






Article

Cross-Corpus Multilingual Speech Emotion Recognition: Amharic vs. Other Languages

Ephrem Afele Retta ¹, Richard Sutcliffe ^{2,*}, Jabar Mahmood ^{3,4}, Michael Abebe Berwo ⁴,
Eiad Almekhlafi ¹, Sajjad Ahmad Khan ⁵, Shehzad Ashraf Chaudhry ^{6,7}, Mustafa Mhamed ^{1,8}
and Jun Feng ¹

- ¹ School of Information Science and Technology, Northwest University, Xi'an 710127, China; afele@stumail.nwu.edu.cn (E.A.R.); e_almekhlafi@stumail.nwu.edu.cn (E.A.); mustafamhamed2099@gmail.com (M.M.); fengjun@nwu.edu.cn (J.F.)
- ² School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
- ³ Faculty of Computing and Information Technology, University of Sialkot, Sialkot 51040, Punjab, Pakistan; jabarmehmood@outlook.com or 2019024906@chd.edu.cn
- ⁴ School of Information and Engineering, Chang'an University, Xi'an 710064, China; 2019024902@chd.edu.cn
- ⁵ Computer Engineering Department, Hoseo University, Asan 31499, Republic of Korea; dr.sajjadkhan19@gmail.com
- ⁶ Department of Computer Science and Information Technology, College of Engineering, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates; ashraf.shehzad.ch@gmail.com
- ⁷ Department of Software Engineering, Faculty of Engineering and Architecture, Nisantasi University, Istanbul 34398, Turkey
- ⁸ College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
- * Correspondence: rsutcl@essex.ac.uk



Citation: Retta, E.A.; Sutcliffe, R.; Mahmood, J.; Berwo, M.A.; Almekhlafi, E.; Khan, S.A.; Chaudhry, S.A.; Mhamed, M.; Feng, J. Cross-Corpus Multilingual Speech Emotion Recognition: Amharic vs. Other Languages. *Appl. Sci.* **2023**, *13*, 12587. <https://doi.org/10.3390/app132312587>

Academic Editor: Douglas O'Shaughnessy

Received: 24 October 2023

Revised: 17 November 2023

Accepted: 19 November 2023

Published: 22 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: In a conventional speech emotion recognition (SER) task, a classifier for a given language is trained on a pre-existing dataset for that same language. However, where training data for a language do not exist, data from other languages can be used instead. We experiment with cross-lingual and multilingual SER, working with Amharic, English, German, and Urdu. For Amharic, we use our own publicly available Amharic Speech Emotion Dataset (ASED). For English, German and Urdu, we use the existing RAVDESS, EMO-DB, and URDU datasets. We followed previous research in mapping labels for all of the datasets to just two classes: positive and negative. Thus, we can compare performance on different languages directly and combine languages for training and testing. In Experiment 1, monolingual SER trials were carried out using three classifiers, AlexNet, VGGE (a proposed variant of VGG), and ResNet50. The results, averaged for the three models, were very similar for ASED and RAVDESS, suggesting that Amharic and English SER are equally difficult. Similarly, German SER is more difficult, and Urdu SER is easier. In Experiment 2, we trained on one language and tested on another, in both directions for each of the following pairs: Amharic↔German, Amharic↔English, and Amharic↔Urdu. The results with Amharic as the target suggested that using English or German as the source gives the best result. In Experiment 3, we trained on several non-Amharic languages and then tested on Amharic. The best accuracy obtained was several percentage points greater than the best accuracy in Experiment 2, suggesting that a better result can be obtained when using two or three non-Amharic languages for training than when using just one non-Amharic language. Overall, the results suggest that cross-lingual and multilingual training can be an effective strategy for training an SER classifier when resources for a language are scarce.

Keywords: speech emotion recognition; multilingual; cross-lingual; feature extraction

1. Introduction

Emotions assist individuals to communicate and to comprehend others' points of view [1]. Speech emotion recognition (SER) is the task of comprehending emotion in a voice signal, regardless of its semantic content [2].

SER datasets are not available in all languages. Moreover, the quantity and quality of the training data that are available varies considerably from one language to another. For example, when evaluated across several datasets, differences in corpus language, speaker age, labeling techniques, and recording settings significantly influence model performance [3,4]. This encourages the development of more robust SER systems capable of identifying emotion from data in different languages. This can then permit the implementation of voice-based emotion recognition systems in real time for an extensive variety of industrial and medical applications.

The majority of research on SER has concentrated on a single corpus, without considering cross-lingual and cross-corpus effects. One reason is that, in comparison to the list of spoken languages, we only have a small number of corpora for the study of speech analysis [5]. Furthermore, even when only considering the English language, accessible resources vary in quality and size, resulting in the dataset sparsity problem observed in SER research. In such instances, learning from a single data source makes it challenging for SER to function effectively. As a result, more adaptable models that can learn from a wide range of resources in several languages are necessary for practical applications.

Several researchers have investigated cross-corpus SER in order to enhance classification accuracy across several languages. These works employed a variety of publicly accessible databases to highlight the most interesting trends [6]. Even though some research has addressed the difficulty of cross-corpus SER, as described in Schuller et al. [6], the challenges posed by minority languages such as Amharic have not been investigated.

Amharic is the second-largest Semitic language in the world after Arabic, and it is also the national language of Ethiopia [7]. In terms of the number of speakers and the significance of its politics, history, and culture, it is one of the 55 most important languages in the world [8]. Dealing with such languages is critical to the practicality of next-generation systems [9], which must be available for many languages.

In our previous work [10], we created the first spontaneous emotional dataset for Amharic. This contains 2474 recordings made by 65 speakers (25 male, 40 female) and uses five emotions: fear, neutral, happy, sad, and angry. The Amharic Speech Emotion Dataset (ASED) is publicly available for download (https://github.com/Ethio2021/ASED_V1 accessed on 17 October 2023). This dataset allows us to carry out the work reported here.

The contributions of this paper are as follows:

- We investigate different scenarios for monolingual, cross-lingual, and multilingual SER using datasets for Amharic and three other languages (English, German, and Urdu).
- We experiment with a novel approach in which a model is trained on data in several non-Amharic languages before being tested on Amharic. We show that training on two non-Amharic languages gives a better result than training on just one.
- We present a comparison of deep learning techniques in these tasks: AlexNet, ResNet50, and VGGE.
- To the best of our knowledge, this is the first work that shows the performance tendencies of Amharic SER utilizing several languages.

The structure of this paper is as follows: Section 2 presents previous work. Section 3 explains our approach, datasets, and feature extraction methods for SER. Section 4 presents the proposed deep learning architecture and experimental settings. Section 5 describes the experiments and outcomes. Finally, Section 6 gives conclusions and next steps.

2. Related Work

Over the last two decades, much important research has been conducted on speaker-independent SER. This work has shown that several factors influence accuracy, including

the dataset utilized, the features extracted, and the classifier network employed to predict emotions. Sailunaz et al. [11] present a thorough survey of SER work. However, while there has been preliminary research on enhancing the robustness of SER by combining multiple emotional speech corpora to form the training set and thereby minimizing data scarcity, there is a shortage of studies on multilingual cross-corpus SER [6,12]. In the following, we first summarize related cross-lingual work. After that, we outline multilingual studies. Information about all the research is shown in Table 1.

Concerning cross-lingual studies, Lefter et al. [13] carried out an early study in which they trained an SER classifier on one or more datasets and then tested it on another. In a cross-lingual setting, training on ENT and testing on DES gave the lowest Equal Error Rate for Anger (29.9%).

Albornoz et al. [9] proposed an SER classifier for emotion detection, focusing on emotion identification in unknown languages. The results showed what could be expected from a system trained with a different language, reaching 45% on average. The standard multi-class SVM performed better than the classifier implemented using Emotion Profiles (EP). The Standard Classifier (SC) reached 56.8%, whereas the Emotional Profile Classifier (EPC) obtained 52.1%.

Xiao et al. [14] examined SER for Mandarin Chinese vs. Western languages such as German and Danish. The authors concentrated on gender-specific SER and attained classification rates that were higher than chance but lower than baseline accuracy. The best classification rate in the cross-language family test on male speech samples (71.62%) was when the Chinese Dual-mode Emotional Speech Database (CDESDB) was used for training and Emo-DB was used for testing.

Sagha et al. [15] utilized language detection to improve cross-lingual SER. They found that using a language identifier followed by network selection, rather than a network trained on all existing languages, was superior for recognizing the emotions of a speaker whose language is unknown. On average, the Language IDentification (LID) approach for selecting training corpora was superior to using all of the available corpora when the spoken language was not known.

Meftah et al. [16] proposed Deep Belief Networks (DBNs) for cross-corpus SER and evaluated them in comparison with MLP via emotional speech corpora for Arabic (KSUEmotions) and English (EPST). Training on one dataset and testing on the other yielded similar results for both directions and both models. The best result was Arabic→English using DBNs (valence 53.22%, arousal 57.2%).

Latif et al. [17] extracted eGeMAPS features from their raw audio data. They used SVM with a Gaussian kernel to classify data into their respective categories. The best result came from training on EMO-DB and then testing on URDU (57.87%).

Latif et al. [18] also used eGeMAPS features, and they employed five different corpora for three different languages to investigate cross-corpus and cross-language emotion recognition using Deep Belief Networks (DBNs). IEMOCAP performed well on EMO-DB compared to FAU-AIBO even though both of the latter datasets are German.

Latif et al. [19] studied SER using languages from various language families, such as Urdu vs. Italian or German. The best cross-lingual results were obtained by training on URDU and testing on EMO-DB (65.3%) and the worst were by training on URDU and testing on SAVEE (53.2%).

Goel et al. [20] used transfer learning to carry out multi-task learning experiments and discovered that traditional machine learning architectures [5,21] can perform as well as deep learning neural networks for SER provided the researchers pick appropriate input features. Training the model on IEMOCAP and testing it on EMO-DB obtained the best performance (65%).

Zehra et al. [22] presented an ensemble learning approach for cross-corpus machine learning SER, utilizing the SAVEE, URDU, EMO-DB, and EMOVO databases. The method employed three of the most prominent machine learning algorithms, Sequential Minimal Optimization (SMO), Random Forest (RF), and Decision Tree (J48), plus a majority voting

mechanism. The ensemble approach was worse than the other classifiers except when training on EMOVO and testing on URDU (62.5%).

Duret et al. [23] used prosody prediction and employed eight different corpora for five European languages to investigate cross-lingual and multilingual emotion recognition using Wav2Vec2XLSR. The multilingual setup outperformed the monolingual one for all selected European languages, except English, by a very small margin.

Pandey et al. [24] proposed an SER classifier for emotion detection, focusing on learning emotions, irrespective of culture. They also used 3D Mel-Spectrogram features (henceforth referred to as MelSpec) and employed five different corpora for five languages to investigate cross-lingual emotion recognition using an Attention-Gated Tensor Factorized Neural Network (AG-TFNN). The best result was Fold2→German using a 3D TFNN. In addition, Fold5→Telugu had better performance than Fold4→Hindi, even though both languages are of Indian origin.

Table 1. Previous work on cross-lingual and multilingual SER (X = cross-lingual, M = multilingual, SVM = Support Vector Machine, SC = Standard Classifier, EPC = Emotional Profile Classifier, SMO = Sequential Minimal Optimization, DBN = Deep Belief Network, MLP = Multi-Layer Perceptron, GAN = Generative Adversarial Network, LSTM = Long Short-Term Memory, LR = Logistic Regression, RF = Random Forest, J48 = Decision Tree).

Ref	Methods Employed	Feature Extraction	Databases and Languages	Expts	Classes
[13]	SVM	Prosodic	EMO-DB (German), DES (Danish), ENT (English), SA (Afrikaans), RML (Mandarin)	XM	3
[9]	SVM, SC, EPC	Prosodic Various MFCC	English, Italian Persian, Punjabi, Urdu	X	6
[14]	SMO	Various MFCC	CDESD (Mandarin), EMO-DB, DES	X	Arousal Appraisal Space
[15]	SVM	Various MelSpec	EU-EmoSS (English, French, German, Spanish), VESD (Chinese), CASIA (Chinese)	X	Arousal Valence Plane
[16]	DBN, MLP	Low-level Acoustic	KSUEmotions (Arabic), EPST (English)	X	2
[17]	SVM	eGeMAPS	SAVEE (English), EMOVO (Italian), EMO-DB, URDU (Urdu)	XM	2
[18]	DBN	eGeMAPS	FAU-AIBO (German), IEMOCAP (English) EMO-DB,	X	2
[19]	GAN	eGeMAPS Various	SAVEE, EMOVO EMO-DB, SAVEE, EMOVO, URDU	XM	2
[20]	LSTM, LR, SVM	ISO9	EMOVO, EMO-DB, SAVEE, IEMOCAP, MASC (Chinese)	X	5

Table 1. Cont.

Ref	Methods Employed	Feature Extraction	Databases and Languages	Expts	Classes
[23]	CNN and Wav2Vec2-XLSR	prosody	IEMOCAP, CREMA-D (English), ESD (English), Synpaflex (French), Oreau (French), EMO-DB, EMOVO, emoUERJ (Portuguese), EMO-DB, eINTERFACE (English), IITKGP-SEHSC (Hindi), IITKGP-SESC (Telugu), ShEMO-DB (Persian)	X	4
[24]	AG-TFNN	MelSpec	SAVEE, URDU, EMO-DB, EMOVO	M	2
[22]	SMO, RF, J48, Ensemble	Spectral Prosodic eGeMAPS	SAVEE, URDU, EMO-DB, EMOVO	X	2

We now consider multilingual approaches in which several datasets in different languages are used for training. In addition to the cross-lingual experiments referred to earlier, Lefter et al. [13] also carried out some multilingual work in which they trained on various pairs or triples of datasets chosen from EMO-DB, DES, and ENT and then tested on each of these individually. The best result was obtained by training on all three and testing on EMO-DB (an Equal Error Rate of 20.5%).

Latif et al. [17] used four different corpora (SAVEE, EMOVO, EMO-DB, and URDU) for four different languages to investigate multilingual emotion recognition using Support Vector Machines (SVM). When training on EMO-DB, EMOVO, and SAVEE and testing on URDU, a result of 70.98% was achieved, which was higher than any pair of these datasets.

Latif et al. [19] also used SAVEE, EMOVO, EMO-DB, and URDU. The best performance was achieved by training on SAVEE, EMOVO, and URDU and testing on EMO-DB (68%). The worst performance was observed when training on the same three datasets and testing on EMOVO (61.8%).

Regarding the model used, Latif et al. [17], Albornoz et al. [9], Lefter et al. [13], and Sagha et al. [15] are all based on SVMs. Meftah et al. [16] and Latif et al. [18] utilized DBNs; Goel et al. [20], Duret et al. [23], and Pandey et al. [24] applied machine learning and deep learning methods; Zehra et al. [22] used ensemble methods; and, lastly, Xiao et al. [14] and Latif et al. [19] applied GAN and SMO, respectively. Concerning the earlier studies, we observe that the SVM algorithm performs poorly on large datasets. It also performs poorly in situations with more characteristics per data point, especially in multi-class situations. When attempting to extract features from a DBN plus low-level acoustic information vs. a DBN with eGeMAPS, the latter significantly outperformed the former. Additionally, deep learning models outperform conventional classifiers. However, the model of Goel et al. [20] extracts features quite well but requires a lot of training time. As previously indicated, an ensemble strategy only provided the best performance in one scenario. Furthermore, the existing techniques in SER lack preprocessing operations. We conclude that, across many datasets, a binary performance outperforms multiple classes. Moreover, none of the previous works have focused on the Amharic language.

Here, we first present the preprocessing strategy before describing the extraction of features from the signal. Second, we propose an architecture, based on the VGG model, which offers good results. Third, we provide a classification benchmark for Amharic and three non-Amharic languages using VGG and two other models. Finally, we contrast the

effectiveness of our novel training scenarios to demonstrate the efficiency of cross-lingual and multilingual approaches.

3. Approach

Many factors influence SER accuracy in a cross-corpus and multilingual context. The dataset utilized, the features extracted from the speech signals, and the neural network classifiers implemented to identify emotion are all essential aspects that may significantly impact the results. Our SER method is summarized in Figure 1. We use four corpora (ASED, RAVDESS, EMO-DB, and URDU) to test SER in Amharic, English, German, and Urdu, respectively.

One difficulty faced with this research is that datasets use different sets of emotion labels, as can be seen in Table 2. Following previous work [25,26], we address this by mapping labels into just two classes, positive valence and negative valence, as indicated in the table. For example, ASED uses five emotions. For this dataset, therefore, we map the two emotions Neutral and Happy to positive valence, and the three remaining emotions, Fear, Sadness, and Angry, to negative valence. Analogous mappings are performed for the other datasets.

Further details are provided below concerning the chosen datasets, the feature extraction approach, and the classifiers used.

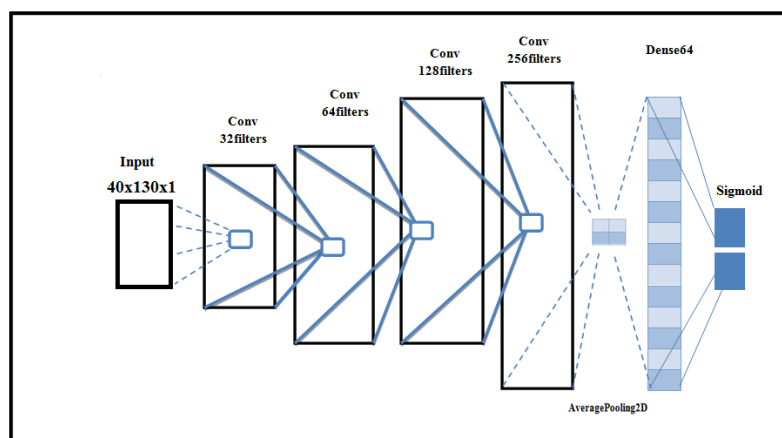


Figure 1. Network architecture of proposed VGG based on well-known VGG model.

3.1. Speech Emotion Databases

ASED [10] is for Amharic and was created by the authors in previous work. It uses five emotions and consists of 2474 recordings made by 65 speakers (25 male, 40 female). Recording was performed at 16 kHz and 16 bits. The ASED dataset is accessible to the public for research purposes (see URL earlier).

RAVDESS [27] is for English, uses eight emotions, and contains just two sentences. The 24 speakers (12 male, 12 female) are professional actors. Interestingly, emotions in this dataset are ‘self-induced’ [28], rather than acted. Moreover, there are two levels of each emotion. There are 4320 utterances. Project investigators selected the best two clips for each speaker and each emotion. Recording was at 48 kHz and 16 bits, and it was carried out in a professional recording studio at Ryerson University.

EMO-DB [29] is for German, uses five emotions, and contains ten everyday sentences: five made of one phrase, and five made of two phrases. There are ten speakers (five male, five female), nine of whom are qualified in acting, and about 535 raw utterances in total. Recording was performed at 16 kHz and 16 bits and was carried out in the anechoic chamber of the Technical Acoustics Department at the Technical University Berlin.

URDU [17] is for Urdu, uses four emotions, and comprises 400 audio recordings from Urdu TV talk shows. There are 38 speakers (27 male, 11 female). Emotions are not acted but occur naturally during the conversations between guests on the talk shows.

Table 2. Datasets used in the experiments. The table also shows the mapping from the emotion labels in each dataset into just two valence labels which can be used across them all: positive and negative.

Aspect	ASED	RAVDESS	EMO-DB	URDU
Language	Amharic	English	German	Urdu
Recordings	2474	1440	535	400
Sentences	27	2	10	-
Participants	65	24	10	38
Emotions	5	8	7	4
Positive valence	Neutral, Happy	Neutral, Happy, Calm, Surprise	Neutral, Happiness	Neutral, Happy
Negative valence	Fear, Sadness, Angry	Fear, Sadness, Angry, Disgust	Anger, Sadness, Fear, Disgust, Boredom	Angry, Sad
References	[10]	[27]	[29]	[17]

3.2. Data Preprocessing

Before proceeding to feature extraction, a number of pre-processing steps were performed on the datasets, as shown in Figure 2. Recordings were first downsampled to 16 kHz and converted to mono. As can be seen in Table 3, most of the sound clips in the datasets are 5 s in length or less. A few are longer than this. Therefore, for our experiments, we extended any shorter clips to 5 s by adding silence to the end. Conversely, any longer clips were cut off in order to make them exactly 5 s long.

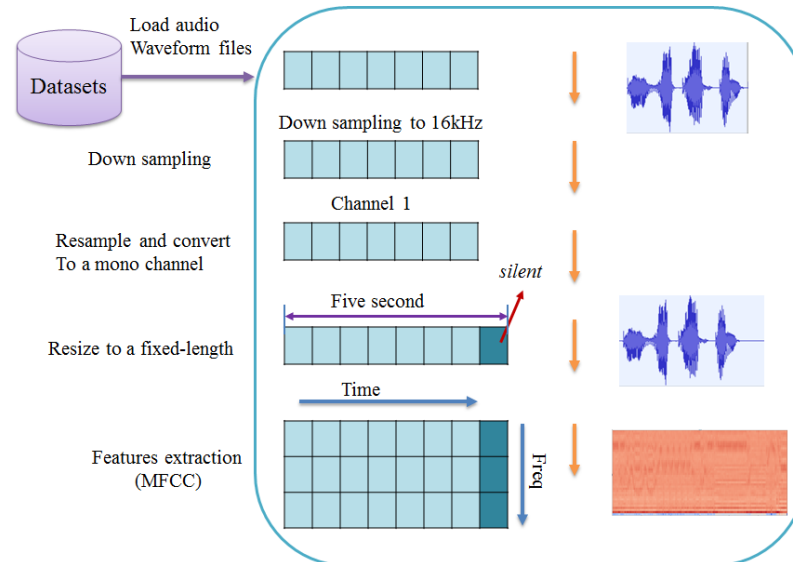


Figure 2. Data preprocessing.

Table 3. Statistics of original clip lengths in seconds for all datasets. In the table, 1–2.0 means $1 \text{ s} \leq d < 2 \text{ s}$.

Duration (s)	ASED	EMO-DB	RAVDESS	URDU
1–2.0		126		
2–3.0	850	224		200
3.0–4	1624	136	1440	200
4.0–5		24		
5.0–6		20		
6.0–7		3		
7.0–8		1		
8.0–9		1		
STD	0.444	1.067	0	0.5
Mean	2.967	2.267	3	2.5

3.3. Feature Extraction for SER

A vast amount of information reflecting emotional characteristics is present in the speech signal. One of the key issues within SER research is the choice of which features should be used.

Previously, traditional feature extraction methods, such as prosodic features, were used for SER [30,31], including the variance, intensity, spoken word rate, and pitch. However, some traditional features are shared across different emotions, as discussed by Gangamohan et al. [30]. For example, as observed in Table 11.2 of Gangamohan et al., angry and happy utterances have similar trends in F0 and speaking rate, compared to neutral speech.

Manually extracted traditional features may work well with traditional classification methods in machine learning, where a set of features or attributes describes each instance in a dataset [32]. In contrast, however, deep learning can itself determine which features to focus on to recognize verbal emotions. Finding some sets of feature vectors or properties that can give a compact representation of the input audio signal has therefore become the main aim of feature extraction methods. The spectrum extraction methods convert the input sound waveform to some discrete shape or feature vector. Normally, the speech signal is not static but when looking at a short period of time, it acts as a static signal. This short, detached snap is called a frame. The acoustic model extracts features from the frames [33,34]. Feature extraction deals with obtaining useful information for reference by removing irrelevant information. These extracted feature vectors are fed into deep learning models. In short, spectrum extraction methods can convert audio signals into vectors that deep learning models can handle. The model can then be trained to learn the features of each emotion and hence classify it. Overall, this is one reason why deep learning models can perform better than machine learning models.

After reviewing many works on SER, it is clear that Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in audio classification and speech emotion recognition [35]. An MFCC is a coefficient that expresses the short-term power spectrum of a sound. It uses a series of steps to imitate the human cochlea, thereby converting audio signals. The Mel scale is significant because it approximates the human perception of sound instead of being a linear scale [36]. In our previous work [10], we compared MFCCs to alternatives and found them to be the best. This is the reason we chose MFCC features for the present study.

4. Architectures and Settings

Most prior research uses CNN-based models for SER [37]. Among such models, the notable ones include AlexNet [38], VGG [39,40], and ResNet50 [41,42]. This section provides a short overview of the models. Our proposed model, VGGE, is a variant of VGG.

- **AlexNet** is one of the famous CNN models used in applications such as image classification and recognition and is widely employed for SER classification [43]. It achieved an outstanding result at the ImageNet competition in 2012 [44].

- **VGG** [39] appeared in 2014, created by the Oxford Robotics Institute. It is well known that the early CNN layers capture the general features of sounds such as wavelength, amplitude, etc., and later layers capture more specific features such as the spectrum and the cepstral coefficients of waves. This makes a VGG-style model suitable for the SER task. After some experimentation, we found that a model based on VGG but using four layers gave the best performance. We call this proposed model VGGE and use it for our experiments. Figure 1 shows the settings for VGGE.
- **ResNet** [42] was launched in late 2015. This was the first time that networks with more than a hundred layers were trained. Subsequently, it has been applied to SER classification [41].

Concerning the experimental setup, the standard code for AlexNet and ResNet50 was downloaded and used for the experiments. For VGGE, the network configuration was altered, as shown in Figure 1. For the other models, the standard network configuration and parameters were used.

In all experiments, the librosa v0.7.2 library [45] was used to extract MFCC features.

We used the Keras deep learning library, version 2.0, with a Tensorflow 1.6.0 backend to build the classification models. The models were trained using a machine with an NVIDIA GeForce GTX 1050. Our model employed the Adam optimization algorithm with categorical cross-entropy as the loss function; training was terminated after 100 epochs, and the batch size was set to 32.

5. Experiments

Two methods are primarily utilized for speaker-independent SER [46]: The first method is Leave One Speaker Out (LOSO) [21,47,48]. Here, when the corpus contains n speakers, we use $n - 1$ speakers for training and the remaining speaker for testing. For cross-validation, the experiment is repeated n times with a different test speaker each time. In the second method, the training and testing sets have been determined previously [17,49,50].

In our work, we followed the second approach. For the first monolingual experiment (train on a corpus, test on the same corpus), the data were split into training, testing, and validation sets randomly five times, ensuring each time that the split sets were speaker-independent. As shown in Table 4, all of the datasets were split into 70% train, 20% test, and 10% validation. In the first experiment, we also carried out a sentence-independent study in which the sentences used for training were not used for testing.

The second and third experiments are the cross-lingual experiment (train on a corpus in one language, test on a corpus in another language) and the multilingual experiment (train on two or three corpora joined together, each in a different non-Amharic language, and test on the Amharic ASED corpus). In these experiments, the speakers in the validation sets are not seen in the training sets. Moreover, the speakers in the testing set are by definition not the same as those in the training and validation sets, as they are from different datasets.

Figure 3 shows a label distribution that is balanced across partitions. The performance of the proposed classification of Amharic language data used in monolingual, cross-lingual, and multilingual SER experiments is evaluated using F1-score and accuracy. We have shared the file names for the audio files that belonged to the train, validation, and test partitions in the experiments (<https://github.com/Ethio2021/File-names> accessed on 17 October 2023). For each experiment, the models were trained five times, and the average result was reported.

Table 4. Class distribution between the train, validation, and test partitions.

Datasets Labels	Train		Test		Validation	
	Positive	Negative	Positive	Negative	Positive	Negative
ASED	693	804	199	230	99	115
EMODB	95	118	27	34	14	17
RAVDESS	456	524	140	160	56	64
URDU	140	138	40	40	20	22

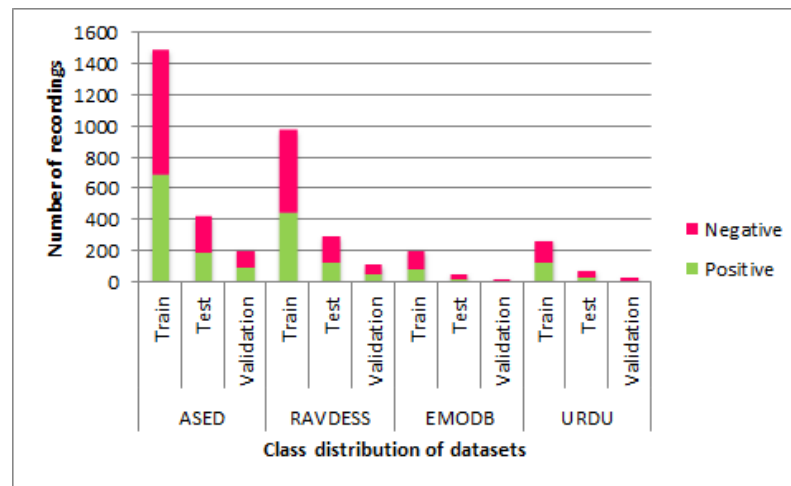


Figure 3. Class distribution within Datasets.

6. Experiment 1: Comparison of SER Methods for Monolingual SER

6.1. Outline

The aim was to carry out an initial comparison of the proposed VGGE model with the two existing models discussed above, AlexNet and ResNet50. Four datasets were used: ASED, RAVDESS, EMO-DB, and URDU. To allow comparison with the other experiments, the emotion labels for each dataset were mapped onto just two labels, positive valence and negative valence, as shown by the scheme in Table 2. This follows the standard approach found in other work [25,26]. When comparing to other papers, we should bear in mind the label mapping that we needed to adopt in order to undertake the later cross-lingual and multilingual experiments. Looking at the table, we can see that the ASED, RAVDESS, EMO-DB, and URDU datasets originally had five, eight, seven, and four emotion classes, respectively, and that these are now being mapped into just two classes: positive and negative emotions. This simplifies the task, which can account for higher performance figures than in other published works.

Experiment 1 has two parts. In Experiment 1.1, the groups of speakers used for training and testing were varied. In Experiment 1.2, the dataset sentences used for training and testing were varied.

6.2. Experiment 1.1: Independence of Speakers

The results of this experiment, expressed as accuracy, are shown in Table 5. Recall that each of the four datasets is monolingual and that we are training and testing on the same language here. We can see that VGGE was the best on ASED (Amharic) and EMO-DB (German), ResNet50 was the best on RAVDESS (English), and AlexNet was the best on URDU (Urdu).

Table 5. Experiment 1.1: Monolingual SER results expressed as accuracy for different models and datasets (train in one language, test in the same language). Training and testing speakers are varied in this experiment. All datasets are monolingual: the languages are Amharic (ASED), German (EMO-DB), English (RAVDESS), and Urdu (URDU).

Model	ASED	EMO-DB	RAVDESS	URDU
AlexNet	78.71	68.52	80.63	93.75
VGGE	84.76	85.19	83.13	70.00
ResNet50	84.13	79.63	84.38	90.00
Average	82.53	77.78	82.71	84.58

It is interesting to look at the average figures on the bottom row of the table. ASED (82.53%) and RAVDESS (82.71%) are very close, EMO-DB (77.78%) is 4.75% lower than ASED, and URDU (84.58%) is 2.05% higher than ASED. Generally, the differences are not that large when we consider that the languages have very different characteristics and that the datasets were created independently by different researchers. Moreover, recall that the original data are being mapped onto two sentiment classes from the original four to eight classes (see Section 6.1 and Table 2).

Subject to these points, we might conclude that Amharic and English monolingual mono-corpus SER are of similar difficulty, German is more difficult, and Urdu is easier. As languages, English and German are perhaps the most similar, since they are both within the Germanic branch of the Indo-European language group. Urdu is also Indo-European but from the Indo-Iranian branch. Finally, Amharic is from the Semitic branch of the Afro-Asiatic group.

6.3. Experiment 1.2: Independence of Sentences

Recall that the datasets all consist of different sentences spoken in every emotion, with the exception of the URDU dataset, based on TV talk show conversation, where individual sentences are not identified. Hence, URDU was not used here.

In this experiment, sentences were either used for training or testing. For each of the datasets shown in the table, the proposed VGGE model, along with AlexNet and ResNet50, was trained using MFCC features. Each model was trained five times using an 80%/20% train/test split, and the average results were computed.

The results are in Table 6. The trends are similar to those of Experiment 1.1. This time, VGGE is the best on ASED and RAVDESS, while ResNet50 is the best on EMO-DB. Concerning the averages, ASED and RAVDESS are fairly close (84.46%, 81.11%), while EMO-DB is lower (66.67%). So, this again suggests that Amharic and English monolingual SER are of similar difficulty and easier, within the context of these particular datasets and this task, while German SER is more difficult.

Table 6. Experiment 1.2: Monolingual SER results, expressed as accuracy, for the different datasets. Training and testing sentences are varied in this experiment.

Model	ASED	EMO-DB	RAVDESS
AlexNet	80.93	55.74	82.22
VGGE	86.63	70.49	83.33
ResNet50	85.82	73.77	77.78
Average	84.46	66.67	81.11

7. Experiment 2: Comparison of SER Methods for Amharic Cross-Lingual SER

The aim was to compare the three models AlexNet, VGGE, and ResNet50 (Section 4) when applied to cross-lingual SER. This time, the systems are trained on data in one

language and then tested on data in another language. Firstly, the three models are trained on ASED and tested on EMO-DB, then trained on EMO-DB and tested on ASED, and so on, for different combinations. To allow this cross-training, dataset-specific emotion labels are mapped into two classes, positive and negative, using the same method as for Experiment 1.

Once again, MFCC features were used for all models. The network configuration for VGGE was the same as in the preceding Experiment (Figure 1). For the other models, the standard configuration and settings were used.

The results are presented in Table 7. As line 1 of the table shows, we first trained on ASED and evaluated on EMO-DB (henceforth written ASED→EMO-DB). VGGE gave the best accuracy (66.67%), followed closely by AlexNet (65.80%) and then ResNet50 (64.06%). For EMO-DB→ASED, VGGE was best (64.22%), also followed by AlexNet (62.39%) and then ResNet50 (58.72%).

Table 7. Experiment 2: Cross-lingual SER results (train in one language, test in another language). The languages are Amharic (ASED), German (EMO-DB), English (RAVDESS), and Urdu (URDU).

Model	Training	Testing	Accuracy	F1-Score
AlexNet	ASED	EMO-DB	65.80	56.85
	EMO-DB	ASED	62.39	58.53
	ASED	RAVDESS	66.00	53.17
	RAVDESS	ASED	65.87	55.57
	ASED	URDU	60.00	56.28
	URDU	ASED	50.67	48.45
Average			61.79%	54.81%
VGGE	ASED	EMO-DB	66.67	52.55
	EMO-DB	ASED	64.22	58.53
	ASED	RAVDESS	59.25	51.85
	RAVDESS	ASED	61.43	62.75
	ASED	URDU	59.69	56.34
	URDU	ASED	60.00	53.94
Average			61.88%	55.99%
ResNet50	ASED	EMO-DB	64.06	50.42
	EMO-DB	ASED	58.72	45.94
	ASED	RAVDESS	61.75	48.68
	RAVDESS	ASED	64.16	52.66
	ASED	URDU	61.56	62.06
	URDU	ASED	61.33	60.03
Average			61.93%	53.30%

Next, for ASED→RAVDESS, AlexNet was best (66.00%), followed by ResNet50 (61.75%) and VGGE (59.25%). For RAVDESS→ASED, AlexNet was best (65.87%), closely followed by ResNet50 (64.16%) and then VGGE (61.43%).

Thirdly, we used ASED→URDU. Here, ResNet50 was best (61.56%), followed by AlexNet (60.00%) and VGGE (59.69%). For URDU→ASED, ResNet50 was best (61.33%), followed by VGGE (60.00%) and AlexNet (50.67%).

It is interesting that for ASED↔EMO-DB, VGGE was best; for ASED↔RAVDESS, AlexNet was best; and for ASED↔URDU, ResNet50 was best. What is more, the figures for AlexNet on ASED↔RAVDESS in the two directions (66.00%, 65.87%, difference 0.13%) were very close, as were those for ResNet50 on ASED↔URDU (61.56%, 61.33%, difference 0.23%), while those for VGGE on ASED↔EMO-DB (66.67%, 64.22%, difference 2.45%) were slightly further apart.

We can therefore conclude that the performance of the three models was very similar overall. This is supported by the average accuracy figures for AlexNet, VGGE, and

ResNet50 (61.79%, 61.88%, 61.93%, respectively) which are also very close (only 0.14% from the smallest to the biggest).

The results in the table also show that the average F1-score performance for VGGE (55.99%) is higher than that for AlexNet (54.81%, 1.18% lower) and ResNet50 (53.30%, 2.69% lower). Hence, it is concluded from these results that the prediction performance of VGGE was best, closely followed by AlexNet and then ResNet50. However, the range of F1-scores is small, only 2.69% from the smallest to the biggest, indicating only a slight difference in performance between different scenarios.

Regarding the results as a whole, two points can be made. First, the accuracy obtained by training on one language and testing on another is surprisingly good. Second, the best language to train on when testing on Amharic seems to vary by model. For AlexNet, it is RAVDESS (65.87%); for VGGE, it is EMO-DB (64.22%); and for ResNet50, it is RAVDESS again (64.16%).

Finally, we can compare our results for this experiment (Table 7) with those given for previous cross-lingual studies in Section 2. Generally, they seem comparable. Our average results are around 62%. In the previous studies, we see 56.8% [9], 57.87% [17], 65.3% [19], and 62.5% [22]. The highest is 71.62% [14]. In looking at these figures, we must remember that the exact methods and evaluation criteria used in previous experiments vary, so exact comparisons are not possible. Many different languages and datasets are used, emotion labels may need to be combined or transformed in different ways, and so on. Please refer to Section 2 for the details regarding these figures.

8. Experiment 3: Multilingual SER

In the previous experiment, we trained in one language and tested in another. In this final experiment, we trained on several non-Amharic languages and then tested on Amharic.

The same three models were used, AlexNet, VGGE, and ResNet50, with the same settings and training regime as in the previous experiments.

Table 8 shows the results. Recall that the languages are Amharic (ASED), German (EMO-DB), English (RAVDESS), and Urdu (URDU). The first three rows for each model show the results when two datasets were used for training, EMO-DB+RAVDESS, EMO-DB+URDU, and RAVDESS+URDU. The fourth row uses all three datasets for training, i.e., EMO-DB+RAVDESS+URDU. In all cases, testing is with ASED.

The best overall performance in the table is for VGGE, training with EMO-DB+URDU (69.94%). The average figure for VGGE over all the dataset training combinations is also the best (66.44%).

When RAVDESS is added to EMO-DB+URDU to make EMO-DB+RAVDESS+URDU, the performance of VGGE falls by 1.53% to 68.41%. In the results presented in Table 9, the upper right-hand column shows the average accuracy, and the lower right-hand column the average F1-score. In this case, we see that the highest figures over all three models are for all three datasets (67.12% and 59.79%, respectively).

However, the most interesting result here is that the best accuracy figure in Table 8 (EMO-DB+URDU→ASED, VGGE, 69.94%) is higher than the best accuracy figure in Table 7 with ASED as the target, (RAVDESS→ASED, AlexNet, 65.87%) by 4.07%. In other words, training on German and Urdu gives a better result for Amharic than training on English alone. Moreover, the best overall average accuracy figure in Table 8 (VGGE, 66.44%) is higher than the best overall average accuracy figure in Table 7 (ResNet50, 61.93%) by 4.51%.

Once again, the results in Table 8 show that the average F1-score performance for VGGE (62.78%) is higher than that for AlexNet (55.81%, 6.97% lower) and ResNet50 (51.65%, 11.13% lower). Furthermore, the best overall average F1-score figure in Table 8 (VGGE, 62.78%) is higher than the best overall average F1-score figure in Table 7 (VGGE, 55.99%) by 6.79%.

Table 8. Experiment 3: Multilingual SER results (train in two or three non-Amharic languages, test in Amharic). The languages are Amharic (ASED), German (EMO-DB), English (RAVDESS), and Urdu (URDU).

Model	Training	Testing	Accuracy	F1-Score
AlexNet	EMO-DB + RAVDESS	ASED	69.06	61.28
	EMO-DB + URDU	ASED	57.23	48.38
	RAVDESS + URDU	ASED	62.46	51.27
	EMO-DB + RAVDESS + URDU	ASED	69.77	62.30
Average			64.63%	55.81%
VGGE	EMO-DB + RAVDESS	ASED	60.50	61.12
	EMO-DB + URDU	ASED	69.94	65.26
	RAVDESS + URDU	ASED	66.89	64.56
	EMO-DB + RAVDESS + URDU	ASED	68.41	60.17
Average			66.44%	62.78%
ResNet50	EMO-DB + RAVDESS	ASED	61.33	43.52
	EMO-DB + URDU	ASED	46.24	44.57
	RAVDESS + URDU	ASED	64.51	56.17
	EMO-DB + RAVDESS + URDU	ASED	63.18	62.32
Average			58.82%	51.65%

Table 9. Experiment 3: Multilingual SER average results. The languages are Amharic (ASED), German (EMO-DB), English (RAVDESS), and Urdu (URDU).

Training	Testing	AlexNet	VGGE	ResNet50	Average Accuracy
EMO-DB + RAVDESS	ASED	69.06	60.5	61.33	63.63
EMO-DB + URDU	ASED	57.23	69.94	46.24	57.80
RAVDESS + URDU	ASED	62.46	66.89	64.51	64.62
EMO-DB + RAVDESS + URDU	ASED	69.77	68.41	63.18	67.12
Training	Testing	AlexNet	VGGE	ResNet50	Average F1-Score
EMO-DB + RAVDESS	ASED	61.00	61.21	43.44	55.31
EMO-DB + URDU	ASED	48.57	65.98	38.96	52.74
RAVDESS + URDU	ASED	51.64	64.63	56.23	57.33
EMO-DB + RAVDESS + URDU	ASED	62.45	60.28	62.99	59.79

These results suggest that by using several non-Amharic datasets for training, we can obtain a better result, by several percentage points, than when using one non-Amharic dataset for training, when testing on Amharic throughout.

Compared with the existing studies discussed in Section 2, there are only three that present multilingual experiments. Leter et al. [13] report that training on three datasets, EMO-DB, DES, and ENT, and testing on EMO-DB gave the best result, better than their

cross-lingual trials. This concurs with our own findings, where the average results for Experiment 3 (Table 8, bottom line) were higher than those of Experiment 2 (Table 7, bottom line). Latif et al. [17] found that training on EMO-DB, EMOVO, and SAVEE and testing on URDU gained a better result than using just two training datasets. Latif et al. [19] also obtained the best result when training on three datasets.

9. Conclusions

In this paper, we first proposed a variant of the well-known VGG model, which we call VGGE, and then applied AlexNet, VGGE, and ResNet50 to the task of speech emotion recognition, focusing on the Amharic language. This was made possible by the existence of the publicly available Amharic Speech Emotion Dataset (ASED), which we created in our previous work [10]. In Experiment 1, we trained the three models on four datasets: ASED (Amharic), RAVDESS (English), EMO-DB (German), and URDU (Urdu). In each case, a model was trained on one dataset and then tested on that same dataset. Speaker-independent and sentence-independent training variants were tried. The results suggested that Amharic and English monolingual SER are almost equally difficult on the datasets we used for these languages, while German is harder, and Urdu is easier.

In Experiment 2, we trained on SER data in one language and tested on data in another language, for various language pairs. When ASED was the target, the best dataset to train on was RAVDESS for AlexNet and ResNet50, and EMO-DB for VGGE. This could indicate that, in terms of SER, Amharic is more similar to English and German than it is to Urdu.

In Experiment 3, we combined datasets for two or three different non-Amharic languages for training and used the Amharic dataset for testing. The best result in Experiment 3 (EMO-DB+URDU→ASED, VGGE, 69.94%) was 4.07% higher than the best result in Experiment 2 (RAVDESS→ASED, AlexNet, 65.87%). In addition, the best overall average figure in Experiment 3 (VGGE, 66.44%) was 4.51% higher than the best overall average figure in Experiment 2 (ResNet50, 61.93%). These findings suggest that if several non-Amharic datasets are used for SER training, the results can be better than if one non-Amharic dataset is used, when testing on Amharic throughout. Overall, the experiments demonstrate how cross-lingual and multilingual approaches can be used to create effective SER systems for languages with little or no training data, confirming the findings of previous studies. Future work could involve improving SER performance when training on non-target languages and trying to predict which combination of source languages will give the best result.

Author Contributions: Conceptualization, E.A.R. and R.S.; methodology, E.A.R.; software, E.A.R. and E.A.; validation, E.A.R., E.A. and M.M.; formal analysis, E.A.R.; investigation, E.A.R., R.S., S.A.K., M.M. and E.A.; resources, E.A.R.; data curation, E.A.R.; writing—original draft preparation, E.A.R., J.M., M.A.B. and S.A.C.; writing—review and editing, R.S., S.A.K. and J.M.; visualization, E.A.R. and E.A.; supervision, J.F. and R.S.; project administration, E.A.R. and R.S.; funding acquisition, J.F. and R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under grant 2020YFC1523302.

Data Availability Statement: Publicly available datasets were analyzed in this study. Our Amharic Speech Emotion Dataset (ASED) is also publicly available at https://github.com/Ethio2021/ASED_V1.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zvarevashe, K.; Olugbara, O. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms* **2020**, *13*, 70. [CrossRef]
2. Khan, M.U.; Javed, A.R.; Ihsan, M.; Tariq, U. A novel category detection of social media reviews in the restaurant industry. *Multimed. Syst.* **2020**, *29*, 1–14. [CrossRef]
3. Zhang, B.; Provost, E.M.; Essl, G. Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5805–5809.

4. Zhang, Z.; Weninger, F.; Wöllmer, M.; Schuller, B. Unsupervised learning in cross-corpus acoustic emotion recognition. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 523–528.
5. Wang, D.; Zheng, T.F. Transfer learning for speech and language processing. In Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, China, 16–19 December 2015; pp. 1225–1237.
6. Wöllmer, M.; Stuhlsatz, A.; Wendemuth, A.; Rigoll, G. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affect. Comput.* **2010**, *1*, 119–131.
7. Mossie, Z.; Wang, J.H. Social network hate speech detection for Amharic language. *Comput. Sci. Inf. Technol.* **2018**, 41–55.
8. Mengistu, A.D.; Bedane, M.A. Text Independent Amharic Language Dialect Recognition using Neuro-Fuzzy Gaussian Membership Function. *Int. J. Adv. Stud. Comput. Sci. Eng.* **2017**, *6*, 30.
9. Albornoz, E.M.; Milone, D.H. Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. *IEEE Trans. Affect. Comput.* **2015**, *8*, 43–53. [[CrossRef](#)]
10. Retta, E.A.; Almekhlafi, E.; Sutcliffe, R.; Mhamed, M.; Ali, H.; Feng, J. A new Amharic speech emotion dataset and classification benchmark. *ACM Trans. Asian -Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–22. [[CrossRef](#)]
11. Sailunaz, K.; Dhaliwal, M.; Rokne, J.; Alhaji, R. Emotion detection from text and speech: A survey. *Soc. Netw. Anal. Min.* **2018**, *8*, 1–26. [[CrossRef](#)]
12. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **2011**, *53*, 1062–1087. [[CrossRef](#)]
13. Lefter, I.; Rothkrantz, L.J.; Wiggers, P.; Van Leeuwen, D.A. Emotion recognition from speech by combining databases and fusion of classifiers. In *International Conference on Text, Speech and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 353–360.
14. Xiao, Z.; Wu, D.; Zhang, X.; Tao, Z. Speech emotion recognition cross language families: Mandarin vs. western languages. In Proceedings of the 2016 International Conference on Progress in Informatics and Computing (PIC), Shanghai, China, 23–25 December 2016; pp. 253–257.
15. Sagha, H.; Matejka, P.; Gavryukova, M.; Povolný, F.; Marchi, E.; Schuller, B.W. Enhancing Multilingual Recognition of Emotion in Speech by Language Identification. *Interspeech* **2016**, 2949–2953.
16. Meftah, A.; Seddiq, Y.; Alotaibi, Y.; Selouani, S.A. Cross-corpus Arabic and English emotion recognition. In Proceedings of the 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Bilbao, Spain, 18–20 December 2017; pp. 377–381.
17. Latif, S.; Qayyum, A.; Usman, M.; Qadir, J. Cross lingual speech emotion recognition: Urdu vs. western languages. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018; pp. 88–93.
18. Latif, S.; Rana, R.; Younis, S.; Qadir, J.; Epps, J. Cross corpus speech emotion classification-an effective transfer learning technique. *arXiv* **2018**, arXiv:1801.06353.
19. Latif, S.; Qadir, J.; Bilal, M. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 732–737.
20. Goel, S.; Beigi, H. Cross lingual cross corpus speech emotion recognition. *arXiv* **2020**, arXiv:2003.07996.
21. Bhaykar, M.; Yadav, J.; Rao, K.S. Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM. In Proceedings of the 2013 National conference on communications (NCC), New Delhi, India, 15–17 February 2013; pp. 1–5.
22. Zehra, W.; Javed, A.R.; Jalil, Z.; Khan, H.U.; Gadekallu, T.R. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.* **2021**, *7*, 1845–1854. [[CrossRef](#)]
23. Duret, J.; Parcollet, T.; Estève, Y. Learning Multilingual Expressive Speech Representation for Prosody Prediction without Parallel Data. *arXiv* **2023**, arXiv:2306.17199.
24. Pandey, S.K.; Shekhawat, H.S.; Prasanna, S.R.M. Multi-cultural speech emotion recognition using language and speaker cues. *Biomed. Signal Process. Control* **2023**, *83*, 104679. [[CrossRef](#)]
25. Deng, J.; Zhang, Z.; Marchi, E.; Schuller, B. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In Proceedings of the 2013 humane association conference on affective computing and intelligent interaction, Geneva, Switzerland, 2–5 September 2013; pp. 511–516.
26. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [[CrossRef](#)]
27. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVD ESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [[CrossRef](#)]
28. Stanislavski, C. An Actor Prepares (New York). *Theatre Art*. **1936**, 38.
29. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.

30. Gangamohan, P.; Kadiri, S.R.; Yegnanarayana, B. Analysis of emotional speech—A review. In *Toward Robotic Socially Believable Behaving Systems-Volume I*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 205–238.
31. Fairbanks, G.; Hoaglin, L.W. An experimental study of the durational characteristics of the voice during the expression of emotion. *Commun. Monogr.* **1941**, *8*, 85–90. [CrossRef]
32. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
33. Dey, N.A.; Amira, S.M.; Waleed, S.N.; Nhu, G. Acoustic sensors in biomedical applications. In *Acoustic Sensors for Biomedical Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–47.
34. Almekhlafi, E.; Moeen, A.; Zhang, E.; Wang, J.; Peng, J. A classification benchmark for Arabic alphabet phonemes with diacritics in deep neural networks. *Comput. Speech Lang.* **2022**, *71*, 101274. [CrossRef]
35. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]
36. Shaw, A.; Vardhan, R.H.; Saxena, S. Emotion recognition and classification in speech using Artificial neural networks. *Int. J. Comput. Appl.* **2016**, *145*, 5–9. [CrossRef]
37. Mustaqeem; Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2020**, *20*, 183.
38. Kumbhar, H.S.; Bhandari, S.U. Speech Emotion Recognition using MFCC features and LSTM network. In Proceedings of the 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 19–21 September 2019; pp. 1–3.
39. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
40. Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; Kautz, J. Pruning convolutional neural networks for resource efficient inference. *arXiv* **2016**, arXiv:1611.06440.
41. George, D.; Shen, H.; Huerta, E.A. Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO. *arXiv* **2017**, arXiv:1706.07446.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
44. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
45. Sharmin, R.; Rahut, S.K.; Huq, M.R. Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network. *Procedia Comput. Sci.* **2020**, *171*, 1381–1388. [CrossRef]
46. Shinde, A.S.; Patil, V.V. Speech Emotion Recognition System: A Review. SSRN 3869462. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3869462 (accessed on 10 October 2023).
47. Deb, S.; Dandapat, S. Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. *IEEE Trans. Cybern.* **2018**, *49*, 802–815. [CrossRef] [PubMed]
48. Wang, K.; Su, G.; Liu, L.; Wang, S. Wavelet packet analysis for speaker-independent emotion recognition. *Neurocomputing* **2020**, *398*, 257–264. [CrossRef]
49. Swain, M.; Sahoo, S.; Routray, A.; Kabisatpathy, P.; Kundu, J.N. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. *Int. J. Speech Technol.* **2015**, *18*, 387–393. [CrossRef]
50. Kuchibhotla, S.; Vankayalapati, H.D.; Anne, K.R. An optimal two stage feature selection for speech emotion recognition using acoustic features. *Int. J. Speech Technol.* **2016**, *19*, 657–667. [CrossRef]

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.