, 2xDense		88.00%		
THIS WORK	3xTDCNN1D, 2xLSTM, 2xDense		87.20%		
THIS WORK	3xCNN1D, 1xLSTM, 2xDense			86.50%	
THIS WORK	3xCNN1D, 1xLSTM, 2xDense			86.50%	
ZHAO ET AL. [5]	3xDCNN, 2xLSTM				86.16%
THIS WORK	3xCNN1D, 1xLSTM, 2xDense		84.70%		
THIS WORK	3xCNN1D, 1xLSTM, 2xDense			84.00%	
THIS WORK	3xCNN1D, 1xLSTM, 2xDense				83.60%
THIS WORK	3xTDCNN1D, 2xLSTM, 2xDense			83.10%	
LI ET AL. [45]	2xCNN, BLSTM, 2xAttention				82.80%
THIS WORK	3xTDCNN1D, 2xLSTM, 2xDense		82.40%		
THIS WORK	3xCNN1D, 1xLSTM, 2xDense				82.30%
THIS WORK	3xCNN1D, 1xLSTM, 2xDense				81.60%
THIS WORK	3xCNN1D, 1xLSTM, 2xDense				81.60%
MEKRUKSAVANICH ET AL. [46]	6xDCNN		75.83%		
MEKRUKSAVANICH ET AL. [46]	6xDCNN			65.77%	
LATIF ET AL. [30]	2xVAE, 4xLSTM 2				64.93%
MIRSAMADI ET AL. [47]	LSTM, ATTN / 4, 3, 3, 4, 4, 4				63.50%
SAHU ET AL. [48]	GAN, SVM				60.29%
KIM ET AL. [7]	LSTM, MTL / 3, 3, 2				56.90%
MEKRUKSAVANICH ET AL. [46]	6xDCNN	55.71%			
HAN ET AL. [44]	3xDCNN				54.30%
LATIF ET AL. [49]	2xLSTM, GAN				53.76%
CHATZIAGAPI ET AL. [50]	DCNN (VGG19), GAN / 19				53.60%
ESKIMEZ ET AL. [51]	CNN, VAE / 5, 6, 4, 10, 5				48.54%

In the next step and for bias assessment, we have decoupled Female specker data samples from Male speakers and compared their performance per speaker. As can be found from Figure 2, the performance of models is not consistent and robust w.r.t. speaker's gender. However, under constrained training setup, i.e., fixed training epochs and batch size, and barely employing our augmentation approach, not only can improve the predictive performance of models but also develop

doi:10.20944/preprints202412.2422.v1

their robustness across different gender and reduce the bias that can be measured by $\Delta(F\_1\ score)$, of "Both", "Male" and "Female" speakers. Additionally, our super corpora augmentation approach is effective regardless of conventional preprocessing approaches, reducing the overall computation cost of SER systems.

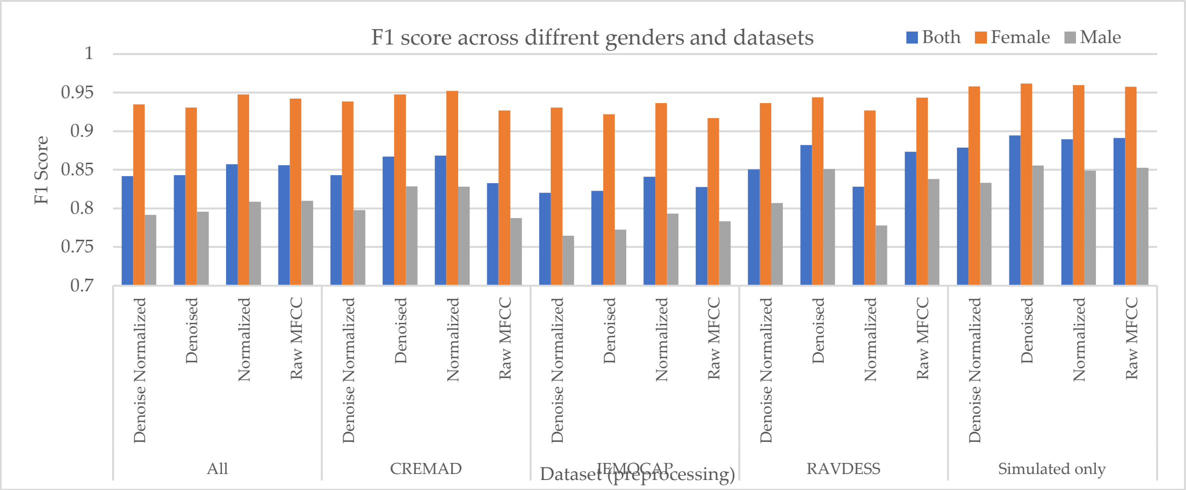

Figure 2. Cross-gender performance of models on each dataset. For each dataset, the effect of normalization and denoising has been reported.

In the following and to demonstrate the cross-corpora effectiveness of our approach, we present the performance of previously mentioned models when trained with each simulated dataset, TESS, RAVDESS, IEMOCAP, and when trained with our "simulated only" super corpora and tested against IEMOCAP as our excluded dataset.

As Figure 3 shows, regardless of the inductive biases that fit each previously mentioned model to the training dataset, all models suffer from exposure to out-of-distribution datasets. This is critical in training and deploying models with a specific dataset to real-world problems. However, training models with a mixture of datasets can improve the generalization of models to out-of-distribution scenarios, even where the model has not seen any data sampled from the test data distributions.

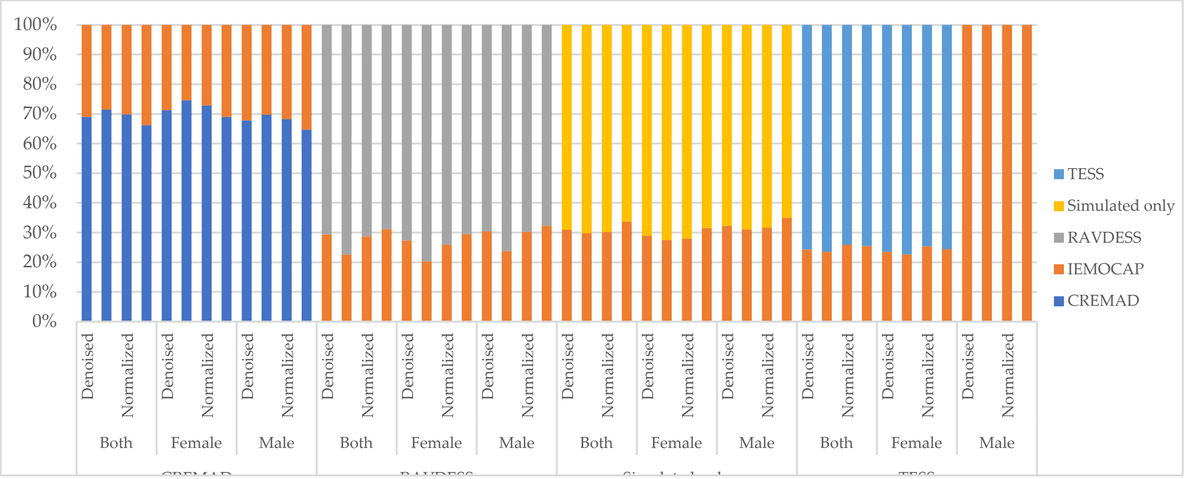

Figure 3. Weighted F1 score of models when tested with train dataset and IEMOCAP as out-of-distribution dataset.

5.2. Emotion-level bias mitigation

At it has been shown in Figures 4–6, our proposed data augmentation approach to creating a super corpus is effective in mitigating biases at the emotion level; By integrating this augmented dataset, the overall fidelity of the data has improved significantly. Specifically, the inter-class performance variation, denoted by

$$\Delta\left(M_{(i,j)}\right); \forall i,j \in \{emotions\} \text{ and } M \in \{F_1, Recall, Precision\}$$

where $emotions = \{Sad, Happy, \ Angry, \ Neutral\}$ is effectively minimized.

This indicates that our approach reduces disparities across emotion categories, resulting in a more balanced and robust model performance. The improvement is evident through consistent evaluation metrics, reflecting enhanced model fairness and reliability.

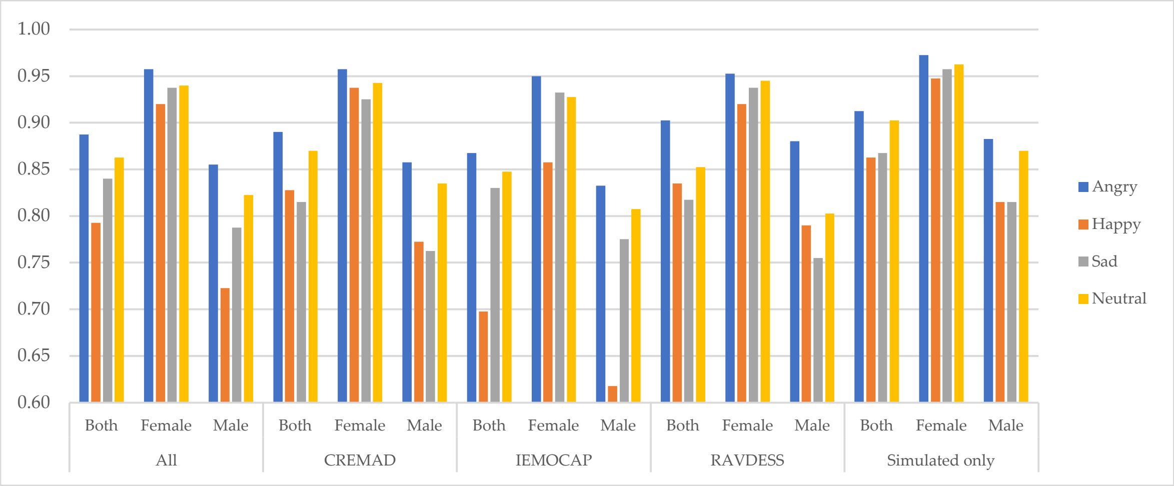

Figure 4. Average F1 score per database and gender across different emotions.

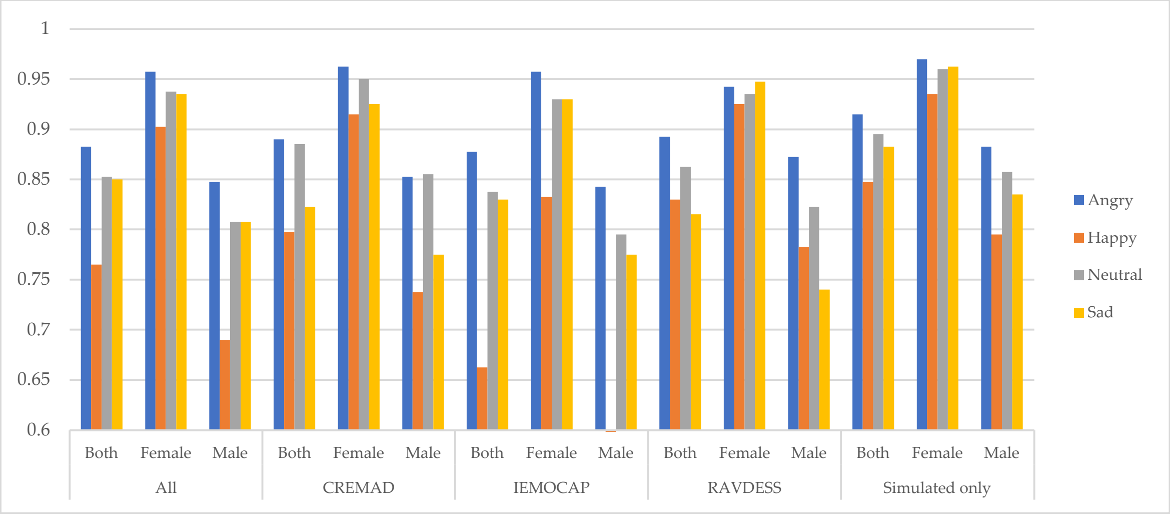

Figure 5. Average recall per database and gender across different emotions.

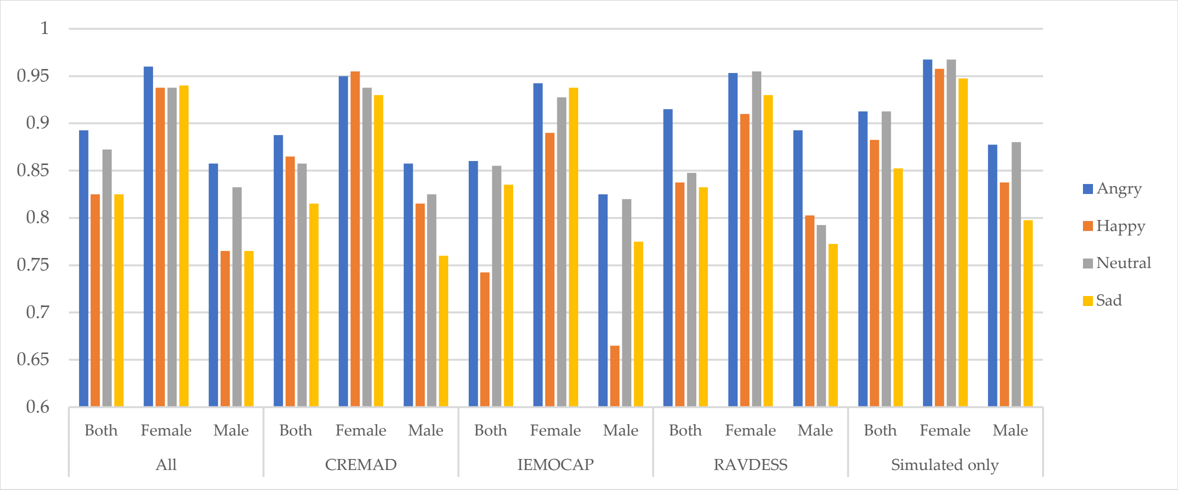

Figure 6. Average precision per database and gender across different emotions.

5.3. Downstream Bias

Finally, in this section, we investigate whether downstream biases of best models, as reported in Section 5.1, vary when trained with individual datasets compared to our augmentation super corpus. To measure downstream biases, we report TPB_y for the best model(according to Table 1) across emotions of our datasets where, as mentioned in Section 4.3, if a classifier predicts, e.g., "Angry" for a Male speaker much more often than for a Female speaker, the TPR ratio for the "Angry" class is low. TPB = 1 implies an unbiased situation.

As Figures 7–10 illustrate, regardless of preprocessing methods and training datasets, we have witnessed lower TPB (0.85 < TPB < 1.2) for Angry and Neutral emotions compared to Happy and Sad(0.6 < TPB < 1.5). This means our best-performing classifier is transferring bias on Happy and Sad emotions, and the training data has a high impact on the performance of models in Happy and Sad detection. Additionally, as can be found from the reported Figures, different preprocessing approaches, i.e., Denoising, Normalization, or both, cannot shift the bias in models. However, our augmentation approach (Simulated only or All) can effectively shift the bias in the model and push biases in the desired direction and, in some cases, fully de-bias the models, e.g., Simulated only/All.

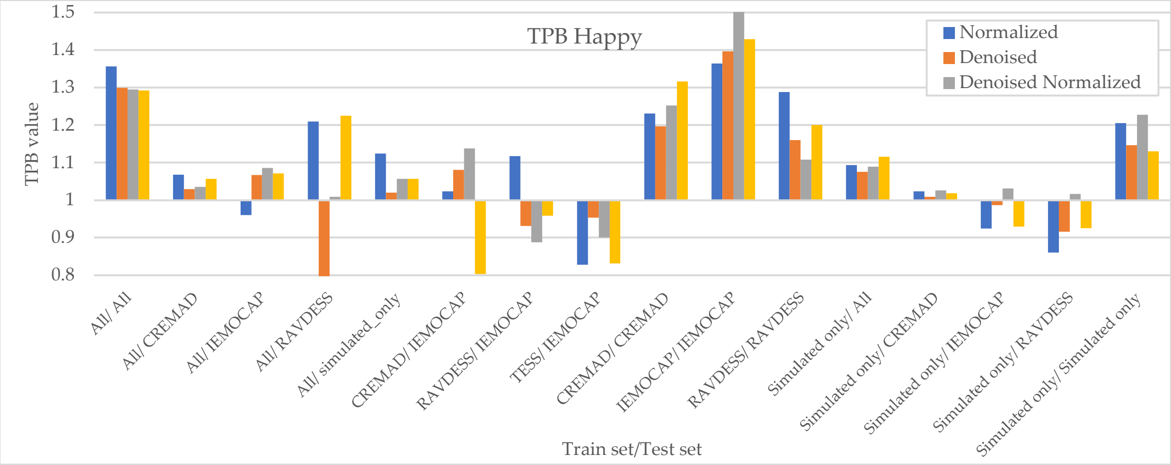

Figure 7. Measured downstream bias per 'Happy' w.r.t. to different preprocessing approaches and across different trainset/test sets.

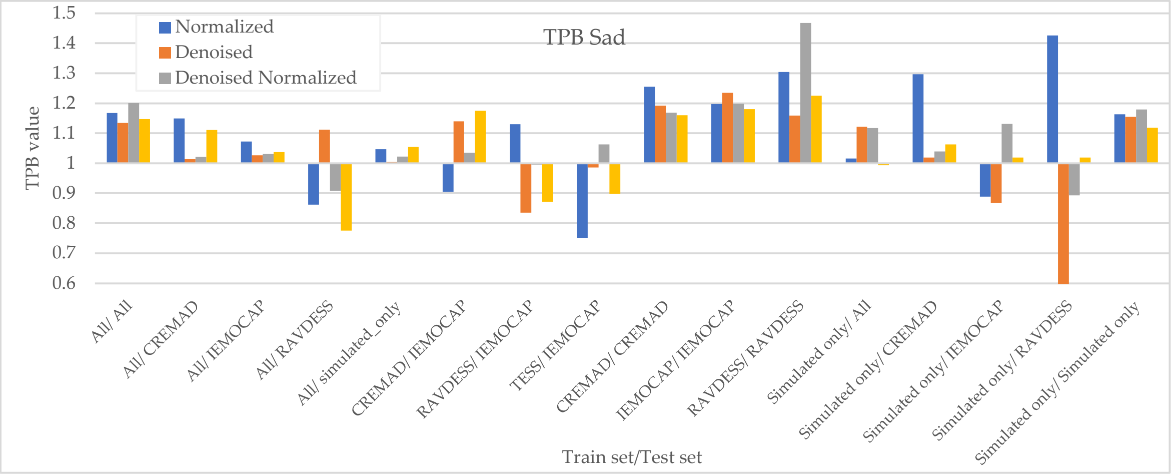

Figure 8. Measured downstream bias per 'Sad' w.r.t. to different preprocessing approaches and across different train/test sets.

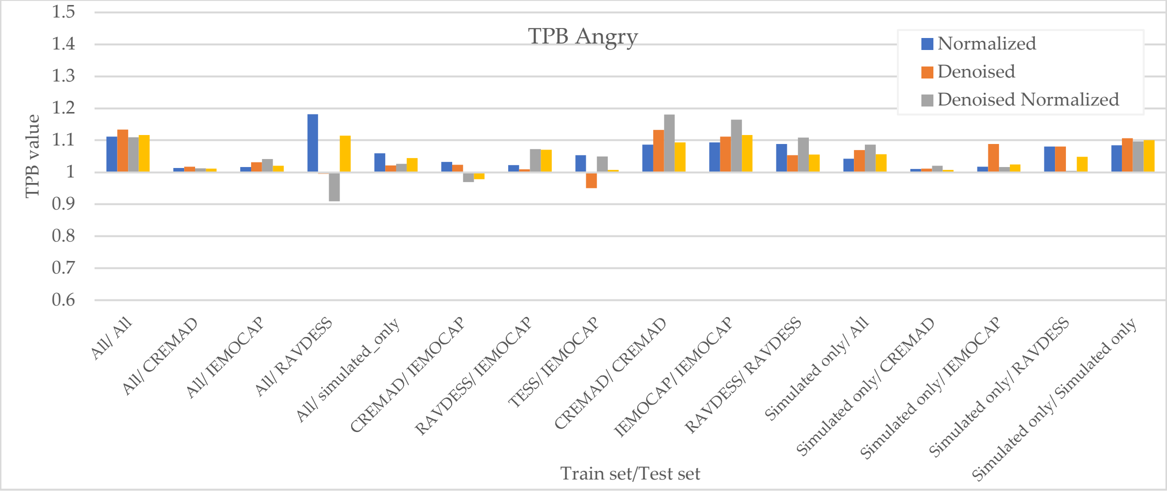

Figure 9. Measured downstream bias per 'Angry' w.r.t. to different preprocessing approaches and across different trainset/test sets.

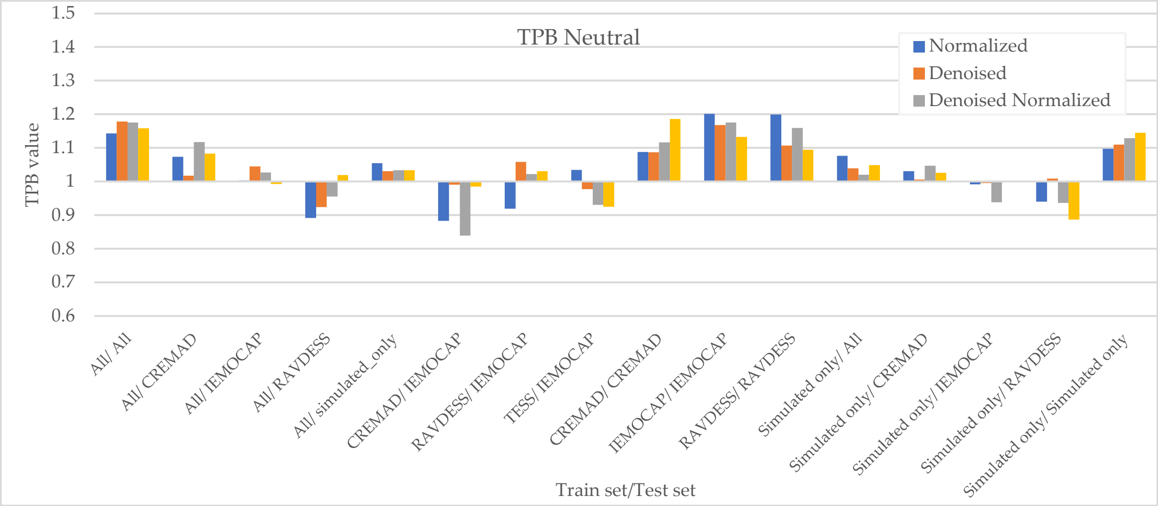

Figure 10. Measured downstream bias per 'Neutral' w.r.t. to different preprocessing approaches and across different trainset/test set.

6. Conclusions

This paper introduces a low-cost data augmentation approach for SER systems, addressing critical gaps in their evaluation and performance. Through extensive experiments, we demonstrate that relying solely on F1 score and accuracy metrics provides an incomplete picture of SER systems' effectiveness. Our findings emphasize the need for a holistic evaluation framework that incorporates additional performance metrics and assesses fairness, robustness, and generalization. This approach highlights limitations in prior work, which often overstate system effectiveness by focusing narrowly on accuracy.

Our analysis reveals that even top-performing models exhibit inconsistent performance across gender groups (Female, Male), struggle with out-of-distribution samples, and show variability when analyzing different emotions within the same dataset distribution. Despite their strong F1 and accuracy metrics, these models continue to reflect biases inherent in the data. Moreover, standard preprocessing methods, such as denoising and normalization, whether used individually or in combination, fail to adequately address these biases.

To tackle these challenges, we introduced a super corpus that significantly augments and diversifies the dataset pool, enabling broader applicability and enhancing robustness in SER models. Additionally, our detailed examination of model robustness in relation to speaker gender and cross-corpora scenarios offers valuable insights into mitigating biases and improving generalization.

Our results further illustrate that the proposed augmentation approach, whether using simulated datasets alone or coupled with preprocessing strategies, effectively reduces or even eliminates bias in some cases, steering the model's behavior in the desired direction. By demonstrating how our super corpus, when integrated with preprocessing strategies, enhances generalization, mitigates gender bias, and improves robustness to out-of-distribution data samples, we establish a comprehensive foundation for advancing SER systems.

This work not only establishes a strong baseline for future research but also underscores the importance of fairness, robustness, and reliability as essential dimensions of SER system evaluation alongside traditional performance metrics.

References

1. B. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," Sensors, vol. 21, no. 4, 2021 Feb 10.
2. E. Furey and J. Blue, "Alexa, Emotions, Privacy, and GDPR," in British HCI, 2018.
3. B. Schüller, G. Rigoll, and M. Lang, "Hidden Markov Model-based Speech Emotion Recognition," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, 2023.
4. B. Schüller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.
5. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Elsevier Biomedical Signal Processing and Control, vol. 47, pp. 312-323, 2019.
6. Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schüller, "Speech Emotion Classification Using Attention-Based LSTM," IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 27, no. 11, pp. 1675-1685, November 2019.
7. J. Kim, G. Englebienne, K. P. Truong and V. Evers, "Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning," in INTERSPEECH, 2017.
8. P. Harár, R. Burget and M. Kishore Dutta, "Speech Emotion Recognition with Deep Learning," in 4th International Conference on Signal Processing and Integrated Networks (SPIN), 2017.
9. M. Welling, "Do we still need models or just more data and compute?" University of Amsterdam, April 2019. Online]. Available: https://staff.fnwi.uva.nl/m.welling/wp-content/uploads/Model-versus-Data-AI-1.pdf.

10. J. Baxter, "Learning, A Model of Inductive Bias," Journal of Artificial Intelligence Research, vol. 12, p. 149–198, 2011.
11. A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 478, no. 2266, 2022.
12. Z. Zhang, F. Weninger, M. Wöllmer and B. Schüller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.
13. B. Schüller, D. Arsic, G. Rigoll, M. Wimmer and Radig, "Audiovisual Behavior Modeling by Combined Feature Spaces," in International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007.
14. B. Schüller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," Image and Vision Computing, vol. 27, no. 12, pp. 1760-1774, November 2009.
15. I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording, and verification of a Danish emotional speech database," in EUROSPEECH, Rhodes, Greece, 1997.
16. O. Martin, I. Kotsia, B. M. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," in 22nd International Conference on Data Engineering Workshops (ICDEW'06), 2006.
17. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, L. Devillers, S. Abrilian, A. Batliner, N. Amir, K. Karpouzis and J.-C. Martin, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 2007.
18. M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in IEEE International Conference on Multimedia and Expo, Hannover, Germany, 2008.
19. R. Milner, M. A. Jalal, R. W. M. Ng, and T. Hain, "A Cross-Corpus Study on Speech Emotion Recognition," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019.
20. W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," Complex & Intelligent Systems, vol. 7, pp. 1845-1854, 2021.
21. P. J. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database," University of Surrey, Guildford, 2014.
22. S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages," in 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2018.
23. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A database of German emotional speech," in INTERSPEECH, Lisbon, Portugal, 2005.
24. G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO Corpus: an Italian Emotional Speech Database," in Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, May 2014.
25. F. Eyben, K. R. Scherer, B. W. Schüller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 2016.
26. N. Braunschweiler, R. Doddipatla, S. Keizer and S. Stoyanchev, "A Study on Cross-Corpus Speech Emotion Recognition and Data Augmentation," Cartagena, Colombia, 2021.
27. 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, p. 335–359, December 2008.
28. 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, 2018.
29. A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

30. S. Latif, R. Rana, S. Khalifa, R. Jurdak and B. Schüller, "Self-Supervised Adversarial Domain Adaptation for Cross-Corpus and Cross-Language Speech Emotion Recognition," IEEE Transactions on Affective Computing, 2022.
31. J. Meyer, L. Rauchenstein, J. D. Eisenberg and N. Howell, "Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications," in 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, 2020.
32. S. Feng, O. Kudina, B. M. Halpern and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," in INTERSPEECH 2021, 2021.
33. C. Liu, M. Picheny, L. Sarı, P. Chitkara, A. Xiao, X. Zhang, M. Chou, and A. Alvarado, "Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022.
34. S. Steidl, "Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech," PhD thesis, Universität Erlangen-Nürnberg, Germany, 2009.
35. I. Sneddon, M. McRorie, G. McKeown and J. Hanratty, "The Belfast induced natural emotion database," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 32-41, 2012.
36. M. Gnjatovic and D. Rosner, "Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus," IEEE Transactions on Affective Computing, vol. 1, no. 2, pp. 132-144, 2010.
37. K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto emotional speech set," Canadian Acoustics - Acoustique Canadienne, vol. 39, no. 3, pp. 182-183, 2011.
38. H. Cao, D. G. Cooper, M. Keutmann, R. 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, October 2014.
39. D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," Journal of the Acoustical Society of America, vol. 90, no. 4, p. 1841–1856, October 1991.
40. M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. Tauman Kalai, "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting," in ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*), 2019.
41. R. Steed, S. Panda, A. Kobren and M. Wick, "Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models," in 60th Annual Meeting of the Association for Computational Linguistics, 2022.
42. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi and N. Ismail, "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks," 2020 6th International Conference on Wireless and Telematics (ICWT), Yogyakarta, Indonesia, pp. 1-6, 2020
43. R. V. Darekara and A. P. Dhande, "Emotion recognition from Marathi speech database using adaptive artificial neural network," Biologically Inspired Cognitive Architectures, vol. 25, pp. 35-42, 2018.
44. K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," in INTERSPEECH, 2024.
45. Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in INTERSPEECH 2019: Training Strategy for Speech Emotion Recognition, 2019.
46. S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative Emotion Recognition using Deep Learning for Thai Language," in Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Pattaya, Thailand, 2020.
47. S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017.
48. S. Sahu, R. Gupta, and C. Espy-Wilson, "On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks," Interspeech 2018, no. http://dx.doi.org/10.21437/Interspeech.2018-1883, 2018.

49. S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study," in INTERSPEECH, 2018.
50. A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data Augmentation Using GANs for Speech Emotion Recognition," in INTERSPEECH 2019: Speech Signal Characterization 1, Graz, Austria, 2019.
51. S. E. Eskimez, Z. Duan and W. Heinzelman, "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.