Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-023-07920-8

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

# Spectro-Temporal Energy Ratio Features for Single-Corpus and Cross-Corpus Experiments in Speech Emotion Recognition

**Cevahir Parlak**[1,3] · **Banu Diri**[2] · **Yusuf Altun**[1]

Received: 17 April 2022 / Accepted: 30 April 2023
© King Fahd University of Petroleum & Minerals 2023

**Abstract**
In this study, novel Spectro-Temporal Energy Ratio features based on the formants of vowels, linearly spaced low-frequency, and logarithmically spaced high-frequency parts of the human auditory system are introduced to implement single- and cross-corpus speech emotion recognition experiments. Since the underlying dynamics and characteristics of speech recognition and speech emotion recognition differ too much, designing an emotion-recognition-specific filter bank is mandatory. The proposed features will formulate a novel filter bank strategy to construct 7 trapezoidal filter banks. These novel filter banks differ from Mel and Bark scales in shape and frequency regions and are targeted to generalize the feature space. Cross-corpus experimentation is a step forward in speech emotion recognition, but the researchers are usually chagrined at its results. Our goal is to create a feature set that is robust and resistant to cross-corporal variations using various feature selection algorithms. We will prove this by shrinking the dimension of the feature space from 6984 down to 128 while boosting the accuracy using SVM, RBM, and sVGG (small-VGG) classifiers. Although RBMs are considered no longer fashionable, we will show that they can achieve outstanding jobs when tuned properly. This paper discloses a striking 90.65% accuracy rate harnessing STER features on EmoDB.

**Keywords** Speech emotion recognition · Filter banks · Feature selection · SVM · Restricted Boltzmann machines · Convolutional neural networks

## 1 Introduction

Emotion is one of the most complicated and distinct features of humankind and plays a significant role in human–human interaction. Humans express their feelings in a variety of ways such as speech (shouting, crying, jitter, shimmer, etc.), facial expressions (mimics, smile), and body language (body, head, upper and lower extremity movements). Therefore, many surveys began digging into this field using unimodal and multimodal affect recognition (Gunes & Piccardi 2008; Zeng & et al. 2008). Emotion recognition or generally pattern recognition may seem like an ordinary task for humans, but when transferred into the machine domain it turns out to be a very formidable process. The human pattern recognition system is a matchless engineering marvel. It consists of 100 billion neurons each with 1000 to 15,000 synaptic connections [1, 2]. Emotions of love, hatred, happiness, sadness, fear, excitement, and boredom shape our lives. Now we begin to interact with the machines through emotional contacts. Speech synthesis techniques have made great improvements, enabling machines to speak as naturally as humans. Embellished with emotionally colored speech, human–machine interaction will be much more attractive, natural, and beneficial for many, including impaired people (Iida et al., 2003). The introduction of emotion in computer science started in the mid-1990s with speech emotion recognition. Most of the earlier emotion recognition work was unimodal and focused on basic emotion types (Cowie et al., 2007). These initial studies sowed the seeds of the foundations of humanoids and have progressed considerably. Today two mainstream definitions of emotion compete. The first is the categorical approach pioneered by Ekman et al. [3], and the second is the continuous definition of emotion proposed by Plutchnik

✉ Cevahir Parlak
cevahirparlak@gmail.com

[1] Computer Engineering Department, Duzce University, Duzce, Turkey

[2] Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey

[3] Information Technologies Department, T C Ataşehir Adıgüzel Vocational School, İstanbul, Turkey

Published online: 27 May 2023

[4], Whissell [5], and Cowie [6]. The categorical approach classifies emotion in certain groups, while the continuous one embraces dimensional classification, which states that emotions can be created by mixing 2 or 3 dimensions, like blending RGB color values to create new colors. Nowadays this theory is used as valence, arousal, and control (dominance) states (Russell et al., 1981; Scherer, 2005) in emotion recognition applications. The main framework of emotion recognition studies can be divided into three stages: creation of emotional databases, feature extraction, and classification. Emotional databases consisting of emotional speech samples can be classified as natural, simulated, and induced according to the recording methods. Natural databases are recorded under natural conditions as they actually occur; they can be difficult to collect, especially with regard to ethical issues as well as technical challenges, including high-quality acoustic recordings. Generally, scientists want to use natural speech samples as much as possible. Simulated or acted databases are recorded under controlled conditions by asking the subjects to utter specific emotions, using various induction procedures, including sometimes even the usage of drugs in a studio or anechoic chamber. Although there are many objections to simulated databases, they are the most used tools in speech emotion research. Some claim [7, 8] that simulated emotions do not carry all the features of natural emotions. Nevertheless, acted databases have proven to be useful for emotional research. Annotation of speech samples is another important and strenuous phase in the creation stage of the databases and involves the labeling of sample classes.

The second stage is the extraction of relevant features for the classification of emotional speech samples. There are two kinds of features when we talk about speech utterances: spectral features and temporal features [9]. Spectral features deal with the frequency components of the utterances, while temporal features deal with time-related features such as jitter, shimmer, duration, voice quality, and their statistical values such as mean, variance, skewness, and kurtosis.

The third stage is the classification of the annotated data. There is no standard way for classifying emotional speech data. In this work, we will try SVM, Deep Belief Net, and sVGGNet.

In the beginning of our research, we noticed that some frequency regions of audible spectrum are closely related to some emotional classes. For instance, the ratio between the linear region (below 1100 Hz) and logarithmic region (above 1100 Hz) of auditory spectrum has significant effect in the discrimination of angry, neutral, and sad emotional classes. This finding led us to extend our research to construct a filter bank strategy. These filter bank energies, ratios between them, and their statistical functionals are used as new features creating a 315-feature space in total. Our novel contributions are listed as follows:

i. We constructed 5 trapezoidal filters centered around the formants of the vowel sounds of human vocal tract. We collect the spectral magnitude under these filters. We utilize overlapping trapezoidal shape because there is a transition region between neighboring vowel sounds where a sound can be perceived like both and then a flat region follows where the actual vowel sound can purely be identified. We also included our fundamental frequency related prosodic features.
ii. We added 2 more trapezoidal filters covering the entire linearly (below 1100 Hz) and logarithmically spaced (above 1100 Hz) regions of the human audible spectrum. We also use the ratio (LHRate) of total spectral magnitude under these frequency regions covered by these two filters as novel feature. This ratio is very effective exclusively in the discrimination of angry, sad, and neutral speech samples.
iii. Statistical functions such as max, min, standard deviation, skewness, kurtosis, and quartiles are also applied to the above-mentioned features throughout the entire frames of the speech samples. Detailed information on the proposed features is given in Section 5.

This paper is organized as follows: Sect. 2 evaluates the current literature, Sect. 3 introduces the datasets used in the experiments, Sect. 4 discusses the experimental setup and methods, Sect. 5 presents proposed STER-based features, Sect. 6 provides the results of single-corpus and cross-corpus experiments, and finally, in Sect. 7, we conclude this paper and discuss future works.

## 2 Literature Review

Emotion recognition has been an exceedingly popular research area during the last two decades. Many datasets have been created in different languages, and numerous applications have been implemented [10–12]. These applications include emotion detection in different languages [13], cross-corpus experiments [14–16], implementation of Nonnegative Matrix Factorization [17], and use of different feature descriptors such as Fourier parameters and spectrograms [18–22]. Autoencoders and GANs (Generative Adversarial Networks) have also been used to automatically generate data samples and enhance speech emotion recognition [23–25].

The primary goal of these studies has evolved from a handful of basic emotions to more than 12 emotions in a single corpus. As computer and artificial intelligence technology came of age, these works began to span a wide range of applications [16, 26]. Despite the extremely high levels of success rates on single-corpus implementations, namely on simulated databases, when moved to the cross-corpora

domain these applications produce disappointing outcomes. Especially natural databases pose exceedingly difficult challenges for researchers.

Recent work with LSTM-based cross-corpus implementation on RECOLA, SEMAINE, and CreativeIT databases was tried by Kaya et al. [27]. Some deep-learning-related work with temporal information and spectrograms [28–31] and end-to-end studies using Convolutional Neural Networks on the RECOLA database can be found in [32, 33]. In [33], the proposed model was trained end-to-end and used Convolutional Neural Network (CNN) to extract relevant features from the raw speech signal, and a 2-layer Long Short-Term Memory (LSTM) was stacked on top of it to capture the contextual information from the data. The model significantly outperforms the state of the art in terms of Concordance-Correlation-Coefficient in the RECOLA database.

In [16], Shah, Chakrabarti, and Spanias performed experiments on the USC IEMOCAP [34] and the SEMAINE [35] datasets that contain acted and spontaneous emotions of four types including neutral, sad, happy, and angry. When IEMOCAP is used as the training set and SEMAINE as the test set, they attained 76.31% success rate on the arousal state of speech samples which are longer than 6 s using the 384-feature set of IS09 (Interspeech 09). On the valence state, 59.36% unweighted average recall rate was achieved using sRSM (supervised replicated softmax model).

In [36], four datasets ACC, IEMOCAP, EMA, and EmoDB in three languages (Chinese, German, English) and 4 emotions (angry, happy, neutral, sad) were used. Three datasets are acted, and, while the English one, is an elicited dataset. OpenSMILE tool was used to extract 1582 features and classification was performed with the SVM-SMO of the Weka tool. An 89.4% accuracy was achieved on single-corpus experiments on the EmoDB dataset.

In [37], Eyben et al. used SmartKOM, Aibo, SAL, and VAM datasets for their cross-corpus experiments, making three of them training sets and the remaining as the test set. These four corpora vary greatly in language, emotion types, age of subjects, annotation, and implementation scenario, which cause large inconsistencies in the experiments. They found unweighted average success rates ranging from 50.4% to 58.1% with the SVM-SMO classifier of the Weka tool.

Schuller et al. also conducted cross-corpus experiments on ABC, AVIC, DES, EmoDB, ENTERFACE, SAL, SUSAS, and VAM datasets in [38]. They used one database as the training set and all the other 7 datasets as the test set. Training on the VAM dataset gave the highest average unweighted accuracy of 67.7% in the arousal state and training on the DES dataset gave 54.8% mean unweighted accuracy in the valence state. They also reported that the valence state is more complicated for the classifier to estimate. Training on the EmoDB dataset produces 62.5% and 51.4% mean unweighted average accuracy rates for the arousal and valence states respectively with the SVM classifier.

In [39], Kismet, BabyEars, DES, and EmoDB databases were used for cross-corpus evaluation. They reported accuracy rates as low as 20.8% on cross-corpus classification despite the 80.7% accuracy rate on single-corpus works with the SVM classifier when EmoDB was used as the training set and DES as the test set. On the other hand, when DES was used as training and EmoDB as test set, they achieved a 52.6% accuracy rate on cross-corpus studies versus the 64.90% accuracy on single-corpus studies with the DES dataset.

Many applications have been developed for emotion recognition for video, speech, and EMG data using RBMs. In [40], an ensemble of random deep belief networks (RDBN) is used for speech emotion recognition. It first extracts the low-level features and then constructs many random subspaces, presenting them to a DBN to extract higher-level features as the classifier input; finally, a majority voting mechanism is used to obtain the final output label. Experiments show that RDBN has better accuracy than its rivals. RBM is applied to video and audio-based databases in [41] and EDA, PPG, and zEMG data in [42].

In [43], authors tried to combine acoustic features with text to improve speech emotion recognition with a DNN-trained SVM classifier. They used IEMOCAP and MSP-IMPROV datasets with acoustics and text transcriptions. They showed that the fusion of acoustic features with text features produces the best results.

In [44], Firdaus et al. present a new large-scale balanced Multimodal Multi-label Emotion, Intensity, and Sentiment Dialogue dataset (MEISD). This dataset is collected from television series and contains textual, audio, and visual features. They also present the results of modality, sentiment, and emotional label classification with CNN, LSTM, and RNN methods.

In [45], the authors introduce the Multimodal EmotionLines Dataset (MELD). This new dataset comprises 13,000 speech samples from 1,433 dialogues of a TV series. Utterances are annotated with emotion and sentiment labels. The MELD dataset contains audio, visual, and textual modalities.

In [46], Yin et al. studied a novel Graph Convolutional Neural Network (GCNN) in conjunction with the LSTM network to capture temporal features on EGG data of the DEAP dataset. The extracted features are fed into a dense layer for final emotional classification to obtain state of the art results.

In [47], Pakyurek et al. discussed the MFCC and histograms of MFCC features on EmoDB and the newly introduced PAU database with an SVM classifier. They showed that combining MFCC and histograms of MFCC

**Table 1** Speaker numbers by emotion type in EmoSTAR

| | Angry | Neutral | Happy | Sad |
|---|---|---|---|---|
| Male | 25 | 25 | 7 | 4 |
| Female | 10 | 14 | 12 | 9 |
| Total = 106 | 35 | 39 | 19 | 13 |

leads to better performance on speech emotion recognition applications.

## 3 Datasets

EmoSTAR [48] is the first database we employ to conduct our experiments. EmoSTAR consists of 393 angry, neutral, happy, and sad speech samples in the Turkish and English languages. Neutral utterances are mostly from television news channels. Therefore, we can assume neutral utterances as natural speech samples, and they are very abundant, easy to obtain compared to the other types of samples. Angry utterances are professionally acted speech samples from TV series and movies. Happy utterances of EmoSTAR were collected from the speeches of award winners of various ceremonies such as the Oscar, Golden Globe, and American Music Awards (AMA) as the actor/actress was happily receiving the Oscar statuette, AMA, or Golden Globe awards in his/her hand and most likely represent examples of natural, spontaneous happiness. Sad samples are from internet videos showing people actually crying and talking, although some are acted samples of sad speech. In most of the sad speech samples speakers who are crying, at the same time are talking; for example, mourning a sister or close friend who died unexpectedly. Obtaining sad speech samples is the most difficult part of the construction of EmoSTAR. The database contains different speech sentences with lengths of the speech samples varying between 2.2 and 14.5 s, spoken by 106 different speakers. As such, it is a remarkably diverse dataset in terms of speaker-specific features and eliminates the need to use the LOSO (Leave One Speaker Out) method. Therefore, experiments on the EmoSTAR dataset can be considered cross-subject. The number of sentences by a speaker varies from 1 to 29. The number of speakers by gender and emotion type in the EmoSTAR dataset is detailed in Table 1.

EmoSTAR is still under development to contain more samples, emotion types, and modalities. It is freely available for research purposes only via email to the author. In the EmoSTAR dataset, most of the samples have been annotated by the owners of the video; however, for our analysis, we also made our own assessments: we listened to the audio signals while watching the video files. If we agreed with the annotations done by the video owners, we extracted the audio part from the video and included these annotated wave samples in our dataset. Detailed numbers of emotion types of EmoSTAR are shown in Table 2.

The other database is the Berlin Emotional Database (EmoDB) [49]. It is comprised of 535 speech samples in 7 emotion types as shown in Table 3. Berlin Emotional Database is in the German language with 5 short and 5 long sentences. The highest weighted accuracy on EmoDB in the classification of 7 emotions is 89.5% which was achieved by Eyben et al. in [50] using 5967 features and tenfold stratified cross-validation with LibSVM. In our work, we show a greater success rate utilizing a far smaller number of features.

In EmoDB, perception tests were carried out by 20 subjects. They presented the speech utterances by computer and were asked to decide the emotion of the speaker as well as how convincing the speaker was in expressing that emotion. They could listen to each speech sample once. Utterances that were recognized as those intended by the speaker at a rate higher than 80% and naturalness higher than 60% were selected to be used in the analysis. In Table 3, A is Angry, N is Neutral, H is Happy, S is Sad, B is Bored, D is Disgust, and F stands for Fear speech utterances.

**Table 2** Emotions in EmoSTAR (E = English, T = Turkish)

| | Angry | Neutral | Happy | Sad |
|---|---|---|---|---|
| Male | 33 E−30 T | 35 E−34 T | 45 E | 12 E |
| Female | 40 E | 37 E−20 T | 37 E | 51 E−19 T |
| Total = 393 | 103 | 126 | 82 | 82 |

**Table 3** Emotions in EmoDB dataset (German)

| | A | N | H | S | B | D | F |
|---|---|---|---|---|---|---|---|
| Male | 60 | 39 | 27 | 25 | 35 | 11 | 36 |
| Female | 67 | 40 | 44 | 37 | 46 | 35 | 33 |
| Total = 535 | 127 | 79 | 71 | 62 | 81 | 46 | 69 |

**Table 4** Number of speech samples for the ENTERFACE, RML, SAVEE, and IEMOCAP datasets

| | ENTERFACE | RML | SAVEE | IEMOCAP |
|---|---|---|---|---|
| Angry | 215 | 120 | 60 | 880 |
| Happy | 202 | 120 | 60 | 461 |
| Neutral | – | – | 120 | 1276 |
| Sad | 205 | 113 | 60 | 476 |

**Table 5** Feature numbers in the Emobase and STER feature sets

| | # of Features |
|---|---|
| Emobase | 988 (26 LLD + 26 Δ) * 19 functionals |
| STER | 315 (10 LLD + 10Δ + 10 ΔΔ)*10 functionals + 15 |
| Emobase + STER | 1303 |

The IEMOCAP dataset contains 10,039 samples, but we eliminated some long sentences. IEMOCAP contains 11 categories; however, we used only 4 of them to make it compatible with the other datasets as shown in Table 4. In the IEMOCAP dataset, categorical emotional annotation was conducted with 6 human evaluators. In the annotation sessions, 3 different annotators evaluated each utterance. They used Fleiss Kappa statistics to analyze the inter-annotator agreement. In an attempt to enlarge our study, we included ENTERFACE 2005 [51], RML [52], and SAVEE [53] datasets that are freely available. The total number of speech samples of these datasets is detailed in Table 4. In our work, we select a different approach by merging 5 datasets to create a larger one. We merged EmoDB, EmoSTAR, ENTERFACE, RML, and SAVEE datasets using all the angry, happy, neutral, and sad samples and named it as 5Mix dataset. This 5Mix dataset is the one we use for cross-corpus experiments with the IEMOCAP dataset, and the total number of speech samples is 2007 in the 5Mix dataset.

## 4 Experimental Setup and Methods

The feature extraction phase of the experiments was conducted with the openSMILE tool [54]. For the main feature set, we selected the Emobase feature set which contains 988 features as shown in Table 5. Detailed information can be found in [54] and also in the openSMILE handbook. Although we conducted experiments using the Emo_large set of the openSMILE tool which contains 6669 features, we do not give its results because they are usually below the results of the Emobase set except for some rare cases. The proposed STER-based features are extracted by EmoSTAR's application. SVM classification tasks are run by the Weka tool [55].

In the Emobase feature set, we have 26 Low-Level-Descriptors (LLD), first-order deltas, and 19 statistical functions of maximum, minimum, range, maximum position, minimum position, arithmetic mean, 1st and 2nd linear regression coefficients, linear regression arithmetic and quadratic errors, standard deviation, skewness, kurtosis, 1st, 2nd, and 3rd quartiles, 1st, 2nd, and 3rd interquartile ranges. Total number of features is 988.

Robust first- and second-order delta parameters can effectively be obtained using the following formulas:

$$\Delta C_k(t) = \frac{\sum_{\tau=-M}^{M} \tau \cdot C_k(t+\tau)}{\sum_{\tau=-M}^{\tau=+M} \tau^2} \tag{1}$$

$$\Delta^2 C_k(t) = \frac{\sum_{\tau=-M}^{M} \tau \cdot \Delta C_k(t+\tau)}{\sum_{\tau=-M}^{\tau=+M} \tau^2} \tag{2}$$

where $M$ is the number of frames before and after the frame $t$. Typically M is taken as 2 and determines the sharpness of the coefficients. A visualization of the delta coefficients of a signal is shown in Fig. 1. Notice the values of delta at the points where sharp changes occur.

In the STER feature set, we have 10 Low-Level descriptors including f0, sum of the magnitude of frequency spectrum under the trapezoidal filters from 1st through 5th shown in Fig. 2, and LHRATIO (Ratio of sum of magnitude of frequency spectrum under the 6th and 7th trapezoidal filters shown in Fig. 3). Total number of features is 315.

The Restricted Boltzmann Machine is an energy-based graphical Markovian Random Field model with visible and hidden binary stochastic units. It is a special variant of the Boltzmann machine without visible-to-visible and hidden-to-hidden connections. RBMs were invented by Paul Smolensky [56] with the name of Harmonium in 1986 and later developed by Geoffrey Hinton and his collaborators after 2006 kick-starting the Deep Learning era [57]. The energy of state {v, h} of an RBM can be formulated in scalar form as follows:

$$E(v, h) = -\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij} v_i h_j - \sum_{i=1}^{M} b_i v_i - \sum_{i=1}^{N} c_j h_j \tag{3}$$

This form is analogous to Hopfield networks and $w_{ij}$ represents the bidirectional symmetric weights between visible layer $i$, and hidden layer $j$, $b_i$, and $c_j$ are biases from visible to hidden and hidden to visible layers, respectively, $M$ is the number of inputs, and $N$ is the number of hidden nodes. Thus, the probability of visible units to be 1 is independent of other visible units and the same applies to the hidden units.

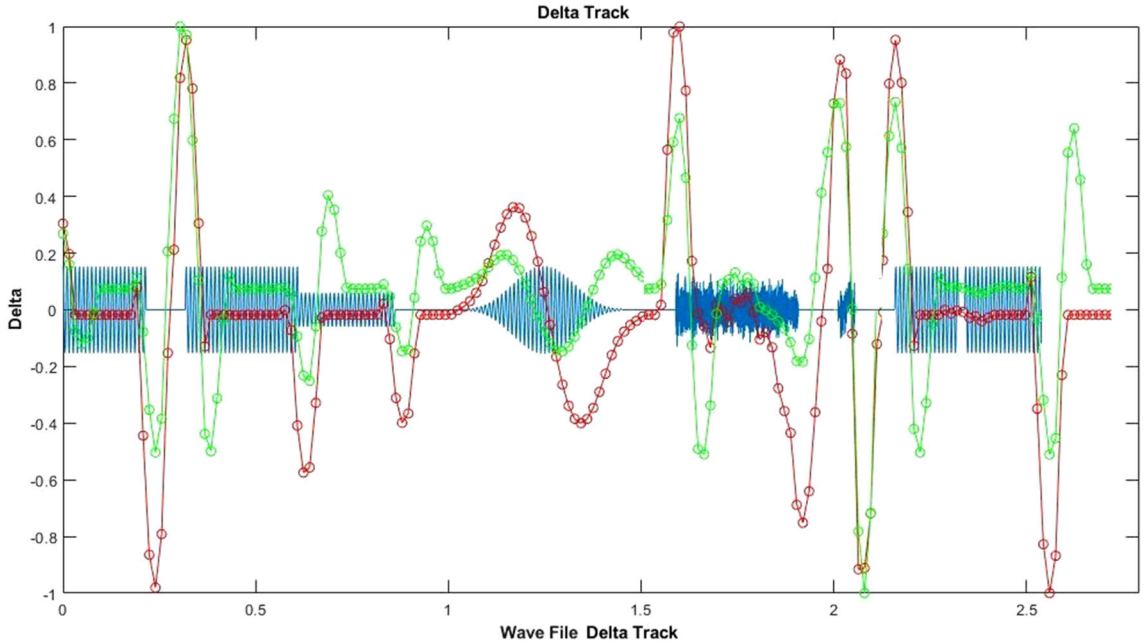



**Fig. 1** A signal (blue) and its corresponding first-order (red) and second-order (green) delta track

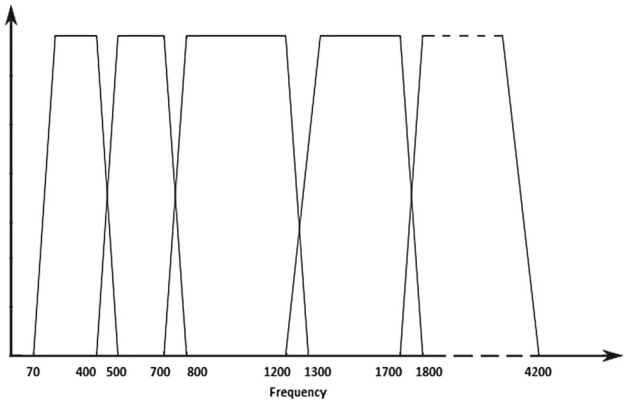



**Fig. 2** First 5 formant-based trapezoidal filter banks designed for speech emotion recognition

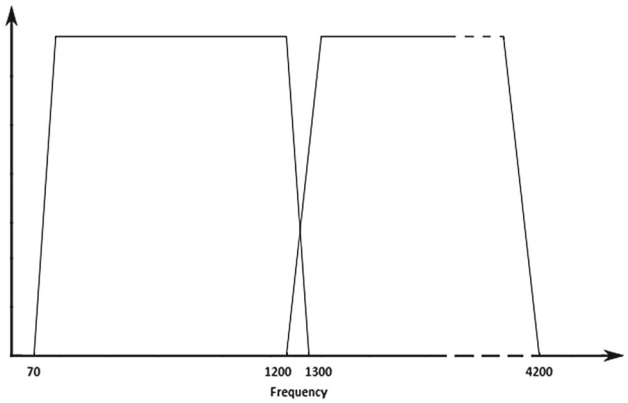



**Fig. 3** 6th and 7th trapezoidal filter banks of linear and logarithmic regions of the human auditory system

The visible and hidden states are independent of one another. This makes sampling remarkably effective, so that we can sample all visible and hidden units simultaneously. The original RBM is binary, but we can relax the Bernoulli constraint to make it suitable for stochastic processes. RBMs are the starting point of deep learning with the extension of stacking RBMs on top of one another to form Deep Belief Nets [58–62]. Although they are considered no longer fashionable, this work will show that they can accomplish outstanding jobs fast and efficiently when tuned properly without using too wide or too deep of network architectures. Weight decay is used to prevent unnecessary deviation, particularly for the sigmoid activation function. Momentum is to avoid unneeded oscillations of weight updates. A typical value for the size of a mini-batch is between 10 and 100, learning rate is empirically determined, weight decay is between 0.01 and 0.00001, momentum is initially 0.5, and then 0.9. In model learning, the derivative of the log-likelihood can be expressed as [58]:

$$\frac{\partial L(\theta)}{\partial w_{ij}} = E_{P_{data}}[v_i h_j] - E_{p_\theta}\left[v_j b_j\right] \tag{4}$$

Here the first term is called the data term (clamped term) and the second term is called the model term (unclamped term). The second term is difficult to compute due to the exponential possibilities. Since RBMs are initialized randomly, we run ten evaluations for each experiment and take the average of them for a fair comparison. All columns of data are normalized between [0, 1] independently and all the experiments are reproducible. In our experiments, we used

**Table 6** Small VGGNet model

| Layer | Output shape | Param # |
|---|---|---|
| Conv2D(32, (3, 3)) | 10, 99, 32 | 320 |
| Activation("relu") | 10, 99, 32 | 0 |
| BatchNormalization | 10, 99, 32 | 128 |
| Conv2D(32, (3, 3)) | 10, 99, 32 | 9248 |
| Activation("relu") | 10, 99, 32 | 0 |
| BatchNormalization | 10, 99, 32 | 128 |
| MaxPooling2D(2, 2) | 5, 49, 32 | 0 |
| Dropout(0.25) | 5, 49, 32 | 0 |
| Conv2D(64, (3, 3)) | 5, 49, 64 | 18,496 |
| Activation("relu") | 5, 49, 64 | 0 |
| BatchNormalization | 5, 49, 64 | 256 |
| Conv2D(64, (3, 3)) | 5, 49, 64 | 36,928 |
| Activation("relu") | 5, 49, 64 | 0 |
| BatchNormalization | 5, 49, 64 | 256 |
| MaxPooling2D(2, 2) | 2, 24, 64 | 0 |
| Dropout(0.25) | 2, 24, 64 | 0 |
| Conv2D(128, (3, 3)) | 2, 24, 128 | 73,856 |
| Activation("relu") | 2, 24, 128 | 0 |
| BatchNormalization | 2, 24, 128 | 512 |
| Conv2D(128, (3, 3)) | 2, 24, 128 | 147,584 |
| Activation("relu") | 2, 24, 128 | 0 |
| BatchNormalization | 2, 24, 128 | 512 |
| MaxPooling2D(2, 2) | 1, 12, 128 | 0 |
| Dropout(0.25) | 1, 12, 128 | 0 |
| Flatten | 1536 | 0 |
| Dense(512) | 512 | 786,944 |
| Activation("relu") | 512 | 0 |
| BatchNormalization | 512 | 2048 |
| Dropout(0.5) | 512 | 0 |
| Dense('softmax') | 4 | 2052 |

the Deep Belief Net implementation of Tanaka [62] with 125 hidden units, 1000 iterations, 0.025 step ratio, 0.0002 weight cost, 0.5 initial momentum, 0.9 final momentum, 25 batch size, and cross-entropy. Despite our rigorous efforts to implement wider and deeper networks, we were not able to obtain better results; therefore, we used one-layer RBM.

In CD-k (k-step Contrastive Divergence) learning, visible units are initialized with input variables, and we sample the hidden and visible layers alternately with infinite repetitions. After infinite repetitions, we will reach an equilibrium state. In [58], it is proven that only a few iterations are sufficient to reach the equilibrium state and usually one step is enough using Contrastive Divergence.

Although RBMs were very popular and ignited the deep learning era, after the developments in Graphical Processing Units, CNNs (Convolutional Neural Networks) started to dominate the field. CNNs are widely used in image classifications, but they can be run for speech processing applications by introducing the speech signal features as 2-D data. In this study, we employ the small-VGG CNN model which is derived from the famous VGGNet (Visual Geometry Group) [63]. Model architecture is unfolded in Table 6. The model is implemented with Python 3.6.5, Tensorflow 1.9.0, Keras 2.2.4, and trained employing stochastic gradient descent with a learning rate of 0.01, batch size 32, reducing the learning rate with factor 0.90, patience 20, minimum learning rate 0.0000001, and 1000 epochs. SVM-SMO experiments are run with tenfold cross-validation. RBM and sVGG experiments are run with 70% training and 30% test set.

## 5 Proposed STER Features

The human auditory system consists of many frequency bands between 20 and 22,050 Hertz. Many filter bank strategies have been proposed to cultivate this issue in speech processing applications. Mel scale and Bark scale are two prominent filter bank implementations that are successfully used so far. Generally, the human auditory system can be divided into two distinct parts in the Mel filter banks [64, 65]: a linearly spaced low-frequency region and a logarithmically spaced high-frequency region. The proposed Spectro-Temporal Energy Ratio (STER) features were designed to cultivate these issues along with the formant-based trapezoidal filter banks as shown in Figs. 2 and 3.

Most of the speech emotion recognition applications use the filter banks of speech processing technology. However, underlying dynamics as well as the characteristics of speech recognition and speech emotion recognition are quite different. In emotional speech, some voiced sounds may get lost as illustrated in Fig. 4. Therefore, designing a specific filter bank strategy solely for emotion recognition is compulsory. Another interesting point when dealing with the human auditory system is the fact that it is more sensitive to high-frequency components, especially to frequencies around

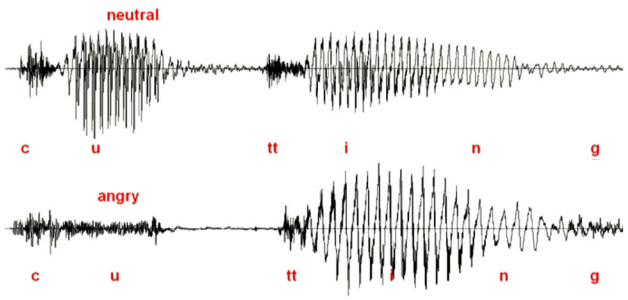



**Fig. 4** Dynamics of speech emotion recognition is quite different than speech recognition. Some voices may get lost in angry utterances (i.e., "u" of the cutting (IPA: kʌtɪŋ) in the angry speech wave signal)

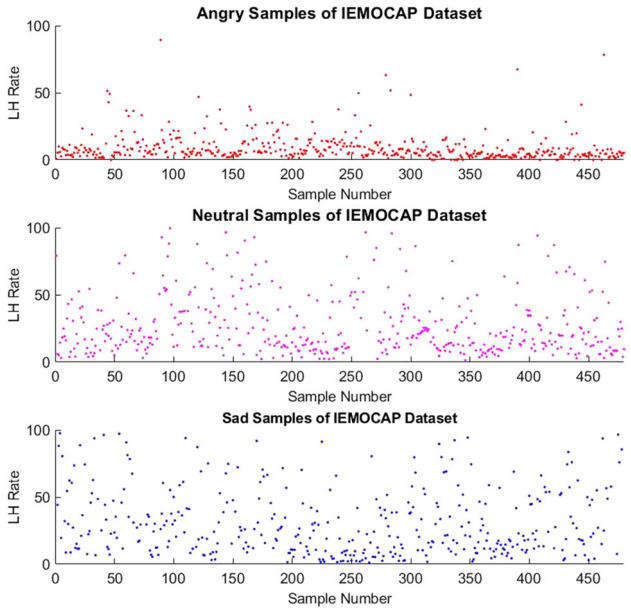



**Fig. 5** Scatter plots of LHRate for angry, neutral, and sad samples in the IEMOCAP dataset

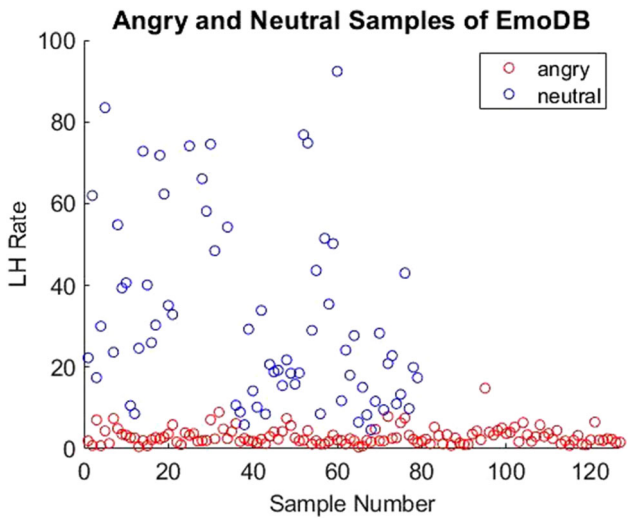



**Fig. 6** Scatter plot of LHRate for neutral and angry utterances in the EmoDB dataset

3000 Hz. A sound signal at 3000 Hz can easily suppress a 100-Hz signal even though its amplitude is 10 times weaker according to the equal loudness curves [66–69]. Another point to be dealt with is the boundary of linear and logarithmic regions which is generally considered as 1000 Hz. We discovered that the energy ratio (LHRate) of linear and logarithmic regions is very distinguishing for certain emotion types as shown in Fig. 5 and Fig. 6. In angry utterances, this ratio is always small but in neutral and sad speech samples, it goes to very high levels. We use ratios of these filter banks as new features without calculating the computationally expensive formants. In [70], authors claim that there are two kinds of sad speech: (1) sad, quiet, passive speech and (2) active grief of the type sometimes experienced in mourning. The first one is studied very extensively, and their findings also support our claims in terms of low $f_0$ and LHRate values.

Some studies have used formants and harmonics in speech emotion recognition (Ververides et al., [10]) but did not construct filter banks. Somewhat related Li et al. use the ARX-LF (AutoRegressive eXogenous with Lilijencrant-Fant) model to assess the cues of source and filter to evaluate arousal and valence states of emotional speech on Japanese vowels [71]. The novel filter bank strategy we are proposing differs from the Mel and Bark scales in terms of both geometrical shapes and frequency regions. Because the transition between formant regions of adjacent acoustic units (phones) is not fully sharp, it uses trapezoidal filter banks instead of triangular filters. There is a transition region between a phone and its acoustic neighbor. After this transition region, there is a flat line that points exactly to the related phone. Trapezoidal filter banks can be formulated as in Eq. (5). Here $n$ is the total number of filters and $f()$ holds the list of filter bank frequencies. We collect the magnitude of FFT under these regions using the value of the filter bank as weight factor for each coefficient.

$$H_n(k) = \begin{cases} 0 & k < f(n-1) \\ \frac{k-f(n-1)}{f(n)-f(n-1)} & f(n-1) \le k \le f(n) \\ 1 & f(n) \le k \le f(n+1) \\ \frac{f(n+1)-k}{f(n+1)-f(n)} & f(n+1) \le k \le f(n+2) \\ 0 & k > f(n+2) \end{cases} \quad (5)$$

As we can see from the scatter plots of Figs. 5 and 6, hot anger speech utterances (as opposed to cold anger ones, as discussed below) usually contain high $f_0$ values compared to the neutral and sad cases. In the proposed features, the maximum $f_0$ value is set to 700 Hertz. There are many algorithms for detecting fundamental frequency such as YAAPT [72], YIN [73], a recent convolutional neural network representation (CREPE) [74], an adaptive autocorrelation [75], wavelet-based [76], and various others can be found in [77–80].

Due to the great variations in human emotional expression, some researchers have defined two categories of anger: hot and cold anger. Broadly speaking, hot anger has high fundamental frequency values compared to neutral speech, as well as cold anger. Compared to cold anger, hot anger has a higher $f_0$, higher intensity, and stronger articulation [81, 82]. In our datasets, we analyzed anger with low and high $f_0$ and found out that the relationship between low- and high-frequency regions holds for hot anger. Further investigation is needed for the cold anger speech samples.

**Table 7** Classification results for EmoDB and EmoSTAR with Emobase set

| | EmoDB | EmoSTAR |
|---|---|---|
| SVM | **87.28** | 96.18 |
| RBM | 86.19 | 93.25 |
| sVGG | 81.98 | **97.45** |

**Table 8** Results of using 315 proposed STER features in EmoDB and EmoSTAR

| | EmoDB | EmoSTAR |
|---|---|---|
| SVM | **70.65** | **88.29** |
| RBM | 67.87 | 86.32 |
| sVGG | 65.21 | 87.28 |

**Table 9** Classification results of using Emobase + STER feature set

| | EmoDB | EmoSTAR |
|---|---|---|
| SVM | **88.22** | **96.69** |
| RBM | 86.62 | 95.55 |
| sVGG | 80.74 | 95.76 |

## 6 Results

In this section, we will introduce single-corpus and cross-corpus experiments on the EmoSTAR, EmoDB, IEMOCAP, and 5Mix datasets using Emobase, STER, and Emobase + STER feature sets. We also deal with the feature selection using Information Gain and Cfs Subset Evaluators of the Weka data mining tool. These experiments are conducted with SVM-SMO, RBM, and sVGG CNN models. These classifiers are the most powerful models of their eras; hence, this will give us an opportunity and insight to compare and evaluate their outputs.

### 6.1 Results of Single-Corpus Experiments

The results of the Emobase feature set, proposed STER (Spectro-Temporal Energy Ratios) features, and merging of Emobase and STER (Emobase + STER) are shown in Tables 7, 8 and 9, respectively. EmoDB results are for seven emotions and EmoSTAR results are for 4 emotion types. STER-based features have low accuracies compared to the Emobase feature set; however, they are performing extremely well on the cross-corpus experiments. In some experiments, STER-based features are surpassing the Emobase feature set very significantly despite the fewer number of features. Moreover, when they are merged with the Emobase set, they improve the performance very significantly in SVM, RBM, and sVGG classifiers most of the time as seen in Table 9. Although Emobase + STER feature set posts degrading performances with the sVGG model, this will change very shortly when we discuss the feature selection results in the next section. All results are in accuracy metric and the best scores are denoted in bold.

**Table 10** Classification results for the IEMOCAP and the 5mix datasets

| | SVM | RBM | sVGG |
|---|---|---|---|
| IEMOCAP-Emobase | 61.94 | 62.26 | 63.47 |
| IEMOCAP-Emobase + STER | 64.85 | 65.12 | **66.81** |
| IEMOCAP-STER | 63.33 | 59.00 | 58.29 |
| 5Mix-Emobase | 79.27 | 84.16 | 86.40 |
| 5Mix-Emobase + STER | 79.52 | 85.13 | **87.56** |
| 5Mix-STER | 68.95 | 68.72 | 75.95 |

We extended our work using IEMOCAP, ENTERFACE, RML, and SAVEE databases. Instead of using them individually, as discussed above in the Datasets section, we combined angry, happy, neutral, and sad samples of ENTERFACE, RML, SAVEE, EmoDB, and EmoSTAR to create a larger dataset called 5mix. In cross-corpus experiments, we use this dataset with IEMOCAP using 4 emotional classes of anger, happy, neutral, and sad. The IEMOCAP dataset has 3093 samples and the 5mix dataset has 2007 samples in 4 emotions.

Classification results can be seen in Table 10 for the IEMOCAP and the 5Mix datasets with SVM, RBM, and sVGG classifiers. We notice that the best accuracies have been achieved by the sVGG classifier with the Emobase + STER feature set. The sVGG classifier has a very significant improvement over SVM and is nearly 2% ahead of RBM in the IEMOCAP and the 5Mix datasets.

### 6.2 Feature Selection Results of Single-Corpus Experiments

Feature selection is applied with the Weka tool, and results are introduced in Table 11. The proposed features are selected by Information Gain and CFS Subset Evaluator with the highest information gain values. The Cfs Subset Evaluator demonstrated dazzling performances in terms of both accuracy and feature numbers in EmoSTAR and EmoDB, reducing the feature space very significantly while preserving the performance. Although we did not include experiments of the openSMILE Emo_large feature set, it would be beneficial to mention it in some rare cases where the best results have been achieved. In EmoSTAR, SVM reached 97.96% accuracy employing only 128 features selected by Cfs Subset

**Table 11** Results of feature selection in EmoDB and EmoSTAR by merging proposed STER features with the Emobase set

| | EmoDB | | EmoSTAR | |
|---|---|---|---|---|
| | Cfs (153) | IG (997) | Cfs (135) | IG (1132) |
| SVM | 87.10 | **89.71** | 96.94 | 96.94 |
| RBM | 86.56 | 86.50 | 97.35 | 96.24 |
| sVGG | 81.36 | 83.85 | **98.30** | 97.45 |

The number of selected features is shown in parentheses

Evaluator out of 6984 features of the Emo_large + STER feature set, surpassing all other classifications except the sVGG. This is more than a 98% decrease in the number of features while boosting the accuracy. Cfs Subset Evaluator selects more than 30 of the proposed features from the STER set in the experiments on the Emobase + STER feature set. In EmoDB, the SVM-SMO classifier performs the best result with 89.71% using 997 features selected by the Information Gain Attribute Evaluator (IG) out of 1303 features of the Emobase + STER feature set outperforming all works currently available.

On the other hand, sVGGNet posts an astonishing 98.30% accuracy on the EmoSTAR utilizing only 135 features from the Emobase + STER feature set. This is a very striking result in terms of both accuracy and feature number and suggests that Deep Networks can do exceptionally well even on small datasets with few features. This finding contrasts sharply with the general idea of the need for large datasets for deep neural networks.

## 7 Results of Cross-Corpus Experiments

After 20 years of work on machine emotion recognition, we are close to the limits of accuracy in single-corpus experiments, thanks to the improvements in artificial intelligence technologies. Current machine learning algorithms can demonstrate near-human-level performances on single-corpus works. But cross-corpus experiments pose a great challenge for researchers given the very low success rates compared to the single-corpus experiments in these applications. In single-corpus works, many environmental variables such as microphone, room acoustics, and speaker group are supposed to be constant. The results of cross-corpus experiments between EmoSTAR and EmoDB datasets are given in Tables 12, 13, and 14, respectively. As we can see, the results are not comparable with the results of single-corpus work, with great differences between the experimental results depending on the training and test sets. Feature selectors are selecting different features for each dataset. In cross-corpus experiments, we are using the entire feature set from both datasets for the integrity of cross-corpus experiments. We have run cross-corpus experiments using one or more databases as the training set and the others as the test set. In some studies, up to 8 databases [38] have been utilized for cross-corpus experiments. Since EmoSTAR has 4 emotion types, cross-corpus implementations were conducted with 4 emotions of EmoDB, namely, angry, neutral, happy, and sad. As can be seen from the tables, STER-based features always improve performance significantly. SVM, RBM, and sVGG classifiers perform better when trained on EmoSTAR. This may suggest the richness and diversity of EmoSTAR, in terms of feature representations. A nearly 10% increase was achieved in the SVM classifier when EmoSTAR was used as the training set with the Emobase + STER feature

**Table 12** Cross-corpus results with Emobase feature set on EmoSTAR and EmoDB datasets

| | Train On EmoSTAR - Test On EmoDB | Train On EmoDB - Test On EmoSTAR |
|---|---|---|
| SVM | 47.19 | **44.27** |
| RBM | **62.62** | 39.26 |
| sVGG | 61.65 | 33.84 |

**Table 13** Cross-corpus results using STER feature set on EmoSTAR and EmoDB datasets

| | Train On EmoSTAR - Test On EmoDB | Train On EmoDB - Test On EmoSTAR |
|---|---|---|
| SVM | **63.12** | 39.18 |
| RBM | 52.97 | **39.59** |
| sVGG | 48.08 | 38.67 |

**Table 14** Cross-corpus results of merging STER and Emobase feature sets on EmoSTAR and EmoDB datasets

| | Train On EmoSTAR - Test On EmoDB | Train On EmoDB - Test On EmoSTAR |
|---|---|---|
| SVM | 56.63 | **46.31** |
| RBM | **63.21** | 44.96 |
| sVGG | 51.32 | 37.65 |

**Table 15** Cross-corpus results of IEMOCAP and 5mix datasets using Emobase feature set

| | Train On IEMOCAP - Test On 5Mix | Train On 5Mix - Test On IEMOCAP |
|---|---|---|
| SVM | 32.38 | 28.25 |
| RBM | 36.29 | **37.19** |
| sVGG | **37.61** | 31.29 |

**Table 16** Cross-corpus results of IEMOCAP and 5mix datasets using STER feature set

| | Train On IEMOCAP - Test On 5Mix | Train On 5Mix - Test On IEMOCAP |
|---|---|---|
| SVM | 35.07 | 27.48 |
| RBM | 29.95 | **30.86** |
| sVGG | **35.22** | 21.75 |

**Table 17** Cross-corpus results of IEMOCAP and 5mix datasets merging STER and Emobase feature sets

| | Train On IEMOCAP - Test On 5Mix | Train On 5Mix - Test On IEMOCAP |
|---|---|---|
| SVM | 31.34 | 29.32 |
| RBM | **37.23** | 30.86 |
| sVGG | 33.43 | **32.13** |

set. Emobase + STER feature set increased the accuracy by more than 5% with the RBM classifier and more than 2% with the SVM classifier when EmoDB was used as the training set.

The STER-based feature set can generate a 63.12% accuracy with the SVM classifier. SVM demonstrates very poor performance in all cross-corpus experiments compared to the single-corpus experiments without the STER-based features. The performance of sVGG looks weaker in these cross-corpus experiments. We see that when EmoSTAR is used as the training set, the accuracy is over 60%, and when EmoDB is used as training set, the accuracy is over 40%. This strongly emphasizes the importance of training datasets in terms of feature representations.

In Tables 15, 16 and 17, we present the results of cross-corpus experiments with the 4-emotional (angry, happy, neutral, and sad) IEMOCAP and 5mix datasets. As we can see, the results are very poor compared to the cross-corpus experiments on EmoSTAR and EmoDB.

In this work, we have so far achieved very high accuracy rates. In the literature, the highest accuracy rate on the EmoDB dataset was 88.8% with 5967 features using SVM classifier with polynomial kernel and tenfold stratified cross-validation by Eyben et al. in [50]. We already outperformed this using our proposed features as mentioned in the previous section, reducing the feature space very significantly. In this part of the manuscript, we went further by increasing the MFCC coefficients from 12 to 24 to gain more accuracy in addition to the STER-based features. The total number of features reached 1760, and the accuracy rate was 87.10%. Applying the Information Gain Attribute Evaluator, we achieved an 88.97% accuracy rate with 1339 features. Next, we removed some features by careful manual elimination reducing the feature number down to 1220. By selecting 60% training data and 40% test data, we reached 90.65% accuracy with the SVM-SMO classifier which is the highest score currently available on the EmoDB dataset for 7 emotion types, as depicted in Table 18.

**Table 18** Feature selection results on EmoDB for 7 emotions after merging 24 MFCC and STER features

| | # of features | Accuracy |
|---|---|---|
| 24 MFCC + STER | 1760 | 87.10 |
| IG | 1339 | 88.97 |
| Expert | 1220 | **90.65** |

## 8 Conclusions

The most striking conclusion of this work is the exceptionally high accuracy rates of cross-corpus experiments that were obtained merging the STER-based feature set with the Emobase feature set. These features surpassed both the Emobase and Emo_large sets in cross-corpus experiments, proving the high integrity and robustness of the STER-based features against the cross-corporal variations. Novel features can close the gap between cross-corpus and single-corpus experiments, helping to increase the accuracy. Another major conclusion of this paper is the moderate recognition rates of cross-corpus results compared to the single-corpus results when RBM and sVGG are employed. The performance of the Emo_large feature set that comprises 6669 features is not superior to the Emobase set with 988 features and therefore is not included in this manuscript except for some rare cases. In single-corpus tests, it is below the performance of the Emobase set in the EmoDB dataset. All the best results of this study are obtained by merging the Emobase feature set and the proposed STER feature set. In the EmoDB dataset SVM posts the best result. In EmoSTAR, IEMOCAP, and the 5Mix datasets, sVGG CNN model has the highest accuracies.

Another promising conclusion is the extremely superior performance of the STER-based features when they are used

with feature selectors. Feature selectors surpass the original full feature set regardless of their enormous number of features. In the EmoDB and EmoSTAR datasets, we can surpass the success rate of the Emo_large configuration using only 128 features. The Cfs Subset Evaluator feature selector shrinks the size of the feature space from 6984 down to 128 which means more than 98% reduction while increasing the accuracy. Although RBM demonstrated slightly weaker performance in some of the single-corpus experiments, it is superior in cross-corpus experiments against SVM, particularly when trained on the EmoSTAR dataset. In single-corpus experiments, STER-based features increase the performance by nearly 3% with sVGG classifier on the IEMOCAP dataset. On the other hand, in cross-corpus experiments, STER-based features improve the performance of the SVM classifier by approximately 15% when trained on EmoSTAR, and the performance of the RBM classifier by up to 5% when trained on EmoDB.

One of the interesting results is that increasing the depth or width of the RBM classifier does not improve performance. A single-layer RBM with fewer hidden nodes can surpass the accuracy of 2, 3, and even 4-layer RBM classifier. Better results are obtained by tuning the hyperparameters properly. Shallower and narrower networks can perform better than wider and deeper networks significantly.

In future work, we will develop EmoSTAR for more emotional speech samples and more emotion types. The STER-based filter banks may need to be improved particularly in the high-frequency range. Too many filters as well as too few filters may lead to overfitting or underfitting.

Today emotion recognition progresses by combining signal features with the features of speech recognition, facial expression, and body language in multimodal works [83]. We are posed to enlarge EmoSTAR as a multimodal database including the aforementioned features for real-life scenarios of emotion recognition to reach our goal of an optimal set of features.

**Author Contributions** CP: Conceptualization, Methodology, Writing, Experiments, BD: Supervising, Reviewing, Writing, Editing, Conceptualization, Validation, and Data Preparation, YA: Supervising, Reviewing, Writing, Editing, Validation.

**Funding** This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Availability of Data and Materials** Data is available on the following site: https://github.com/cevparlak/speech-emotion-recognition.

**Code Availability** The source code is available on the following site: https://github.com/cevparlak/speech-emotion-recognition.

## Declarations

**Conflict of interest** There is no conflict of interest between the authors.

**Human and Animal Rights** This article does not contain any studies with human participants or animals performed by any of the authors.

**Consent to Participate and for Publication** We confirm that we did not use participants in our study.

## References

1. Herculano-Houzel, S.: The human brain in numbers: a linearly scaled-up primate brain. Front. Hum. Neurosci. (2009). https://doi.org/10.3389/neuro.09.031.2009
2. Nguyen, T.: Total number of synapses in the adult human neocortex. Undergrad. J. Math. Model. One+ Two **3**(1), 26 (2010). https://doi.org/10.5038/2326-3652.3.1.26
3. Ekman, P.E.; Davidson, R.J.: The Nature of Emotion: Fundamental Questions. Oxford University Press (1994)
4. Plutchik, R.: The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. **89**(4), 344–350 (2001)
5. Whissel, C.M.: The Dictionary of Affect in Language: Emotion: Theory, Research and Experience. Academic Press, New York (1989) https://doi.org/10.1016/B978-0-12-558704-4.50011-6
6. Cowie, R.; Cowie, E.D.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G.: Emotion recognition in human-computer interaction. Signal Process. Mag. IEEE **18**(1), 32–80 (2001). https://doi.org/10.1109/79.911197
7. Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D.: Recognizing realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9–10), 1062–1087 (2011). https://doi.org/10.1016/j.specom.2011.01.011
8. Wang, W.: Machine Audition: Principles, Algorithms, and Systems, p. 1–554. IGI Global (2011)
9. Wu, S.; Falk, T.H.; Chan, W.: Automatic speech emotion recognition using modulation spectral features. Speech Commun. **53**(5), 768–785 (2010). https://doi.org/10.1016/j.specom.2010.08.013
10. Ververidis, D.; Kotropoulos, C.: Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006). https://doi.org/10.1016/j.specom.2006.04.003
11. Ramakrishnan, S.: Recognition of emotion from speech: a review. Int. J. Speech Technol. **15**(2), 99–117 (2012)
12. He, L.: Stress and Emotion Recognition in Natural Speech in the Work and Family Environments, PhD Thesis, RMIT University (2010)
13. Neumann, M.: Cross-lingual and multilingual speech emotion recognition on English and French. In: 2018 International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5769–5773. Calgary, AB, Canada IEEE (2018). https://doi.org/10.1109/ICASSP.2018.8462162
14. Zhang, B.; Provost, E.M.; Essl, G.: Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences. IEEE Trans. Affect. Comput. **10**(1), 85–99 (2019). https://doi.org/10.1109/TAFFC.2017.2684799
15. Song, P.: Transfer linear subspace learning for cross-corpus speech emotion recognition. IEEE Trans. Aff. Comput. **10**(2), 265–275 (2017). https://doi.org/10.1109/TAFFC.2017.2705696
16. Shah, M.; Chakrabarti, C.; Spanias, A.: Within and cross-corpus speech emotion recognition using latent topic model-based features. EURASIP J. Audio Speech Music Process. (2015). https://doi.org/10.1186/s13636-014-0049-y
17. Song, P., et al.: Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. Speech Commun. **83**, 34–41 (2016). https://doi.org/10.1016/j.specom.2016.07.010

18. Wang, K., et al.: Speech emotion recognition using Fourier parameters. IEEE Trans. Affect. Comput. **6**(1), 69–75 (2015). https://doi.org/10.1109/TAFFC.2015.2392101
19. Yenigalla, P.; et al.: Speech emotion recognition using spectrogram & phoneme embedding. In: Interspeech 2018, September, pp. 3688–3692. Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-1811
20. Mao, Q., et al.: Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. Speech Commun. **93**, 1–10 (2017). https://doi.org/10.1016/j.specom.2017.06.006
21. Kamińska, D.; Sapiński, T.; Anbarjafari, G.: Efficiency of chosen speech descriptors in relation to emotion recognition. EURASIP J. Audio Speech Music Process. (2017). https://doi.org/10.1186/s13636-017-0100-x
22. Seng, K.P.; Li-Minn, A.; Ooi, C.S.: A combined rule-based & machine learning audio-visual emotion recognition approach. IEEE Trans. Affect. Comput. **9**(1), 3–13 (2016). https://doi.org/10.1109/TAFFC.2016.2588488
23. Phan, D.A.; Matsumoto, Y.; Shindo, H.: Autoencoder for semisupervised multiple emotion detection of conversation transcripts. IEEE Trans. Affect. Comput. **12**(3), 682–691 (2018). https://doi.org/10.1109/TAFFC.2018.2885304
24. Deng, J., et al.: Semisupervised autoencoders for speech emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 31–43 (2017). https://doi.org/10.1109/TASLP.2017.2759338
25. Sahu, S.; Gupta, R.; Espy-Wilson, C.: On enhancing speech emotion recognition using generative adversarial networks. Preprint https://arxiv.org/abs/1806.06626 (2018). https://doi.org/10.48550/arXiv.1806.06626
26. Oflazoglu, C.; Yildirim, S.: Recognizing emotion from Turkish speech using acoustic features. EURASIP J. Audio Speech Music Process. (2013). https://doi.org/10.1186/1687-4722-2013-26
27. Kaya, H.; et al.: LSTM Based cross-corpus and cross-task acoustic emotion recognition. In: Interspeech 2018, September, pp. 521–525, Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-2298
28. Rouast, P.V.; Adam, M.; Chiong, R.: Deep learning for human affect recognition: insights and new developments. IEEE Trans. Affect. Comput. **12**(2), 524–543 (2019). https://doi.org/10.1109/TAFFC.2018.2890471
29. Cho, J.; et al.: Deep neural networks for emotion recognition combining audio and transcripts. In: Interspeech 2018, September, pp. 247–251, Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-2466
30. Kim, J.; Saurus, R.A.: Emotion recognition from human speech using temporal information and deep learning. In: Interspeech 2018, September, pp. 937–940, Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-1132
31. Ma, X.; et al.: Emotion recognition from variable-length speech segments using deep learning on spectrograms. In: Interspeech 2018, September, pp. 3683–3687, Hyderabad, India (2018). https://doi.org/10.21437/Interspeech.2018-2228
32. Trigeorgis, G.; et al.: Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: 2016 International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016 March, pp. 5200–5204, Shanghai, China. IEEE (2016). https://doi.org/10.1109/ICASSP.2016.7472669
33. Tzirakis, P.; Zhang, J.; Schuller, B.W.: End-to-end speech emotion recognition using deep neural networks. In: 2018 International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, April, pp. 5089–5093, Calgary, AB, Canada. IEEE (2018). https://doi.org/10.1109/ICASSP.2018.8462677
34. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
35. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M.: The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**(1), 5–17 (2012). https://doi.org/10.1109/T-AFFC.2011.20
36. Jeon, J.H.; Le, D.; Xia, R.; Liu, Y.: (2013). A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception. In: Interspeech 2013, August, pp. 2837–2840 Lyon, France (2013). https://doi.org/10.21437/Interspeech.2013-246
37. Eyben, F.; Batliner, A.; Schuller, B.; Seppi, D.; Steidl, S.: Cross-Corpus classification of realistic emotions–some pilot experiments. In: Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, pp. 77–82Valetta, Malta (2010)
38. Schuller, B.; Zhang, Z.; Weninger, F.; Rigoll, G.: Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In: Proc. Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel (2011)
39. Shami, M.; Verhelst, W.: An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun. **49**(3), 201–212 (2007). https://doi.org/10.1016/j.specom.2007.01.006
40. Wen, G., et al.: Random deep belief networks for recognizing emotions from speech signals. Comput. Intell. Neurosci. (2017). https://doi.org/10.1155/2017/1945630
41. Kahou, S.E., et al.: Emonets: Multimodal deep learning approaches for emotion recognition in video. J. Multimodal User Interfaces **10**, 99–111 (2016). https://doi.org/10.1007/s12193-015-0195-2
42. Hassan, M.M., et al.: Human emotion recognition using deep belief network architecture. Inf. Fusion **51**, 10–18 (2019). https://doi.org/10.1016/j.inffus.2018.10.009
43. Atmaja, B.T.; Akagi, M.: Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM. Speech Commun. **126**, 9–21 (2021). https://doi.org/10.1016/j.specom.2020.11.003
44. Firdaus, M.; Chauhan, H.; Ekbal, A.; Bhattacharyya, P.: MEISD: A Multimodal Multi-Label Emotion, Intensity, and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations. In: Proceedings of the 28th International Conference on Computational Linguistics, 2020, December, pp. 4441–4453, Barcelona, Spain (2020). https://doi.org/10.18653/v1/2020.coling-main.393
45. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. Preprint https://arxiv.org/abs/1810.02508 (2018). https://doi.org/10.48550/arXiv.1810.02508
46. Yin, Y.; Zheng, X.; Hu, B.; Zhang, Y.; Cui, X.: EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. Appl. Soft Comput. **100**, 106954 (2021). https://doi.org/10.1016/j.asoc.2020.106954
47. Pakyurek, M.; Atmis, M.; Kulac, S.; Uludag, U.: Extraction of novel features based on histograms of MFCCs used in emotion classification from generated original speech dataset. Elektronika ir Elektrotechnika **26**(1), 46–51 (2020). https://doi.org/10.5755/j01.eie.26.1.25309
48. Parlak, C.; Diri, B.; Gürgen, F.: A Cross-Corpus Experiment in Speech Emotion Recognition. In: Proc. International Workshop on Speech, Language and Audio in Multimedia (SLAM 2014), pp. 58–61, Penang, Malaysia, (2014)

49. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B.: A database of German emotional speech. In: Interspeech 2005, September, pp. 1517–1520 Lisbon, Portugal (2005). https://doi.org/10.21437/Interspeech.2005-446
50. Eyben, F.; Wollmer, M.; Schuller, B.: OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops ACII 1–6, 2009, September, Amsterdam, Netherlands. IEEE (2009). https://doi.org/10.1109/ACII.2009.5349350
51. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I.: The eNTERFACE'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06) 8–8, April 2006, Atlanta, GA, USA. IEEE (2006). https://doi.org/10.1109/ICDEW.2006.145
52. Wang, Y.; Guan, L.: Recognizing human emotional state from audiovisual signals. IEEE Trans. Multimedia **10**(5), 936–946 (2008). https://doi.org/10.1109/TMM.2008.927665
53. Haq, S.; Jackson, P.J.B.: Multimodal Emotion Recognition, In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems. IGI Global Press, ISBN 978–1615209194 Chapter 17, pp. 398–423 (2010). https://doi.org/10.4018/978-1-61520-919-4.ch017
54. Eyben, F.; Wöllmer, M.; Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. ACM Multimedia (MM), ACM, ISBN 978-1-60558-933-6, pp. 1459–1462, Florence, Italy, (2009). https://doi.org/10.1145/1873951.1874246
55. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. Newsl. **11**(1), 10–18 (2009). https://doi.org/10.1145/1656274.1656278
56. Smolensky, P.: Chapter 6: Information processing in dynamical systems: foundations of harmony theory. Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, MIT Press, pp. 194–281 (1987)
57. Hinton, G.E.; Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006). https://doi.org/10.1126/science.112764
58. Salakhutdinov, R.: Learning deep generative models. Annu. Rev. Stat. Appl. **2**, 361–385 (2015). https://doi.org/10.1146/annurev-statistics-010814-020120
59. Hinton, G.E.: A practical guide to training restricted Boltzmann machines. Neural networks: Tricks of the trade **7700**, 599–619 Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35289-8_32
60. Krizhevsky, A.; Hinton, G.E.: Learning multiple layers of features from tiny images **1**(4) Technical report, University of Toronto (2009)
61. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8), 1771–1800 (2002). https://doi.org/10.1162/089976602760128018
62. Tanaka, S.M.; Okutomi, M.: A novel inference of a restricted Boltzmann machine. In: 22nd International Conference on Pattern Recognition, 2014, August, pp. 1526–1531, Stockholm, Sweden IEEE (2014). https://doi.org/10.1109/ICPR.2014.271
63. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint https://arxiv.org/abs/1409.1556 (2014). https://doi.org/10.48550/arXiv.1409.1556
64. Davis, S.; Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
65. Skowronski, M.D.; Harris, J.G.: Improving the filter bank of a classic speech feature extraction algorithm. In: Proceedings of the 2003 International Symposium on Circuits and Systems ISCAS'03, 2003, May, IV-IV, Bangkok, Thailand. IEEE (2003). https://doi.org/10.1109/ISCAS.2003.1205828
66. Fletcher, H.; Munson, W.A.: Loudness, its definition, measurement and calculation. Bell Syst. Tech. J. **12**(4), 377–430 (1933). https://doi.org/10.1002/j.1538-7305.1933.tb00403.x
67. Robinson, D.W.; Dadson, R.S.: A re-determination of the equal-loudness relations for pure tones. Br. J. Appl. Phys. **7**(5), 166 (1956). https://doi.org/10.1088/0508-3443/7/5/302
68. Suzuki, Y.; Mellert, V.; Richter, U.; Møller, H.; Nielsen, L.; Hellman, R.; Takeshima, H.: Precise and full-range determination of two-dimensional equal loudness contours. Tohoku University, Japan (2003)
69. Suzuki, Y.; Takeshima, H.: Equal-loudness-level contours for pure tones. The J. Acoust. Soc. Am. **116**(2), 918–933 (2004). https://doi.org/10.1121/1.1763601
70. Erickson, D.; Yoshida, K.; Menezes, C.; Fujino, A.; Mochida, T.; Shibuya, Y.: Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica **63**(1), 1–25 (2006). https://doi.org/10.1159/000091404
71. Li, Y.; Li, J.; Akagi, M.: Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space. J. Acoust. Soc. Am. **144**(2), 908 (2018). https://doi.org/10.1121/1.5051323
72. Zahorian, S. A.; Dikshit, P.; Hu, H.: A spectral-temporal method for pitch tracking. In: Ninth International Conference on Spoken Language Processing, 2006, September, paper 1910-Wed2A1O.5. Pittsburgh, Pennsylvania, USA (2006). https://doi.org/10.21437/Interspeech.2006-475
73. De Cheveigné, A.; Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002). https://doi.org/10.1121/1.1458024
74. Kim, J.W.; et al.: Crepe: A convolutional representation for pitch estimation. In: 2018 International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 161–165. Calgary, AB, Canada, IEEE (2018). https://doi.org/10.1109/ICASSP.2018.8461329
75. Staudacher, M., et al.: Fast fundamental frequency determination via adaptive autocorrelation. EURASIP J. Audio Speech Music Process. (2016). https://doi.org/10.1186/s13636-016-0095-8
76. Goh, Y.H.; et al.: Fast Wavelet-based Pitch Period Detector for Speech Signals. In: 2016 International Conference on Computer Engineering and Information Systems, 2016, November, pp. 494–497, Shanghai, China. Atlantis Press (2016). https://doi.org/10.2991/ceis-16.2016.101
77. Stone, S.; Steiner, P.; Birkholz, P.: A time-warping pitch tracking algorithm considering fast f0 changes. In: Interspeech 2017, August, pp. 419–423 Stockholm, Sweden (2017). https://doi.org/10.21437/Interspeech.2017-382
78. Aneeja, G.; Yegnanarayana, B.: Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 829–838 (2017). https://doi.org/10.1109/TASLP.2017.2666425
79. Ardaillon, L., & Roebel, A.: Fully-convolutional network for pitch estimation of speech signals. In: Interspeech 2019, September, pp. 2005–2009, Graz, Austria, (2019). https://doi.org/10.21437/Interspeech.2019-2815
80. Wang, D.; Yu, C.; Hansen, J.H.L.: Robust harmonic features for classification-based pitch estimation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(5), 952–964 (2017). https://doi.org/10.1109/TASLP.2017.2667879
81. Kim, J.; Erickson, D.; Lee, S.; Narayanan, S.: A study of invariant properties and variation patterns in the converter/distributor model for emotional speech. In: Interspeech 2014, September, pp. 413–417 Singapore (2014). https://doi.org/10.21437/Interspeech.2014-95

82. Whiteside, S.P.: Simulated emotions: an acoustic study of voice and perturbation measures. In: Fifth International Conference from Spoken Language Processing (ICSLP 1998), November, paper 0153, Sydney Convention Centre, Sydney, Australia (1998). https://doi.org/10.21437/ICSLP.1998-141
83. Gunes, H.; Piccardi, M.; Pantic, M.: From the lab to the real world: Affect recognition using multiple cues and modalities. InTech Education and Publishing, pp. 185–218 (2008). https://doi.org/10.5772/6180

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.