A New Amharic Speech Emotion Dataset and Classification Benchmark

A New Amharic Speech Emotion Dataset and Classification Benchmark

EPHREM AFELE RETTA, School of Information Science and Technology, Northwest University, China
EIAD ALMEKHLAFI, School of Information Science and Technology, Northwest University, China
RICHARD SUTCLIFFE*, School of Information Science and Technology, Northwest University, China and School of Computer Science and Electronic Engineering, University of Essex, UK
MUSTAFA MHAMED, School of Information Science and Technology, Northwest University, China
HAIDER ALI, School of Information Science and Technology, Northwest University, China
JUN FENG†, School of Information Science and Technology, Northwest University, China

In this paper we present the Amharic Speech Emotion Dataset (ASED), which covers four dialects (Gojjam, Wollo, Shewa and Gonder) and five different emotions (neutral, fearful, happy, sad and angry). We believe it is the first Speech Emotion Recognition (SER) dataset for the Amharic language. 65 volunteer participants, all native speakers of Amharic, recorded 2,474 sound samples, two to four seconds in length. Eight judges (two for each dialect) assigned emotions to the samples with high agreement level (Fleiss kappa = 0.8). The resulting dataset is freely available for download. Next, we developed a four-layer variant of the well-known VGG model which we call VGGb. Three experiments were then carried out using VGGb for SER, using ASED. First, we investigated which features work best for Amharic, FilterBank, Mel Spectrogram, or Mel-frequency Cepstral Coefficient (MFCC). This was done by training three VGGb SER models on ASED, using FilterBank, Mel Spectrogram and MFCC features respectively. Four forms of training were tried, standard cross-validation, and three variants based on sentences, dialects and speaker groups. Thus, a sentence used for training would not be used for testing, and the same for a dialect and speaker group. MFCC features were superior under all four training schemes. MFCC was therefore adopted for Experiment 2, where VGGb and three well-known existing models were compared on ASED: RESNet50, AlexNet and LSTM. VGGb was found to have very good accuracy (90.73%) as well as the fastest training time. In Experiment 3, the performance of VGGb was compared when trained on two existing SER datasets – RAVDESS (English) and EMO-DB (German) – as well as on ASED (Amharic). Results are comparable across these languages, with ASED being the highest. This suggests that VGGb can be successfully applied to other languages. We hope that ASED will encourage researchers to explore the Amharic language and to experiment with other models for Amharic SER.

*Corresponding author.
†Corresponding author.

Authors' addresses: Ephrem Afele Retta, School of Information Science and Technology, Northwest University, P.O. Box 1212, Xi'an, China, afele@stumail.nwu.edu.cn; Eiad Almekhlafi, School of Information Science and Technology, Northwest University, Changan, Xi'an, China, e_almekhlafi@stumail.nwu.edu.cn; Richard Sutcliffe, School of Information Science and Technology, Northwest University, Xi'an, China and School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, UK, rsutcl@nwu.edu.cn, rsutcl@essex.ac.uk; Mustafa Mhamed, School of Information Science and Technology, Northwest University, Xi'an, China, mustafamhamed@stumail.nwu.edu.cn; Haider Ali, School of Information Science and Technology, Northwest University, Xi'an, China, alihaider@stumail.nwu.edu.cn; Jun Feng, School of Information Science and Technology, Northwest University, Xi'an, China, 710127, fengjun@nwu.edu.cn.

CCS Concepts: • **Applied computing → Sound and music computing**; • **Computing methodologies → Supervised learning by classification**.

Additional Key Words and Phrases: Speech emotion recognition; Amharic dataset; Classifiers; Feature extraction.

1 INTRODUCTION

Emotion plays a significant role in everyday human interactions [22] and conveys considerable information about the mental state of an individual. This has opened up a new research field called automatic emotion recognition. Various approaches have been explored in prior studies to recognize emotional states using facial expressions, speech, physiological signals, etc. [22]. Several inherent advantages make speech signals a good source for affective computing. For example, compared to many other biological signals (e.g., electrocardiograms), speech signals can usually be acquired more readily and economically. This is why many researchers are interested in Speech Emotion Recognition (SER).

SER is an important research area that has been active for more than two decades [41]. The results of SER can already be seen in many application fields, including entertainment, computer games, audio monitoring, online learning, clinical research, polygraph tests, and call centers [4, 22, 28]. Even though SER has many benefits, it is still a difficult task to perform with high accuracy [27]. One key problem is choosing the right features; an incorrect choice can lead to moderate performance [16].

Audio spectrum features are usually divided into two domains; time-domain features and frequency-domain features. The time-domain functions are elementary to extract and allow easy analysis of audio signals. In the case of small audio datasets, the frequency domain features will show deeper patterns, which may help distinguish the signal's basic emotion. Frequency-domain features include spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectral roll-off, spectral entropy, and Chroma coefficients [49]. In this paper, a comprehensive analysis of each feature was performed during the exploratory data analysis. However, for the purpose of this work, we limited ourselves to three principal features, FilterBank, Mel Spectrograms and MFCC.

In recent years, numerous speech datasets have been created to support deep learning for SER. The languages of these datasets include English, Chinese, Spanish, French, German and many others [23]. However, no SER dataset has been created for Amharic yet. Amharic is the second-largest Semitic language in the world after Arabic and the national language of Ethiopia [34]. 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 [32]. However, despite this, Amharic and its dialects have very few language resources compared to other languages such as English. It is for this reason that we have created a new SER dataset for Amharic.

Unlike English, Amharic is a syllabic language; each character represents one syllable [5]. The language uses a script derived from the Ge'ez alphabet [5]. It has 33 main characters, and each consonant-vowel combination has seven forms. Compared with English, Amharic has some unique phonemes. Gemination in Amharic is one of the most distinctive features of the speech's cadence, and it has great semantic and syntactic functional weight. In contrast to English, where the rhythm is mainly characterized by stress (loudness), the rhythm of Amharic is mainly characterized by longer and shorter syllables depending on the germination of consonants and certain characteristics of the phrase [2]. Amharic gemination is either lexical or morphological. In the speech synthesis field, it is vital to introduce emotional features to give greater expression to a machine, so that it speaks more like a human. Additionally, expressions and intonation for emotions and mental states differ markedly in their nature and significance from language to language.

In summary, therefore, the two main challenges for Amharic are the limited availability of datasets and the language's complex morphological characteristics [35]. This paper addresses both points by introducing a new Amharic dataset suitable for training and testing SER systems, and by presenting experiments which show how frequency-domain features can be effectively used for Amharic SER.

In addition, we have developed a SER model called VGGb, based on the well-known VGG SER architecture, and used it to carry out three experiments. The first experiment was to choose an appropriate method from the FilterBank, Mel Spectrogram and MFCC technologies for extracting features from recordings in our ASED dataset, using VGGb. Four train-test schemes were used. The first was standard cross-validation, withholding 10% of training utterances for testing. The second trained on just five of the seven sentences in ASED, testing on the remaining two. The third trained only with utterances spoken in three of the four dialects in ASED, testing on utterances in the remaining dialect. The fourth trained using utterances by participants in two of the three speaker groups in ASED, testing on utterances in the remaining speaker group. Under all four schemes, MFCC was found to be the most effective in terms of accuracy and training time.

The second experiment was to compare the classification performance of VGGb and three other popular models using MFCC features. VGGb achieved a high accuracy (90.73%) as well as having the shortest training time. In the third experiment, we evaluated VGGb on the English RAVDESS and German EMO-DB datasets. The results for English and German were comparable to those achieved for Amharic on the ASED dataset.

The contributions of this paper are as follows:

- We create for the very first time a SER dataset for Amharic, called ASED. There are 65 speakers and 2,474 recordings, 522 neutral, 510 fearful, 486 happy, 470 sad, and 486 angry.
- All four Amharic dialects (Gojjam, Wollo, Shewa and Gonder) are included in the dataset.
- Eight judges evaluate the recordings, and agreement between them is high (Fleiss kappa = 0.8). So the data is of high quality.
- We analyze ASED with respect to nine other well-known SER datasets and show that it compares very well in terms of the number of participants, the amount of data produced, the method of quality control, the number of judges, and their agreement level.
- We develop a high-performing variant of the VGG SER model which has just four CNN layers. We call this VGGb.
- Using VGGb and ASED data, we compare FilterBank, Mel Spectrogram and MFCC features and show experimentally in a SER task that MFCC leads to higher accuracy. We show that the superiority of MFCC is independent of training sentence, Amharic dialect and speaker group.
- We apply VGGb and three other architectural models to the SER task, working with ASED data, and show that VGGb is very effective, and by far the fastest.
- Using VGGb and working with MFCC, we train models to recognise five emotions in Amharic, English and German, using the ASED, RAVDESS, and EMO-DB datasets respectively. VGGb shows excellent performance.

The rest of this paper is organized as follows: Section 2 presents the ASED dataset, describing the rationale behind its design and the method by which it was created. A detailed comparison with nine other datasets is also included (four for English, one each for Chinese, German, Greek, Gujarati, Hindi). Section 3 discusses feature extraction for speech emotion recognition, and gives a short overview of FilterBank, Mel Spectrograms and MFCC. Section 4 describes the VGGb architecture and settings used for our experiments. Sections 5-7 describe the experiments. Section 8 gives conclusions and next steps.

2 ASED DATASET FOR AMHARIC

2.1 Existing SER Datasets

Before presenting our proposed dataset, we briefly describe two well-known existing datasets which are used in our experiments. Table 1 shows the main data.The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [30] contains audio and video recordings of English sentences spoken by twelve males and twelve females in eight emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. The total number of

Table 1. Comparison between RAVDESS, Emo-DB and ASED Speech Emotion Recognition datasets.

Aspect	RAVDESS	Emo-DB	ASED
Language	English	German	Amharic
Number of recordings	1,440	535	2,474
Number of sentences	2	10	27
Number of participants	24	10	65
Number of emotions	8	7	5

Table 2. Examples of Amharic sentences for each of the five emotions in the ASED dataset.

No.	Emotions	Sentences
1	Neutral	ነገ ስብሰባ አለኝ (I have a meeting tomorrow)
2	Fearful	ኧረ በሩ ተንኳኳ ሌባ ነው መሰለኝ (I think he was a thief who knocked on the door)
3	Happy	የአበበ ሰርግ በጣም ደስ ይላል (Abebe's wedding is very nice)
4	Sad	የአክስቱ ልጅ በመኪና አደጋ ሞተች (My cousin died in a car accident)
5	Angry	ዞር በል ጥፋልኝ ከአሁን በኋላ አንዳላይህ (Turn away and let me not see you again)

speech file utterances is 1,440. Recordings are three seconds in length at a sampling rate of 48 kHz. The Berlin Emo-DB dataset [6] contains audio recordings of German sentences made by five males and five females in seven different emotions, neutral, fear, anger, happiness, sadness, disgust, and boredom. The speech material comprises about 535 sentences. Recordings are two to three seconds in length at a sampling rate of 16 kHz.

2.2 Design of ASED

Dialects: Amharic can be considered a challenging language for SER because of its huge lexical variety and complicated morphology [2]. There are four main Amharic dialects, namely Gojjam (Gojjamegna), Wollo (Wollogna), Shewa (Shewagna), and Gonder (Gonderegna) [32]. We wished the proposed dataset to contain examples of all four dialects.

Emotions: RAVDESS contains eight emotions (see above) while Emo-DB has seven. It was decided to adopt five emotions which are common to the other datasets. So, relative to RAVDESS, calm, surprise and disgust are omitted, while relative to Emo-DB, disgust and boredom are omitted. The use of a common subset allows direct SER comparisons to be made across languages.

Test sentences: For each of the five emotions in ASED, five sentences expressing that emotion were composed in Amharic. For example, for Happy we have 'Abebe's wedding is very nice' (Table 2), a sentence which expresses a happy concept. Similarly, 'My cousin died in a car accident' is a Sad sentence. Furthermore, the dataset contains two common sentences which express no strong emotion, e.g. 'My sister is coming by plane'. The total number of sentences in the dataset is thus 27, 5×5 emotion-specific sentences plus two common sentences.

Approach to generating emotion: Three methods can be used to collect the recordings in SER datasets: Simulated, Induced, and Natural [12, 23]. For simulated emotions, participants are asked to read a sentence while expressing a stated emotion. So if the sentence was 'My sister is coming by plane', one participant could be asked to read it in an angry way, while another read it in a sad way. In the Induced approach, the participant is made to feel the required emotion before reading the sentence. In the Natural approach, a recording must be made when a speaker happens to be feeling a particular emotion. The Simulated approach is the most practical to implement, and nearly 60% of speech datasets are collected using this method [23]. For this reason, the Simulated approach was also adopted for ASED data collection.

Table 3. Classification of the 65 participants who made the ASED dataset recordings.

Category	Values
Occupation	Postgraduate (21), Undergraduate (21), Business person (23)
Expertise	Professional (20), Semi-professional (26), Amateur (19)

2.3 Creation of ASED

Judges:

All judges were Ethiopian postgraduate students of Computer Science or Management at universities in Xi'an, China. All were native speakers of one Amharic dialect. Two judges were from Xidian University (西安电子科技大学) and spoke Shewa and Gonder dialects respectively; two were from Chang'an University (长安大学), (Wollo, Gojjam); one was from Xi'an Shiyou University (西安石油大学), (Shewa); two were from Xi'an Jiaotong University (西安交通大学), (Gonder, Gojjam); finally, one was from Northwest University (西北大学), (Wollo). Judges were responsible for the quality control of the dataset (see below).

Participants: There were three classes of participant, undergraduate students, postgraduate students and business people. The undergraduate and postgraduate students came from Ethiopia to China in order to study. The business people came to China for professional reasons concerning their work. In order to take part, participants had to be native speakers of one of the four Amharic dialects, and they had to be capable of speaking in different emotions, according to the opinions of the judges, following some initial tests. During the selection process, the judges assigned each participant to one of three groups, depending on their expertise in the task: Professional, Semi-professional and Amateur (Table 3). Participants who were clearly experienced at acting and were excellent at expressing the different emotions, in the opinion of the judges, were assigned to the Professional group. Those who were judged very good at expressing the emotions were assigned to the Semi-professional group. Finally, participants who were not experienced at acting but were nevertheless judged good enough to participate in the task were assigned to the Amateur group. In total there were 65 participants (25 female, 40 male), aged from 20 to 40 years.

Recording: In order to record the speech, we used six Huawei Nova 4 mobile phones, on which an Android-based speech recording software app [46] had been installed. Mobile phones were used because professional audio equipment was not available to us. The software was set up to capture the speech utterances utilizing a 16 kHz sampling rate at 16 bits, resulting in a mono .Wav file. The recording software displayed the text for them one sentence at a time and indicated the required emotion. They then recorded the sentence onto the phone. When they finished speaking, the recorded audio file was saved. They then recorded the same sentence with the same emotion a second time. The recording was done at the School of Information Science and Technology, Northwest University, in a quiet room in order to obtain speech signals with minimum noise. The distance between the speaker's mouth and the microphone of the mobile phone was 25 cm. We used the Audacity audio editing software [3] to reduce the background noise of the speech signal. Audacity is very reputable and is also open source.

Data creation: We adopted a semi-supervised recording approach: The participants were given a short description of the purpose of the study and the recording system, and were then allowed to choose a convenient time to conduct the recording. Each participant was first asked to record all 25 emotion-specific sentences. They were provided with the sentence and the required emotion, for example, 'Turn away and let me not see you again' to be spoken in an Angry way. Each recording was made twice. Next on the list could be the sentence

Table 4. Utterance counts in the ASED dataset over sentiment classes, broken down by gender.

Emotions	Gender		Total
	Male	Female	
Neutral	332	190	522
Fear	316	194	510
Happy	300	186	486
Sad	296	174	470
Angry	304	182	486
Totals	1,548	926	2,474

Table 5. Counts of utterances in the ASED dataset, broken down by emotion and Amharic dialect.

Id	Emotion	Amharic dialect				Number of recordings
		Gojjam	Wollo	Shewa	Gonder	
1	Neutral	104	208	140	70	522
2	Fearful	59	170	200	81	510
3	Happy	68	150	188	80	486
4	Sad	70	135	135	130	470
5	Angry	111	120	140	115	486
Totals		412	783	803	476	2,474

'I think he was a thief who knocked on the door', to be spoken in a Fearful way, and so on. For these emotion-specific sentences, the required emotion was always the same for a given sentence, for all participants. After the 25 sentences, the participant was then asked to record the two common sentences, for example 'my sister is coming on a plane'. Each common sentence was recorded twice for each of the five emotions, Neutral, Fearful, Happy, Sad and Angry. So they would first record the sentence in a Neutral way (twice), then record the same sentence in a Fearful way (twice) and so on. Participants always spoke in their own Amharic dialect. A complete set therefore comprised $25 \times 2 = 50$ recordings of emotion-specific sentences, and $2 \times 5 \times 2 = 20$ recordings of common sentences, 70 recordings in all. Table 4 shows the distribution of utterances over sentiment classes and also includes separate counts by gender.

Table 6. Numbers of recordings in the ASED dataset contributed by participants. Example: Twenty participants contributed between 43 and 56 recordings to the final dataset.

Number of recordings	Number of participants
1-14	19
15-28	9
29-42	0
43-56	20
57-70	17

Judgments: Every recording was independently reviewed by all eight judges. Concerning emotion-specific sentences, each judge decided whether the recording expressed the emotion adequately or not, and made a binary decision, Accept or Reject. For the common sentences, the judge did not know the intended emotion of

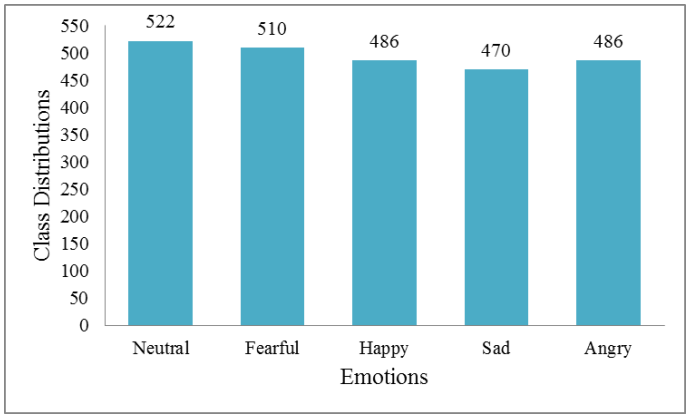


Fig. 1. Distribution of utterances in the ASED dataset across the five emotions.

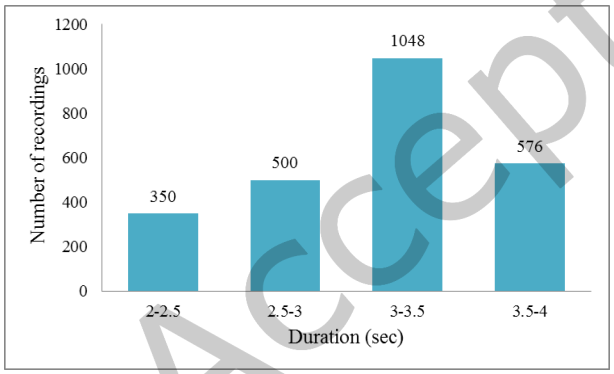


Fig. 2. Distribution of utterances in the ASED dataset, based on duration ranges. $2 - 2.5$ in the figure means $2s <= d < 2.5s$ where d is the utterance length, and the same for the other ranges.

the recording. They decided which emotion the recording expressed. If it was unclear, their decision was Reject, otherwise their decision was one of the five emotions.

For emotion-specific sentences, a recording was only accepted for inclusion in the ASED dataset if five or more judges returned the decision Accept. Similarly, a common sentence was only accepted if five or more judges assigned the same emotion to it.

Because there were 65 participants each of whom made 70 recordings, we would expect ASED to contain 4,550 recordings. However, because of the above selection process, many of these were rejected, resulting in 2,474 recordings in the dataset.

Inter-annotator agreement: Since we had eight judges, the Fleiss kappa [38] coefficient was used to calculate the pairing agreement between participants.

$$\kappa = \frac{\overline{p}_0 - \overline{p}_e}{1 - \overline{p}_e} \tag{1}$$

The factor $1 - \overline{p}_e$ gives the degree of agreement that is attainable above chance, and $\overline{p}_0 - \overline{p}_e$ gives the degree of agreement actually achieved above chance: $k = 1$, if all the raters are in complete agreement. Evaluation of the

Table 7. Some datasets for Speech Emotion Recognition (Sp+Utt = Speakers and Utterances, Sampl = Sampling Frequency, Quant = Quantization, Env = Environment, Mod = Modalities, Sim = Simulated, Nat = Natural, Ind = Induced, Lab = Laboratory Environment, F/TV = Film/TV, A = Audio, V = Visual). See Table 8 for the full names of the datasets.

Database	Lang.	Sp. (Utt.)	Sampl. (kHz)	Quant. (Bits)	Emotions	Type	Env.	Mod.
AESDD	Greek	5 (500)	44.1	16	anger, disgust fear, happiness sadness.	Sim	Lab	A
ASED	Amharic	65 (2,474)	16	16	neutral, fearful happy, sad angry.	Sim	Lab	A
CHEAVD	Mandarin	238 (2,600)	44.1	16	26 emotions: surprise, happy sad, angry	Nat	F/TV	A/V
EGSC	Gujarati	9 (1,296)	44.1	–	fearful, neutral sadness, surprise anger, disgust fear, happiness.	Sim	Lab	A
EMO-DB	German	10 (535)	16	16	anger, boredom disgust, fear sadness, neutral happiness.	Sim	Lab	A
IEMOCAP	English	10 (1,150)	48	16	neutral, happiness anger, sadness frustration.	Sim Ind	Lab	A/V
IITKGP-SEHSC	Hindi	10 (12,000)	16	16	anger, disgust fear, happy neutral, sad sarcastic, surprise.	Sim	Lab	A
RAVDESS	English	24 (4,320)	48	16	neutral, calm happiness, sadness anger, fear surprise, disgust.	Sim	Lab	A/V
SAVEE	English	4 (480)	44.1	16	anger, disgust fear, happiness sadness, surprise neutral, common.	Sim	Lab	A/V
TESS	English	2 (2,800)	24.4	16	anger, disgust neutral, fear happiness, sadness pleasant, surprise.	Sim	Lab	A

inter-rater agreement for our dataset in terms of Fleiss kappa is 0.8. This value shows a high agreement among our eight raters.

Table 8. Dataset names and abbreviations.

Name	Abbreviation
Amharic Speech Emotion Dataset	ASED
Acted Emotional Speech Dynamic Database	AESDD
Chinese Natural Emotional Audio–Visual Database	CHEAVD
Emotional Gujarati Speech Corpus	EGSC
Berlin Database of Emotional Speech	EMO-DB
Interactive Emotional Dyadic Motion Capture Database	IEMOCAP
Hindi Speech Corpus for Emotion Analysis	IITKGP-SEHSC
Ryerson Audio-Visual Database of Emotional Speech and Song	RAVDESS
Surrey Audio-Visual Expressed Emotion	SAVEE
Toronto Emotional Speech Set	TESS

Files and labeling: Each file was then labeled in the form a5-02-01-01-12.wav. The first part of the name indicates the emotional state (n1 = neutral, f2 = fearful, h3 = happy, s4 = sad, a5 = angry), the second part indicates the sentence number (01-07), the third part indicates the repetition (01 = 1st repetition, 02 = 2nd repetition), the fourth part indicates the gender (01 = female, 02 = male) and the fifth part indicates the anonymized participant ID (01 to 65).

The final dataset consists of 2,474 recordings, each between two and four seconds in length, 522 neutral, 510 fearful, 486 happy, 470 sad, and 486 angry. Recorded phrases were stored in five different folders, one for each emotion. Table 5 shows the final breakdown of utterances across emotions and dialects. The ASED dataset is evenly distributed across all emotion classes, as shown in Fig. 1. The contribution of recordings by participants is shown in Table 6. As we have described above, each participant was asked to make 70 recordings. However, some were unable to do this, and moreover, not all recordings were assessed as satisfactory by the judges. The distribution of utterance counts based on length ranges can be seen in Fig. 2. The mode is 3-3.5 seconds. 1,048 utterances (42% of the whole dataset) fall within this range. In the final stage, ASED was split into training and testing sets randomly. The training set contains 90% of the whole dataset. The test set contains the rest of the data. The ASED dataset is freely available for download[1].

2.4 Comparison of ASED to Other Datasets

Before presenting our experiments, we briefly compare ASED to nine other datasets (Table 7). There are four for English, and one each for Chinese, German, Greek, Gujarati, and Hindi. The full names of each are shown in Table 8. First, we describe the main parameters and creation methods for each dataset. Second, we compare them to ASED.

AESDD [50] is for Greek, uses five emotions and contains nineteen emotionally neutral utterances derived from theatrical plays plus one improvised utterance. Each of these twenty is spoken in all five emotions by each actor. There are five professional actors and 500 utterances in total. No quality control or annotation process is mentioned. Recording is done at 44.1 kHz and 16 bits. Recordings were made at the sound studio of the Laboratory of Electronic Media.

CHEAVD [29] is for Mandarin Chinese and uses six basic emotions plus an additional twenty emotions. There are 2,600 segments selected from 34 films, two tv series, two tv shows, one speech and one talk show. Each segment involves one speaker and there are 238 different speakers in all. Thus there are no recruited participants, and all emotions are naturally occurring. Annotation of emotions is carried out by four judges. Pairwise kappa

[1]https://github.com/Ethio2021/ASED_V1

coefficients range between 0.41 and 0.58. However, there are 26 emotions rather than the usual six, so this accounts for the low agreement level. Recordings are all from preexisting films etc and are at 44.1 kHz and 16 bits.

EGSC [48] is for Gujarati, uses six emotions and is based on 24 individual words. Each recording is for just one word, spoken with one of the emotions. There are nine speakers who are experts in drama, and about 1,296 utterances in total. Recording is done at 44.1 kHz using mobile phones. A quiet room was used.

EMO-DB [6] (as already discussed) is for German, uses five emotions and contains ten everyday sentences, five made of one phrase, five made of two phrases. There are ten speakers, nine qualified in acting, and about 800 raw utterances in total. All recordings were judged by twenty listeners. Around 300 of the 800 utterances had a recognition rate greater than 80%. 535 utterances were finally selected for the database. Recording is done at 16 kHz and 16 bits, and was carried out in the Anechoic chamber of the Technical Acoustics Department at the Technical University Berlin.

IEMOCAP [7] is for English, uses five emotions and contains three scripts selected from plays, plus improvised emotional dialogues. Seven speakers are professional actors, three are students. Recordings were judged by six evaluators using a majority voting system. There are 10,039 dialogue turns in the dataset (5,255 scripted turns and 4,784 spontaneous turns). Recording is done at 48 kHz and 16 bits, and took place in the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC).

ITKGP-SEHSC [24] is for Hindi, uses eight emotions and contains fifteen sentences. Ten professional artists each recorded every sentence with every emotion, all in one session. The dataset was prepared in ten sessions. The total number of utterances in the database is 12,000 (10 sessions x 15 sentences × 10 speakers × 8 emotions). Recording was done at 16 kHz and 16 bits, working in a quiet room. Recordings were judged by 25 postgraduate and research students of IIT Kharagpur.

RAVDESS [30] (as already discussed) is for English, uses eight emotions and contains just two sentences. The 24 speakers are professional actors. Interestingly, emotion in this dataset is 'self-induced' [45], rather than Acted. Moreover, there are two levels of each emotion. There are 4,320 utterances. Project investigators first selected the best two clips for each speaker and each emotion. The selected recordings were then judged by 247 naive judges. The average Fleiss Kappa inter-rater score was 0.57. Recording was at 48 kHz and 16 bits, and it was carried out in a professional recording studio at Ryerson University.

SAVEE [17] is for English, uses six emotions and contains three common sentences plus two emotion-specific sentences and ten generic sentences (different for each emotion and phonetically-balanced). There are four speakers (postgraduates and researchers, not professional actors) and there are 480 utterances in total. Recordings were judged by ten evaluators (all students) and the average classification accuracy was 66.5%. Recording was at 44.1 kHz and 16 bits, and took place in the 3D vision laboratory at University of Surrey.

TESS [9, 36] is for English, uses seven emotions and contains 200 individual words. These are spoken in a carrier sentence 'Say the word X'. There are two female speakers who are professional actors, and 2,800 utterances. Fifty-six judges identified the emotion in each recording with an average accuracy of 82% [9]. 'Pleasant' was the least well recognized, and 'Anger' the most. Recording was at 24.4 kHz and 16 bits, and took place in a sound attenuating booth at University of Toronto.

A number of interesting comparisons can be made between ASED and the datasets mentioned above. Firstly, concerning participants, we can see that mostly professional actors are used for the recordings. The only exceptions are SAVEE which uses members of the campus community, and CHEAVD which has no participants as such, but uses preexisting speech in tv shows etc. For ASED we have a mixture of postgraduates, undergraduates and business persons making the recordings. It is an interesting assumption among all the papers that actors are the best choice. However, the way in which emotions are portrayed in drama is surely entirely different from everyday life. Actors exaggerate and distort the facets of the emotion they portray. Usually they also do this in a way which has come to be expected within the particular entertainment genre, but which is not real. So we

would argue that alternatives should be explored if the ultimate goal is to train SER systems for practical tasks such as detecting an upset customer in a call center. The number of participants used for the datasets varies dramatically from just two (TESS) up to 231 (CHEAVD). Excluding CHEAVD, the average is 9.2, so the use of 65 speakers in ASED is well above average, covers all the main dialects in Ethiopia and can thus be considered a representative sample.

For emotions, mostly the standard five are chosen as the basis and ASED conforms to this norm. Expression of the emotions is generally of the Acted form, i.e. simply asking the participant to speak in, say, an angry way, and leaving it to them how to do it. However, RAVDESS uses self-inducement [31] whereby the actor conjurs up the emotion in themselves over a period of some minutes before speaking. IEMOCAP uses scripts selected from plays, where the emotion is developed in the dialogue over several turns. They also try improvised dialogues based around an idea, e.g. for 'sad' the idea is that a close friend has recently died and one person is comforting another. This can allow the emotion to develop more naturally. At the other extreme, TESS expresses the emotion in a single word, with no preparation. For ASED, we use the simulated approach (e.g. please speak in a sad way) and we rely on our quality control mechanism to ensure that the results are satisfactory.

The choice of sentences which participants must speak varies greatly across datasets. EGSC uses single words, while TESS uses 'Say the word X' where X is a single word. RAVDESS has just two neutral sentences, spoken with different emotions. EMO-DB has ten, ITKGP-SEHSC fifteen, AESDD twenty and SAVEE 120; by contrast, IEMOCAP is using drama scripts plus dialogues, so this is a more naturalistic approach. At the extreme, CHEAVD has 2,600 tv show scripts, so this is an extrapolation of the IEMOCAP idea. For ASED, we are using 25 emotion-specific sentences plus two neutral sentences which are common to all emotions, so this is closest to EMO-DB and AESDD.

Considering the amount of data produced, ASED (2,474 utterances) compares well with the other datasets; the minimum is SAVEE (480), the maximum ITKGP-SEHSC (12,000 dialogue turns), and the average is 3,680. So, for a first Amharic dataset, ASED provides enough data to perform basic training tasks, as we show later. Finally, we consider the judges and agreement level. The number of judges used for CHEAVD (4), IEMOCAP (6) and SAVEE (10) are comparable to ASED (8). EMO-DB (20), ITKGP-SEHSC (25), TESS (56) and RAVDESS (247) use greater numbers. Considering the sampling frequency, both EMO-DB and ITKGP-SEHSC use 16 kHz which is the same as ASED. The majority of datasets are recorded in a quiet laboratory environment, the approach adopted for ASED as well.

The eight ASED judges showed a high level of inter-rater agreement (Fleiss kappa 0.8 – see Section 2.3 above). RAVDESS report a kappa of 0.57 over their 247 judges, while CHEAVD report kappa 0.41-0.58 over four judges and 26 emotions. Obviously, these figures are not directly comparable, for example the CHEAVD task is much harder because there are many more emotions, however, the ASED agreement level seems very good.

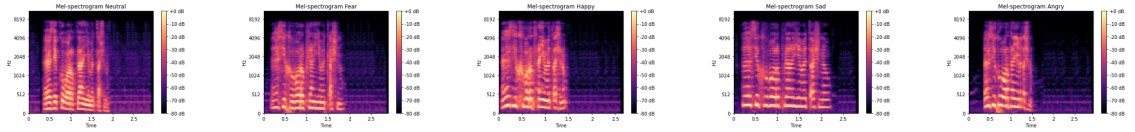
In conclusion, the proposed ASED SER dataset for Amharic appears to compare well with the others we mention here. In the following sections, we present some initial experiments where the ASED data is used for Amharic SER.

3 FEATURE EXTRACTION FOR SER

3.1 Background

The speech signal contains a large number of parameters reflecting emotional characteristics. One of the key issues within SER research is the choice of features which should be used.

Gangamohan et al. investigate how emotional speech differs from normal speech in terms of excitation source characteristics [11]. Excitation can be computed using linear prediction analysis and zero frequency filtering, performed at the sub-segmental level (1-3 ms) and this was shown to contain information relating to emotion. In further work, a 2D feature space, comprising spectral band magnitude energy ratio (β) vs. strength of excitation



(a) Neutral (b) Fearful (c) Happy (d) Sad (e) Angry

Fig. 3. Examples of Mel Spectrogram features for each emotion in the ASED dataset.

(SoE), is used to distinguish Anger from Happiness [10]. Different regions of the space correspond to different emotions.

Focusing around glottal closure instants, Kadiri et al. extract linear prediction residual and weighted linear prediction cepstral coefficients, using zero frequency filtering and linear prediction analysis [21]. These features are fed to an auto-associative neural network in order to classify inputs as either neutral or emotional. The authors extend this work using instantaneous fundamental frequency, strength of excitation and energy of excitation, once again extracted using zero frequency filtering and linear prediction analysis [20]. A decision tree approach is used to classify emotional speech into four classes, neutral, sad, angry and happy. Finally, instantaneous fundamental frequency, strength of excitation and energy of excitation are used as features to detect emotion, using only recordings of neutral speech as training examples [19]. This is done by mapping pairs of features into three different 2D spaces, and distinguishing regions in those spaces.

Meghanani et al. use log Mel Spectrograms and MFCC features to recognize dementia in speech [31]. A CNN-LSTM model achieves the best result using MFCC features. Chen et al. use log Mel Spectrograms, MFCC, delta MFCC and delta-delta MFCC for SER using a BiLSTM having a dual attention mechanism [8]. Experiments with the IEMOCAP dataset attain the best results with delta and delta-delta MFCCs as well as log Mel Spectrograms. Xia et al. experiment with a number of different traditional features for SER, including log Mel Spectrogram, pitch pulse, and modulation patterns [51]. These are compared with Wav2vec [40]. Evaluations use a CNN BiLSTM model trained on the IEMOCAP dataset and Wav2vec features are shown to be the best. Xu et al. [52] use log Mel Spectrograms as the features in their CNN model with attention. Data augmentation with vocal tract length perturbation (VTLP) [18] is also incorporated in the model, and evaluation is with the IEMOCAP dataset.

After reviewing many works on emotion recognition, it is clear that various forms of FilterBank, Mel Spectrograms and Mel-Frequency Cepstral Coefficients (MFCC) are broadly utilized in audio classification and speech emotion recognition. The following is a short overview of these methods.

3.2 FilterBanks

A Mel FilterBank is a triangular filter bank that works similarly to the human ear's perception of sound; thus it is more discriminative at lower frequencies and less discriminative at higher frequencies. Mel FilterBanks are used to provide a better resolution at low frequencies and less resolution at high frequencies [1, 37].

3.3 Mel Spectrograms

The spectrogram is the relationship between the time and frequency of the audio signal. Different emotions can show different patterns in the energy spectrum. Mel Spectrograms represent the audio signal in Mel-scale. The logarithmic form of the spectrogram can be better understood because humans perceive sound on a logarithmic scale. The human ear is observed to act as a sub-band filter bank. These filters overlap and are unevenly spaced on the frequency axis. In audio processing, the signal is considered stationary within 10 to 30 ms, and therefore a window with a shorter duration is selected [47]. Sample Mel Spectrogram plots for each of the five emotions in the ASED dataset are shown in Fig. 3.

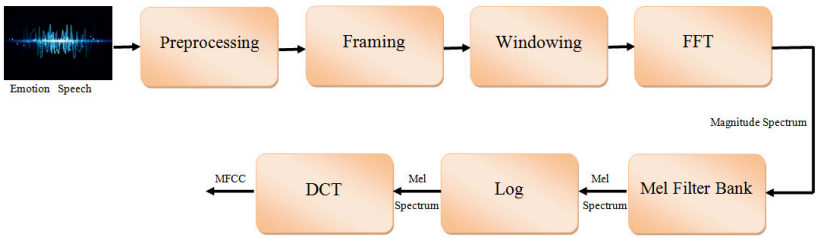
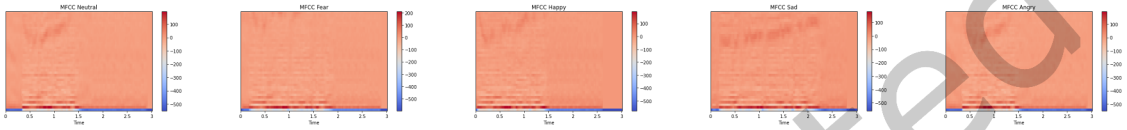


Fig. 4. Steps in producing MFCC features.



(a) Neutral (b) Fearful (c) Happy (d) Sad (e) Angry

Fig. 5. Examples of MFCC features for each emotion in the ASED dataset.

3.4 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficient (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 as shown in Fig. 4. The Mel scale is significant because it approximates the human perception of sound instead of being a linear scale [43]. Sample MFCC plots for each of the five emotions are shown in Fig. 5.

4 NETWORK ARCHITECTURES AND SETTINGS

4.1 Deep Learning Architectures

As discussed earlier, most of the previous studies employ CNN-based models for SER [28]. Among such models, the notable ones include AlexNet [25, 39], VGG [33, 44], and ResNet50 [14, 15], as well as LSTM [26].

VGG was one of the first CNN models used for signal processing. 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 VGGb and used it for our experiments. Fig. 6 and Table 9 show the architecture of VGGb. AlexNet, ResNet50 and LSTM were also used for comparison.

4.2 Experimental Setup

The standard code for the AlexNet, VGG, ResNet50 and LSTM models was used for the experiments. The VGG code was adapted to create VGGb. For the other models, the original network configuration and parameters were used.

We used the Keras deep learning library, version 2.0, with Tensorflow 1.6.0 backend [26] to build the emotion recognition models. The models were trained using a 2.30 GHz (CPU) Intel(R) Xeon(R) CPU. The Adam optimization algorithm was used to train our model, with categorical cross-entropy as the loss function; training stopped after 100 epochs, and the batch size was set to 16. Fig. 7 shows an outline of the experiments.

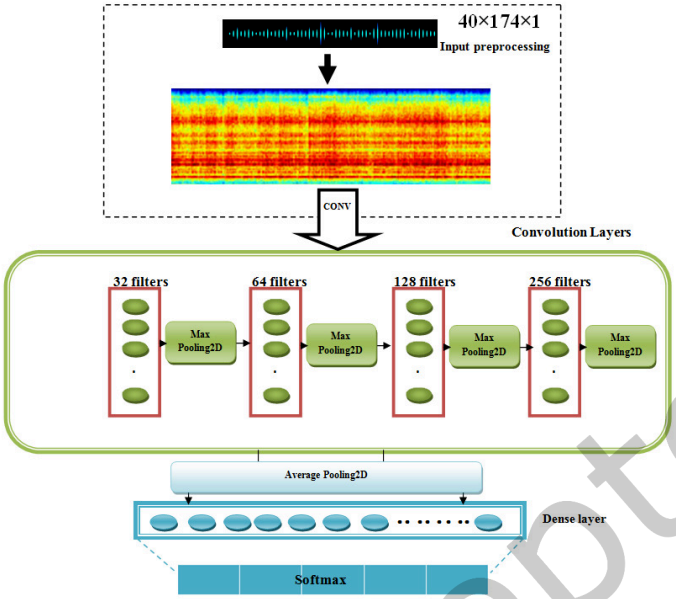


Fig. 6. VGGb architecture used in the SER experiments.

Table 9. Details of the VGGb architecture.

Input 40 x 174 x 1
Conv2D 1 (32 filters + kernel size = 2 + Relu) MaxPooling 2D (2,2) + Dropout (0.15)
Conv2D 2 (64 filters + kernel size = 2 + Relu) MaxPooling 2D (2,2) + Dropout (0.15)
Conv2D 3 (128 filters + kernel size = 2 + Relu) MaxPooling 2D (2,2) + Dropout (0.15)
Conv2D 4 (256 filters + kernel size = 2 + Relu) MaxPooling 2D (2,2) + Dropout (0.15)
AveragePooling2D Global Dense (64, Relu)
Dense (Soft Max 5)

5 EXPERIMENT 1: CHOICE OF FEATURES FOR AMHARIC SER

5.1 Outline

As we have mentioned, FilterBank, Mel Spectrograms and MFCC are three forms of feature which are widely used within SER systems for other languages. We therefore wished to determine which of these was most suitable for Amharic SER. Experiment 1 has four parts. First, a direct comparison of FilterBank, Mel Spectrograms and MFCC was carried out, by substituting each one in turn into the VGGb architecture and determining the resulting SER

performance. Second, the dataset sentences used for training and testing were varied in order to show that this did not affect results. Third, the dialects used for training and testing were varied. Fourth, the groups of speakers used for training and testing were varied. The effect of steps 2-4 was to validate the performance of the two feature types by performing a kind of cross-validation based on sentences, dialects and speaker groups.

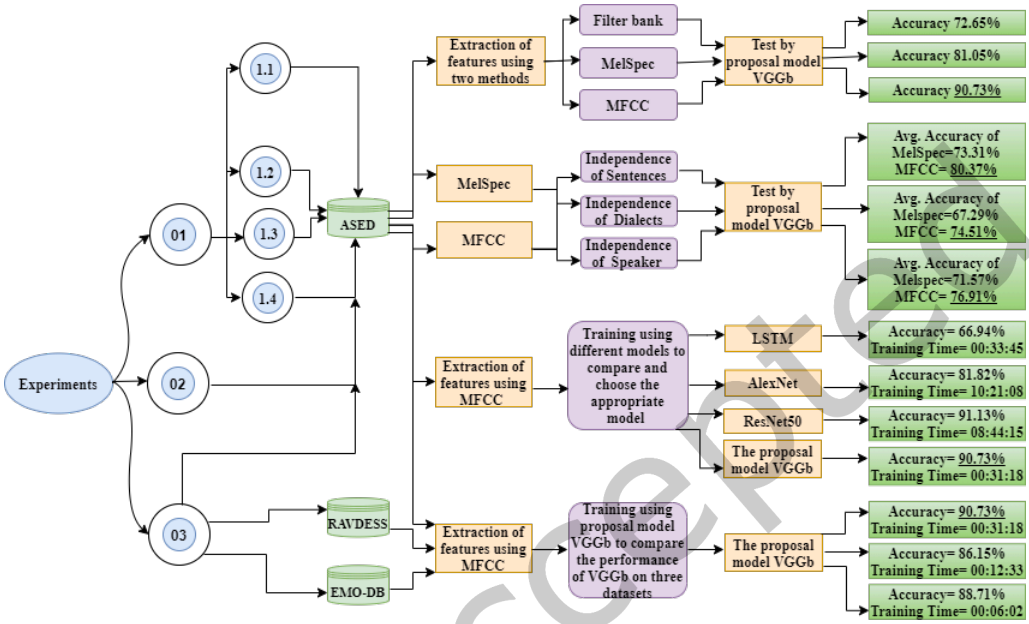


Fig. 7. General outline of the experiments that were performed in this study.

5.2 Experiment 1.1: Initial Comparison

We adopted the VGGb model for our experiments. Training and testing were performed with ASED data. We extracted the FilterBank features using the Python Speech Features library v0.6. Mel Spectrogram and MFCC features were obtained with the librosa v0.7.2 library [42]. The Mel Spectrogram was extracted with 128 bands, and MFCC with 40 bands in DMFCC (2D) form, according to the standard settings of the tool.

The model was trained in three forms, using just FilterBank, Mel Spectrogram, and MFCC features respectively. Data was split 90% train and 10% test. Each form of the model was trained five times with a different random train/test split, and the average result was reported.

The results are shown in Table 10. MFCC has proven to be effective at extracting the important features as has been demonstrated in previous experiments with VGG CNN-based SER models [13, 30, 33, 39]. Our results with VGGb confirmed this trend. The classification accuracy was 90.73% for MFCC as compared with 81.05% for Mel Spectrograms and 72.65% for FilterBank. As expected, FilterBank was the worst performing, so we continued our analysis with just the first two.

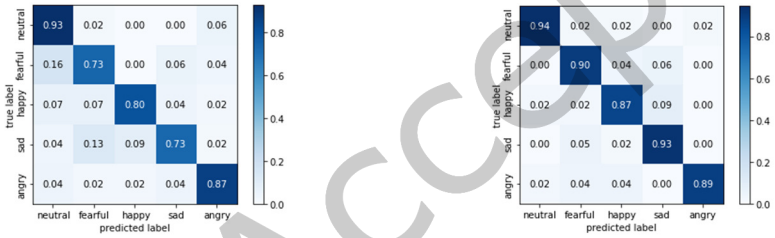
The confusion matrices for Mel Spectrogram and MFCC features are shown in Fig. 8 (a) and (b) respectively. The diagonal values show the percentage of true neutral examples predicted as neutral, and the same for the other emotion classes. We can immediately see that those diagonal values for MFCC (Fig. 8 (b)) are all higher than the corresponding values for Mel Spectrogram (Fig. 8 (a)). Moreover, it can be seen that the Mel Spectrogram performance for the fearful and sad classes is markedly lower (73% vs. 90%, and 73% vs. 93% respectively). This

is mainly because Mel Spectrogram predicts 13% of sad cases as fearful and 16% of fearful examples as neutral. Confusion between sad and fearful might be expected, as these are both strong emotions. However, it is striking that the Mel Spectrogram VGGb model also incorrectly predicts the happy class as fearful 7.0% of the time, happy as neutral, 7.0%, sad as fearful, 13.0%, and sad as happy, 9.0%, even though these emotional states are opposite to each other.

The Confusion matrices show that certain emotions are often mistaken for other emotions. Likewise, a few emotions are easier to recognize. This could be because the test data is based on participants pretending to speak with designated emotions, and people found it difficult to imitate certain emotions. It can be seen that neutral, fearful, sad, and angry are the emotions that are less difficult to recognize in Amharic speech as opposed to happy which is the most difficult.

Table 10. Experiment 1.1: Accuracy of the VGGb model trained on ASED data, using FilterBank, Mel Spectrogram and MFCC Features

Dataset	Features			Approach
ASED	FilterBank	Mel Spectrogram	MFCC	VGGb
	72.65%	81.05%	90.73%	



(a) Mel Spectrogram (b) MFCC

Fig. 8. Experiment 1.1: Confusion matrices for the VGGb model trained on ASED data using (a) Mel Spectrogram and (b) MFCC features.

5.3 Experiment 1.2: Independence of Sentences

Recall that the dataset consists of five sentences, one for each emotion, and two common sentences, which are spoken in every emotion. In this experiment, sentences were either used for training or for testing according to the scheme in Table 11. This was to determine how dependent the recognition of emotion is on the sentence used to express it. For each of the schemes shown in the table, VGGb was trained once using Mel Spectrogram and once using MFCC. The average result for each feature type is shown at the bottom. As can be seen in Table 11, the average performance for MFCC (80.37%) is higher than that for Mel Spectrogram (72.12%). The standard deviations for these values are 1.60 and 3.42 respectively. In each training scenario, MFCC is better than Mel Spectrogram; moreover, the standard deviations are low, indicating that there is only a small difference in performance between different scenarios, i.e. choices of sentences for training and testing.

5.4 Experiment 1.3: Independence of Dialects

ASED consists of recordings in four dialects, Gojjam, Wollo, Shewa and Gonder. In this experiment, the model is trained with utterances spoken in three dialects, and then tested with utterances in the fourth dialect, according

Table 11. Experiment 1.2: Accuracy of the VGGb model trained on ASED data using Mel Spectrogram and MFCC features. Training and testing sentences are varied in this experiment.

Training Sentences	Testing Sentences	Mel Spectrogram	MFCC
1+2+3+4+5	6+7	75.51%	82.10%
1+2+3+6+7	5+4	75.54%	81.08%
1+4+5+6+7	2+3	69.46%	78.37%
2+3+4+5+6	1+7	72.73%	79.91%
Average		**73.31%**	**80.37%**
St. Dev.		**2.89**	**1.60**

Table 12. Experiment 1.3: Accuracy of the VGGb model trained on ASED data using Mel Spectrogram and MFCC features. Speaker dialects used for training and testing are varied in this experiment.

Training Dialects	Testing Dialect	Mel Spectrogram	MFCC
Wollo+Shewa+Gonder	Gojjam	64.67%	72.03%
Gojjam+Shewa+Gonder	Wollo	68.36%	75.32%
Wollo+Gojjam+Gonder	Shewa	69.22%	76.56%
Wollo+Shewa+Gojjam	Gonder	66.91%	74.12%
Average		**67.29%**	**74.51%**
St. Dev.		**1.99**	**1.93**

to the scheme in Table 12. Similar to Experiment 1.2, this was to determine how dependent the emotion found is on the dialect used to express it. In Table 12, the average performance for MFCC (74.51%) is higher than that for Mel Spectrogram (67.29%). The standard deviations for these values are 1.93 and 1.99 respectively, suggesting that dialect is only making small changes to the performance figures and is not affecting the overall superiority of MFCC under all training scenarios.

5.5 Experiment 1.4: Independence of Speaker Category

ASED was recorded with three groups of speakers, Professional, Semi-professional and Amateur. To investigate how results might be affected by the speaker category, training was done with speakers from two groups and then tested with speakers of the third group, according to the scheme in Table 13. The results in that table once again show that performance for MFCC (76.91%) is higher than that for Mel Spectrogram (71.57%). The standard deviations for values are 1.53 and 2.50 respectively, indicating that the choice of speaker groups is not changing the result very much. Moreover, it is interesting that Amateur speakers do not give the lowest emotion recognition results. It is often stated that theatre actors should be used for generating emotion datasets, but the low standard deviations shown here do not lend support to that viewpoint.

Overall, Experiments 1.2-1.4 suggest that the advantage of MFCC over Mel Spectrogram as measured by emotion recognition accuracy is robust with respect to sentence choice, dialect and speaker category.

6 EXPERIMENT 2: COMPARISON OF SER METHODS FOR AMHARIC

Experiment 1 demonstrated, under four different measures, that MFCC features are better than Mel Spectrogram for use in Amharic SER. The aim of Experiment 2 was to determine which of the models discussed in Section 4 would give the best performance when applied to ASED data using MFCC features.

Table 13. Experiment 1.4: Accuracy of the VGGb model trained on ASED data using Mel Spectrogram and MFCC features. Speaker categories used for training and testing are varied in this experiment.

Training Groups	Testing Group	Mel Spectrogram	MFCC
Semi-professional + Amateur	Professional	74.36%	78.18%
Professional + Amateur	Semi-professional	69.52%	75.21%
Professional + Semi-professional	Amateur	70.84%	77.34%
Average		**71.57%**	**76.91%**
St. Dev.		**2.50**	**1.53**

Table 14. Experiment 2: Comparison of VGGb with other CNN and RNN models, all trained using the ASED dataset.

No.	Model	Training Time	Accuracy
1	LSTM	00:33:45	66.94%
2	AlexNet	10:21:08	81.82%
3	ResNet50	08:44:15	91.13%
4	VGGb	00:31:18	90.73%

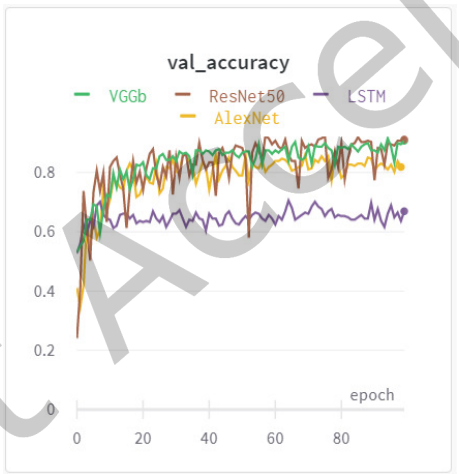


Fig. 9. Experiment 2: Val-accuracy training curves for the AlexNet, ResNet50, LSTM and VGGb models on the ASED dataset.

Recall that the four models are: AlexNet, VGGb, ResNet50 and LSTM. The network configuration for VGGb was the same as in the previous Experiment (Fig. 6, Table 9). For the other models, the standard configuration and settings were used. Once again, the librosa v0.7.2 library [42] was used to extract MFCC features to input to the models. Each model was trained five times using a 90%/10% test/train split and the average results were computed.

Results are presented in Table 14. Although ResNet50 had the highest accuracy (91.13%), VGGb was just 0.4% behind (90.73%). Moreover, VGGb was much faster than ResNet50 (00:31:18 vs. 08:44:15), showing that it is more efficient and hence more suitable for applying to large datasets.

The computational simplicity of VGGb in terms of training time and overall accuracy is again shown in Fig. 9 which represents the val-accuracy curve for the implemented models, AlexNet, ResNet50, LSTM and VGGb. The models are trained for 100 epochs. The curves show that after the 20th epoch, the val-accuracy starts stabilizing. The curve for VGGb looks like a better fit, while that for ResNet50 shows more noisy movements than the other models.

Table 15. Experiment 3: Performance of the VGGb model when trained on different datasets.

Model	Dataset	Training Time in min	Accuracy
VGGb	RAVDESS	00:12:33	86.15%
	EMO-DB	00:06:02	88.71%
	ASED	00:31:18	90.73%

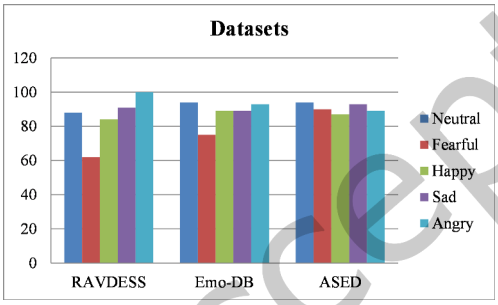


Fig. 10. Experiment 3: Performance of the VGGb model on the RAVDESS, Emo-DB and ASED datasets, broken down by emotion.

7 EXPERIMENT 3: COMPARISON OF ENGLISH, GERMAN AND AMHARIC SER

The aim of Experiment 3 was to compare the performance of VGGb on three datasets, RAVDESS (English), EMO-DB (German) and ASED (Amharic). Naturally there are interesting SER databases for many different languages (Table 7 earlier). However, in previous work, RAVDESS and EMO-DB have been extensively used for SER research, and are frequently quoted as baselines. This is why we chose to use these datasets, rather than others, for the comparison. Earlier, in Table 1 (Section 2.1) we provided summary information for the three datasets, which we also analyzed in detail in Section 2.4.

The VGGb model was trained using each of the three datasets. Network configuration and settings were there same as for previous experiments (Fig. 6, Table 9). Input features were MFCC, extracted by librosa v0.7.2. In each case, training was carried out five times with a 90%/10% test train split and the average results were computed.

Table 15 shows the results. Performance on ASED is the best (90.73%) but all accuracy figures are within a range of 4.2%. From this we can conclude that SER can be successfully performed on Amharic speech, and that the accuracy which can be attained is similar to that for English and German. The training time is longer (00:31:18) relative to EMO-DB (00:06:02) and RAVDESS (00:12:33) but this is because ASED is more than double the size of the other datasets. Fig. 10 shows the model's performance across the three datasets and across the five emotions. We can see that the fearful and sad classes are more clearly distinguished in ASED than in the other datasets. Conversely, the angry class is not as clearly distinguished as in the others.

8 CONCLUSIONS

In this paper, we first collected the Amharic Speech Emotion Dataset, which we believe to be the very first SER dataset for the Amharic language. It is based on five emotions, neutral, fearful, happy, sad and angry, contains 2,474 recordings and was created by 65 volunteer participants. We then conducted three experiments, based on a four-layer version of VGG which we call VGGb. Experiment 1 aimed to determine whether FilterBank, Mel Spectrogram or MFCC features were most suitable for SER in Amharic and to establish how robust this difference was relative to sentence choice, dialect and speaker group. Experiment 1.1 first compared the performance of VGGb on the ASED dataset using three forms of feature. Experiments 1.2-1.4 then tested the results using a form of cross-validation relative to sentences, dialects and speaker groups using Mel Spectrogram and MFCC. The results showed that MFCC features were superior to both Mel Spectrogram and FilterBank for Amharic SER and that this result was robust relative to the variations made.

In the second experiment, working with MFCC features and the ASED data, we compared the performance of four different models when applied to the SER task, AlexNet, ResNet50, LSTM, and VGGb. While ResNet50 was the best (91.13%), VGGb was very close (90.73%) and was considerably faster (training time for ResNet50 08:44:15, for VGGb 00:31:18).

The third experiment compared the performance of VGGb on three datasets in different languages, RAVDESS (English), EMO-DB (German) and ASED (Amharic). The accuracy values in each case were similar, suggesting that VGGb is equally applicable to the three languages.

Future work on ASED will include enlarging the scale of the database, using further elicitation techniques and adopting a dimensional emotion annotation strategy. We also plan to build a better SER model for Amharic, to study the extent to which emotion recognition is dependent on language, and to employ hybrid features (prosody with spectral) to train deep learning models.

ACKNOWLEDGMENTS

This work was supported by The National Key Research and Development Program of China under grant 2020YFC1521503.

REFERENCES

[1] Eiad Almekhlafi, AL-Makhlafi Moeen, Erlei Zhang, Jun Wang, and Jinye Peng. 2022. A classification benchmark for Arabic alphabet phonemes with diacritics in deep neural networks. *Computer Speech & Language* 71 (2022), 101274.

[2] Tadesse Anberbir, Michael Gasser, Tomio Takara, and Kim D Yoon. 2011. Grapheme-to-phoneme conversion for Amharic text-to-speech system. In *Proceedings of conference on human language technology for development, Bibliotheca Alexandrina, Alexandria.*

[3] audacity. 2020. audacity. http://www.audacityteam.org/ [Online; accessed 31-October-2020].

[4] Saikat Basu, Arnab Bag, Md Aftabuddin, Md Mahadevappa, Jayanta Mukherjee, and Rajlakshmi Guha. 2016. Effects of emotion on physiological signals. In *2016 IEEE Annual India Conference (INDICON).* IEEE, 1–6.

[5] Birhanu Belay, Tewodros Habtegebrial, Million Meshesha, Marcus Liwicki, Gebeyehu Belay, and Didier Stricker. 2020. Amharic OCR: An End-to-End Learning. *Applied Sciences* 10, 3 (2020), 1117.

[6] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology.*

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[8] Qiupu Chen and Guimin Huang. 2021. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence* 102 (2021), 104277.

[9] Kate Dupuis and M Kathleen Pichora-Fuller. 2011. Recognition of emotional speech for younger and older talkers: behavioural findings from the toronto emotional speech set. *Canadian Acoustics* 39, 3 (2011), 182–183.

[10] P Gangamohan, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and B Yegnanarayana. 2014. Excitation source features for discrimination of anger and happy emotions. In *Fifteenth Annual Conference of the International Speech Communication Association.*

[11] P Gangamohan, Sudarsana Reddy Kadiri, and B Yegnanarayana. 2013. Analysis of emotional speech at subsegmental level.. In *INTERSPEECH*, Vol. 2013. 1916–1920.
[12] P Gangamohan, Sudarsana Reddy Kadiri, and B Yegnanarayana. 2016. Analysis of emotional speech—A review. *Toward Robotic Socially Believable Behaving Systems-Volume I* (2016), 205–238.
[13] María Teresa García-Ordás, Héctor Alaiz-Moretón, José Alberto Benítez-Andrades, Isaías García-Rodríguez, Oscar García-Olalla, and Carmen Benavides. 2021. Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network. *Biomedical Signal Processing and Control* 69 (2021), 102946.
[14] Daniel George, Hongyu Shen, and EA Huerta. 2017. Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO. *arXiv preprint arXiv:1706.07446* (2017).
[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[16] Dias Issa, M Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* 59 (2020), 101894.
[17] Philip Jackson and S Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).
[18] Navdeep Jaitly and Geoffrey E Hinton. 2013. Vocal tract length perturbation (VTLP) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, Vol. 117. 21.
[19] Sudarsana Reddy Kadiri, P Gangamohan, Suryakanth V Gangashetty, Paavo Alku, and B Yegnanarayana. 2020. Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference. *Circuits, Systems, and Signal Processing* 39, 9 (2020), 4459–4481.
[20] Sudarsana Reddy Kadiri, P Gangamohan, Suryakanth V Gangashetty, and Bayya Yegnanarayana. 2015. Analysis of excitation source features of speech for emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.
[21] Sudarsana Reddy Kadiri, P Gangamohan, and B Yegnanarayana. 2014. Discriminating neutral and emotional speech using neural networks. In *Proceedings of the 11th International Conference on Natural Language Processing*. 214–221.
[22] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. 2019. Automatic Speech Emotion Recognition Using Machine Learning. In *Social Media and Machine Learning*. IntechOpen.
[23] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* 7 (2019), 117327–117345.
[24] Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao. 2011. IITKGP-SEHSC: Hindi speech corpus for emotion analysis. In *2011 International conference on devices and communications (ICDeCom)*. IEEE, 1–5.
[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[26] Harshawardhan S Kumbhar and Sheetal U Bhandari. 2019. Speech Emotion Recognition using MFCC features and LSTM network. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 1–3.
[27] Soonil Kwon et al. 2020. CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network. *Mathematics* 8, 12 (2020), 2133.
[28] Soonil Kwon et al. 2020. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* 20, 1 (2020), 183.
[29] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. 2017. CHEAVD: a Chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing* 8, 6 (2017), 913–924.
[30] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.
[31] Amit Meghanani, CS Anoop, and AG Ramakrishnan. 2021. An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 670–677.
[32] Abrham Debasu Mengistu and Mamo Abebe Bedane. 2017. Text Independent Amharic Language Dialect Recognition using Neuro-Fuzzy Gaussian Membership Function. *International Journal of Advanced Studies in Computers, Science and Engineering* 6, 6 (2017), 30.
[33] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. 2016. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440* (2016).
[34] Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. *Computer Science & Information Technology* (2018), 41–55.
[35] Wondwossen Mulugeta and Michael Gasser. 2012. Learning morphological rules for Amharic verbs using inductive logic programming. *Language Technology for Normalisation of Less-Resourced Languages* 7 (2012).
[36] M. Kathleen Pichora-Fuller and Kate Dupuis. 2010. Toronto emotional speech set (TESS). https://doi.org/10.5683/SP2/E8H2MF
[37] VM Praseetha and PP Joby. 2021. Speech emotion recognition using data augmentation. *International Journal of Speech Technology* (2021), 1–10.
[38] Justus J Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission* (2005).

[39] Muhammad Sajjad, Soonil Kwon, et al. 2020. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access* 8 (2020), 79861–79875.
[40] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).
[41] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 5 (2018), 90–99.
[42] Riffat Sharmin, Shantanu Kumar Rahut, and Mohammad Rezwanul Huq. 2020. Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network. *Procedia Computer Science* 171 (2020), 1381–1388.
[43] Akash Shaw, Rohan Kumar Vardhan, and Siddharth Saxena. 2016. Emotion recognition and classification in speech using Artificial neural networks. *International Journal of Computer Applications* 145, 8 (2016), 5–9.
[44] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[45] Constantin Stanislavski. 1936. An Actor Prepares (New York. *Theatre Arts* (1936), 38.
[46] Siniša Suzić, Darko Pekar, and Vlado Delić. 2014. On the realization of AnSpeechCollector, system for creating transcribed speech database. In *2014 22nd Telecommunications Forum Telfor (TELFOR)*. IEEE, 533–536.
[47] Rishabh N Tak, Dharmesh M Agrawal, and Hemant A Patil. 2017. Novel phase encoded mel filterbank energies for environmental sound classification. In *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 317–325.
[48] Vishal P Tank and SK Hadia. 2020. Creation of speech corpus for emotion analysis in Gujarati language and its evaluation by various speech parameters. *International Journal of Electrical and Computer Engineering* 10, 5 (2020), 4752.
[49] Gustav Sto Tomas. 2019. Speech Emotion Recognition Using Convolutional Neural Networks. (2019).
[50] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. 2018. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society* 66, 6 (2018), 457–467.
[51] Yangyang Xia, Li-Wei Chen, Alexander Rudnicky, and Richard M Stern. 2021. Temporal Context in Speech Emotion Recognition. In *Proc. Interspeech*, Vol. 2021. 3370–3374.
[52] Mingke Xu, Fan Zhang, Xiaodong Cui, and Wei Zhang. 2021. Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6319–6323.