Neurocomputing 528 (2023) 1–11

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An ongoing review of speech emotion recognition

Javier de Lope [a,*], Manuel Graña [b]

[a] Department of Artificial Intelligence, Universidad Politécnica de Madrid (UPM), Spain
[b] Computational Intelligence Group, University of the Basque Country (UPV/EHU), Spain

ARTICLE INFO

Article history:
Received 7 August 2022
Revised 5 December 2022
Accepted 8 January 2023
Available online 18 January 2023

Keywords:
Speech emotion recognition
Deep learning
Datasets
Machine learning
Speech feature extraction

ABSTRACT

User emotional status recognition is becoming a key feature in advanced Human Computer Interfaces (HCI). A key source of emotional information is the spoken expression, which may be part of the interaction between the human and the machine. Speech emotion recognition (SER) is a very active area of research that involves the application of current machine learning and neural networks tools. This ongoing review covers recent and classical approaches to SER reported in the literature.

1. Introduction

Inside the broad field of Automated Emotion Recognition (AER) [83], Speech Emotion Recognition (SER) may be considered a branch of Automatic Speech Recognition (ASR) [30,101,142,152] exploiting the same kind of signal, feature extraction processes, and potential application of diverse machine learning techniques, such as deep learning (DL) architectures, that are also applied in the field of Natural Language Processing (NLP) [61] that shares with SER the sequential nature of the data. Works already developed for ASR on feature extraction, such as the exploitation of Mel Frequency Cepstral Coefficients (MFCC) [27] for classification and pattern recognition, have been seamlessly transferred into SER research. Also, DL approaches developed in ASR or NLP, like several flavors of convolutional neural networks (CNN) and recurrent neural networks (RNN),are evaluated for SER practical applications. In the future, SER is called to be a key technology in the development of innovative Human Computer Interaction (HCI), Human Machine Interaction (HMI), Human Robot Interaction (HRI) systems [129], and affective computing in general [146] which are becoming more important in the upcoming era of the Internet of Things (IoT) when pervasive ambient intelligence will be in a permanent dialog with the human user. Other kind of applications of SER are in the psychological domain, such as the detection of emotional valences in speech for automated depression detection [77].

Most common data sources for AER are physiological data from wearable sensors, electroencephalographic data (EEG) [20], facial imaging, and speech based data. Though some researchers work on multimodal data fusion [146], most current research is focused on a single data modality. The fusion of imaging and voice data is the multimodal approach most often found in the literature.

The increasing activity on SER research is reflected in recent reviews [5,4,14,39,65,121,122,134,147], some of them with special focus on DL approaches [1,74]. This paper contributes some updated information relative to these recent reviews, which are usually structured along three axes: the available data repositories, the feature extraction processes, and the classification and pattern recognition approaches. However, DL approaches usually encompass both the feature extraction and classification phases. We follow this typical structure in this paper. The structure of the paper is as follows: Section 2 reviews the most cited data repositories for SER. Section 3 collects conventional machine learning approaches for SER. Section 4 collects recent DL approaches for SER. We provide a table with the spelling of the most used acronyms in Appendix A.

2. Audio-visual databases

Some general works on emotion recognition provide an exhaustive enumeration of existing dataset for SER. For instance, Pitterman et al. [105] enumerate and briefly describe over 100 emotion databases, while Swain et al. [134] describes in depth about 60 emotion databases. Additionally, Douglas-Cowie et al. [33] offer some guidelines on how to compile appropriate datasets

[*] Corresponding author.

https://doi.org/10.1016/j.neucom.2023.01.002
0925-2312/© 2023 The Author(s). Published by Elsevier B.V.

for SER according to their experience in similar projects. Dataset construction needs to takeinto account factors such as the recording prompts, the speaker selection procedure, the recording setup and the quality control [15]. In this section, we briefly describe the most cited datasets, which are summarized in Table 1. Table 2 summarizes the reference works that use each dataset.

DES
(Danish Emotional Speech) [36,37] is a Danish emotional database funded by the European Union. It contains recordings from 4 professional actors (2 male and 2 female and comprises about 30 min of speech. The recordings consist of two isolated words ("yes" and "no"), nine short phrases (four of which are questions) and two passages. Each utterance was spoken emulating each of the following five emotional states: neutral, surprise, happiness, sadness, and anger. It has been one of the most used databases for researchers in this field. Its application range from testing SVM approaches. alone [23] or in combination with hidden markov models [76]. It has been also applied to gender based emotion recognition [67,139], and to novel deep learning approaches [53] and cross-corpus validation [62].

EMODB
[18] is a German database with high-quality audio recordings from 10 actors (5 male and 5 female) whom produce 10 German utterances (5 short and 5 longer sentences) with 7 emotions (anger, neutral, anger, boredom, happiness, sadness, disgust). Some emotional expressions have two versions recorded by the same author. Thus, the database provides about 800 sentences. Everyday sentences were used in order to achieve a more natural form. Moreover, actors can speak them from memory without need of memorizing or reading them off a paper. Two examples of sentences are "Tonight I could tell him" (short) and "What are the bags standing there under the table?" (long). It is one of the most commonly referred in the technical literature. The works include bioinspired real time speech emotion recognition [81], conventional machine learning classifiers that used features as MFCC (Mel Frequency Cepstral Coefficients) [28,41,51,54,93,100,141,140,158,86,167], LPCC (Linear Prediction Cepstral Coefficient) [40,63,64,67,98,100,116,123,133] and other features [10,11,21,60,62,82,110,114,125,132,144,149,154,161,165,137] as well as deep learning based approaches [9,12,53,59,72,75,91,113,124,130,151,147,45,13,102,119].

eNTERFACE 89
is an audio-visual emotion database. It contains six emotions (happiness, sadness, surprise, anger, disgust and fear) expressed by 42 subjects from 14 different nationalities (81% male and 19% female) as reaction to six successive short stories, each of them eliciting a particular emotion. All the experiments were driven in English. In order to avoid ambiguities, two human experts discarded the samples in which the emotion was not well expressed. It has been used for testing conventional machine learning approaches [62,98,125,25] as well as deep learning models [145].

IEMOCAP
(Interactive Emotional Dyadic Motion Capture Database) [19] is a database collected by the Speech Analysis and Interpretation Laboratory at the University of South California. It was recorded from ten actors in dyadic sessions with markers on face, head and hands during scripted and improvised scenarios. It contains approximately 12 h of data. The audio which is split into segments between 3–15 s and labeled by 3–4 human evaluators. Although

Table 1
Most cited SER databases. #em = number of emotions, AV = audio-visual data, Nat. = natural versus actors recording, #ms #fs = number of male and female subjects, #ut = number of utterances, set = number of sentences.

Database	#em	language	AV	Nat.	#ms	#fs	#ut	#sent
DES [36,37]	5	Danish	n	actors	2	2	13	
EMODB [18]	7	German	n	actors	5	5	10	800
eNTERFACE [89]	6	English	y	natural	34	8	-	-
IEMOCAP [19]	10	English	y	actors	5	5	-	-
SAVEE [47]	7	English	y	actors	4	-	120	480
Thai DB [128]	6	Thai	y	actors	3	3	972	5832
INTER1SP [87]		Spanish	n	actors	1	1	184	6040
TESS [34]	7	English	n	actors	-	2	200	2800
RAVDESS [79]	8	English	y	actors	12	12	2	1440
JL-Corpus [57]	10	NZ English	n	actors	4	0	-	-
MSP-PODCAST [80]	8	English	n	natural	-	-	-	18000

Table 2
Publications per database

Database	reference work
DES [36,37]	[53,76,23,110,62,139]
EMODB [18]	[12,59,60,100,110,40,161,141,140,149,82,132,11,51,63,137,167,86,133,98,123,81,154,165,56,10,54,158,93,125,21,41,116,72,124,144,64,28,62,9,53,91,113,75,130,151,147,102,13,45,119]
eNTERFACE [89]	[62,125,98,63,25,145]
IEMOCAP [19]	[13,164,137,95,56,88,28,26,51,138,8,9,70,91,151,163,50,143,73,45,96,32,166,17,102]
SAVEE [47]	[53,72,66,55,100,25,28,143,92,84]
Thai DB [128]	[116]
INTER1SP [87]	[38,64]
TESS [34]	[107,124,68,43]
RAVDESS [79]	[124,55,40,126,68,157,155,118,90,56,104,52,106,156,92,117,7,6,94,43,131,84]
JL-Corpus [57]	no Refs.
MSP-PODCAST [80]	[8]
VAM [44]	[140]

the database was initially designed to target anger, sadness, happiness, frustration and neural state, it was expanded to 10 categories (the basic emotions according psychologists [35] anger, sadness, happiness, disgust, fear, and surprise, plus frustration, excited, neutral and other). It is a well-known database that has been studied with several classic machine learning techniques [26,28,51,88,95,137]. Recently it has been also used to experiment with modern convolutional neural networks models [9,91,163,164,96,70,73,17,166,151,45,13,102,50,143] and as part of multimodal studies [32].

SAVEE
(Survey Audio-Visual Expressed Emotion) [47] contains video and audio recordings of 4 English male actors in 7 emotions (neutral, anger, disgust, fear, happiness, sadness and surprise). Each actor played 120 utterances, which makes 480 sentences in total. For the visual features, 60 markers were painted on the actors' faces. The recordings consist of 15 phonetically-balanced sentences per emotion (3 common sentences, 2 emotion specific sentences and 10 generic sentences). It has been used to test wavelet based feature extraction methods [66,100], deep learning approaches [53,84], gradient boosting classifiers [55], and multisource information fusion [72,25,92,143].

Thai DB
(Audiovisual Thai Emotion Database) [128] is composed by recordings of six basic emotions: happiness, sadness, surprise, anger, fear and disgust. The utterances were spoken by six drama-students. They were asked to read 1,000 most commonly used Thai words with one to seven syllables. Recordings where human listeners were not able to recognize the emotion were deleted. The final list has 972 words. Frequently it is referred as one of the easier datasets to achieve good performances upon, because of it only contains isolated words rather than longer passages as other databases. However, the use of isolated words is a common practice when compiling this kind of databases. It has been used to test conventional machine learning techniques, such as SVM, to classify the emotions [116].

INTER1SP
[87] is composed by recordings of one male and one female Spanish professional actors. It contains 184 utterances that include isolated words, digits and sentences. It is about 4 h of data from each speaker and contains 6,040 samples. The emotions anger, disgust, fear, happiness, sadness, surprise and neutral are considered. It was created within the scope of an European project. It is used as dataset for both conventional machine learning [63,64] and deep learning [38] approaches.

TESS
(Toronto Emotional Speech Set) [34] includes about 2,800 samples of 200 isolated words by two professional actresses. Seven emotions are considered: anger, disgust, fear, happiness, pleasant surprise, sadness and neutral. It is frequently used for testing a kind of deep learning models [124,107,43] although it is also used with conventional machine learning methods [68].

RAVDESS
(Ryerson Audio-Visual Database of Emotional Speech and Song) [79] is composed by 7356 audio and video clips (roughly 25 GB). It provides samples of speech and songs, thus it also allows to train models for the analysis of musical recordings. It contains 1440 samples of speech audio recordings, which are recorded by 24 professional actors (12 male and 12 female) that read two semantically neutral US English phrases while revealing eight emotions (neutral, calm, happiness, sadness, anger, fear, disgust, surprise). The phrases are "Kids are talking by the door" and "Dogs are sitting by the door" and they were selected according the length in syllables and word frequency and familiarity. Two levels of emotional intensity are considered in each phrase (only one intensity in the neutral emotional state) and there is just one sample for each recording. The audio files are provided as lossless wave format at 48 kHz to avoid the usual artifacts and other alterations of the original signal for the compression. Each one is about 3 s long. As it is one of the most complete databases, it has been extensively used to test conventional machine learning approaches, including SVM, gradient boosting and the hybridization of feature extraction methods base on wavelets and spectral features [40,55,90,118,155,157,68,126] as well as shallow neural networks [124], and deep learning approaches including the classical CNNs and LSTMs[56,104,52,58,106,156,7,6,94,43,131,84,117,92].

JL-Corpus
[58] is a database specifically adapted for the New Zealand English that include 5 primary and 5 secondary emotions. The latter are important in the Human-Robot Interaction (HRI) domain and they are not usually covered by other databases. Besides the original proposal, we found no references of its exploitation to test machine learning approaches.

MSP-PODCAST
[87] is an approach to effectively build a large, naturalistic emotional database with balanced emotional content. It relies on existing spontaneous recordings obtained from audio-sharing websites published under public licenses. It provides more than 18,000 natural emotional sentences, over 27 h, from multiple speakers with audio segments between 2.75–11 s. Then, a group of almost 300 evaluators annotated the samples with emotional attributes (arousal, valence, dominance) and an extended list of categorical emotions. Only the labeling with emotional attributes is balanced. It has been used to test a listener dependent approach to apply deep learning architectures to SER [8].

3. Conventional machine learning

Table 3 contains the map of the found publications relative to the feature extraction process and the classification method.It is undeniable that SVM are the most cited conventional machine learning approach in the SER domain. Gao et al. [40] use a linear kernel SVM with sequential minimal optimization emotions classification. They compute pitch, intensity, MFCC (mel frequency cepstral coefficients), LSP (line spectral pairs) and ZCR (zero crossing rate) as local features of windows in the range from 20–100 ms on the audio files. They also use a depth first search to obtain values for the duration and overlapping of windows. Then, they apply a smoothing and normalization and obtain some statistics of all speech local features extracted from each utterance, which are used as global features. Dahake et al. [27] compare the use of several kernel functions in SVM to classify emotions utterances in an in-housedatabase. They use MFCC and energy and pitch based features. The study is carried out by classifying each emotion with a different kernel, so the results can not be easily generalized. However, the quadratic kernels appear to outperform other alternatives. Milton et al. [93] apply SVM classifiers with linear and RBF kernels and MFCC features to recognize emotions. They improve the state-of-art with their proposal of using a 3-stage SVM. Sinith et al. [123] use SVM and a combination of MFCC, pitch and energy based features to classify speech emotions. Yang et al. [150] propose the use of Twins SVM and compare its performance against standard SVM for speech emotion recognition. They use a set of featured computed in time- and frequency-domains. MFCC are also

Table 3
Conventional machine learning approaches found in the literature. VK SVM = various kernels for the SVM, RSVM = radial basis function kernel SVM, TSVM = twins SVM, LR = logistic regression, MLP = multilayer perceptron, HuM = Hu moments, ELMDT = extreme learning machine decision tree, BN = Bayes networks, GMM = Gaussian mixture model, EDT = Ensembles of decision trees, kNN = k nearest neighbors, NNMF = non negative matrix factorization.

Feat. ext.	Classifiers									
	LSVM	VSVM	RSVM	TSVM	LR	MLP	ELMDT	BN	GMM	EDT	kNN
MFCC	[40,93,88]	[63,108,46,133,23,27,123,24, 41,116,2,161,90,161,160,109]	[98,93]	[150]	[63,155,167]	[24,54,109,55,22,28]			[97,158,160,127,66]	[41,112,54]	[161,25,2,160,16,111,140]
delta MFCC		[132]			[167]						
LPCC		[132,116]	[100]							[31]	
energy	[40]	[27,108,116]				[24]					
LSP	[40]										
ZCR	[40]	[41,90,116]								[41]	
pitch		[27,108]				[28]			[97]		
arbitrary/ optimized features		[3,149,23,64,10,114,148,95,137,110]	[165]	[150]	[72]	[3,22,165]	[62]	[3]	[11,115,148,82]	[165,60,148]	[165,81]
LDA		[24,78,85]	[98]			[78,24]	[78]				
PCA		[24,10]	[98]			[24]					
HuM		[132]									
Wavelet	[144]	[2,118]	[100,144]						[66]		[2]
NNMF					[125]						[51]

considered in a second experiment, where they report an increased performance.

Zhu-Zhou et al. [167] define a speech-based emotion recognition system that uses MFCCs as features while different real-world scenarios with noise and reverberation are simulated. Mariooryard and Busso [88] propose an emotion recognizing system based on SVM with linear kernel. They put the emphasis on factorizing the speaker characteristics, verbal content and expressive behaviors by applying a metric to quantify the dependency between acoustic features and communication traits. Chen et al. [24] compares two different classifiers such as SVM and MLP to recognize speech emotion. They extract several energy and spectral based instantaneous features from the utterances. Apart from these features, the first and second derivatives are also computed as well as a number of statistics of the instantaneous features. They start with 288 candidate features, which are reduced by applying two different methods (LDA and PCA). Thus, they propose four different models by combining the classifiers and dimensionality reduction methods. Generally, they achieve better global performance with LDA + SVM based methods. Sun et al. [132] propose the use of Hu moments as features to detect emotions in speech and compare the results with other usual features such as MFCC and LPCC as unique set of features or as acombination with others. They use a SVM with polynomial kernel. Liu et al. [78] use an ELM decision tree to classify emotions samples from which a set of features has been extracted and selected by applying LDA and correlation analysis. The experimental results are also compared to other conventional classifiers as SVM, MLP and kNN. Akash et al. [3] use three different classic machine learning methods (SVM, MLP and Bayes networks) to classify emotions. They use a combination of features extracted from the original waveform and its spectrogram. Kishore and Satish [66] compare the use of MFCC and wavelet features using GMM classifiers to classify emotions reporting that the wavelet features outperforms the MFCC. Garg et al. [41] employ a hierarchical decision tree method with SVM, BLG and SVR classifiers to recognize emotions. They use 16 low-level descriptors, which cover prosodic, spectral and voice quality features as, for example, MFCC and ZCR, reporting results comparable to the state-of-the-art. Zhang et al. [158] study the influence of several features on the accuracy of speech emotion recognition. The MFCC and ACFC techniques are compared with GMM. Neiberg et al. [97] use GMM to compare three different features sets: standard MFCC, MFCC-low, which are calculated between 20 and 300 HZ to model pitch, and plain pitch features. They use two different databases in English and Swedish languages. They report both MFCC variants are comparable and outperform the pitch features. Seehapoch and Wongthanavasu [116] recognize and classify the speech emotion. They propose energy, F0, ZCR, LPC and MFCC as features and SVM as classifiers. The experimental results offer an better accuracy than state-of-art when a combination of F0, MFCC and energy-based features are used. Daneshfar et al. [28] propose a classifier based on a neural network. The features are extracted from each frame taking into account both spectral (i.e., MFCC) and prosodic (i.e., F0) approaches. Then, the dimensionality is reduced by means of a novel particle swarm optimization (PSO) based method. Li and Akagi [72] present a schema for multilingual speech emotion recognition, in which they emphasize the importance of feature selection. Logistic model trees (LMT) are used to recognize the emotion categories, reporting an improvement over other previously tested techniques such as ANFIS [71]. Kerkeni et al. [64] propose a global approach for speech emotion recognition based on a method to select an optimal combination of features. They use SVM and RNN as classifiers. Wang et al. [144] compare the performance improvement by using wavelet packet coefficient (WPC) over the conventional MFCC to extract features from emotion utterances. They use SVM classifiers with linear and radial basis function kernels.

Zhang et al. [161], compare classic classifiers such as kNN and SVM against a kernel isometric mapping-based classifier to recognize emotions. Rieger et al. [111] propose the use of MFCC and other spectral-based features and a ensemble of kNN classifiers to determine the emotion in speech recordings. Abdel-Hamid [2] applies several conventional techniques for feature extraction (prosodic-based, MFCC and wavelet) over a Egyptian Arabic to recognize emotions on speech database. Both SVM and kNN classifiers are used. Chavan and Gohokar [23] propose SVM classifiers with linear, polynomial, RBF and sigmoid kernels to recognize emotion in speech with MFCC and other features (periodicity histogram and fluctuation pattern). Kerkeni et al. [63] also compare several machine learning classifiers for emotion detection and recognition, namely: multivariate linear regression (MLR), SVM and RNN, while Zamil et al. [155] use a logistic model tree (LMT) classifier and MFCCs as features (the related ones to spectral details of excitation and periodicity of the wave). The signal extracted features are the MFCCs and MS computed with and without speaker normalization. Zhao et al. [165] present an approach of robust emotion recognition in noisy environment by using a weighted sparse representation based on the maximum likelihood estimation. They compare results with six conventional machine learning classifiers (kNN, DT, radial basis function MLP, SVM and sparse representation classifier). Matin and Valles [90] propose SVM and a combination of MFCC and ZCR to classify emotions to be applied in the context of autism treatment. The aim of their system is to aid children with ASD to recognize emotions. The model is used to train these children in such a way they are able to identify human emotions in oral conversations. Iliou and Anagnostopoulos [54] compare RF and MLP to classify emotions using MFCC as features. Xiao et al. [149] present a preliminary study on selection of features for speech emotion recognition. The features include statistics of F0, the first 3 formants and energy. Hou et al. [51] propose the use of non-negative matrix factorization (NMF) to reduce the dimensionality in feature space. They use a conventional kNN classifier. They report improvements over other dimensionality techniques.

Sunitha and Ponnusamy [133] compare the results achieved by applying a SVM classifier and MFCC as features to the EMODB [18] and a Tamil language database created by themselves. The results achieved in both databases are comparable. Rajasekhar and Hota [108] use SVM and a combination of MFCC, pitch and amplitude as features to classify the samples of a database compiled by themselves with a reduced set of emotions. They put the emphasis on the preprocessing of the data in order to improve the recognizing results. Han and Wan [46] propose a speech emotion recognition model based on Proximal SVM to recognize a quite reduced set of emotions in a database recorded by themselves in Chinese language. The method improves the accuracy achieved by common SVM based approaches as well as the response time. Zhang [160] propose a fuzzy SVM to recognize emotions. The reported results improve other SVM classifiers and other classic machine learning methods as GMM and kNN. Atassi and Esposito [11] propose a model based on a combination of classic techniques (GMM and SFFS) to classify emotions. Lugger and Yang [82] propose a model based on GMM to classify emotions. They extract an initial set of 216 suprasegmental and segmental features, which is reduced by applying a SFFS algorithm. They report a detailed study on the relevance of the feature groups for classifying different emotion dimensions. Arias et al. [10] propose low level descriptors such as F0 contours to detect emotions. The model uses PCA for dimension reduction before the emotion classification. Song et al. [125] proposed a speech emotion recognition system based on the non-negative matrix factorization method, which is popular in computer vision and pattern recognition domains. Basically, they consider the discrepancies between source and target data to determine the class of a sample.

Shegokar and Sircar [118] utilize a type of continuous wavelet transform (Morlet wavelet) as features, and quadratic SVM classifiers to recognize all the eight emotions. Rajisha et al. [109] present MLP and SVM classifiers over a Malayala language database. They use the popular MFCC and other energy based features. They report a good performance although the database contains a limited number of emotions. Mao et al. [86] propose a hybrid approach based on HMM and ANN, combining the dynamic time warping of HMM and the pattern recognition of ANN. The utterances are considered as a series of voiced segments, which are used to extract the feature vectors for the ANN. Lin and Wei [76] use two methods (HMM and SVM) to classify emotions. They find the best subset of instantaneous features by applying SFS and then the results are compared to the results when MFCC are used features. They compute several recognition rates while grouping the samples according to several criteria as for example the gender and they found male samples get a higher recognition rate than female or gender independent cases. Yun and Yoo [154] consider a method for speech emotion recognition that incorporates a loss function based on a the Watson and Tellegen emotion model [136]. Each emotion is modeled by a single HMM that maximizes the minimum separation margin between emotions. The margin is scaled by the loss function. Chenchah and Lachiri [26] present a system that uses HMM as classifier and a combination of classical features, such as MFCC, and other features more robust to noise and reverberation distortions, such as PNCC, to recognize emotions in real-life conditions. Yu [153] reports that results achieved by HMM outperform other classification techniques (kNN and LDA) in a Chinese Mandarin language database with a reduced number of emotions. Several features related to pitch, energy, and spectrum, such as first formants, LPCC and MFCC, are used. Vlasenko et al. [141] use MFCC plus speed and acceleration coefficients as features to classify emotions. They use several classifiers: GMM for basic emotion classification, HMM since multiple states are considered, and finally SVM when other low-level descriptors based on prosodic and articulatory aspects are also employed. Schuller et al. [115] apply GMM and continuous HMM to classify instances from two emotion databases in English and German languages. First they classify the utterances by using global statistics and the help of GMM, then the HMM models are applied over low-level instantaneous features. Schuller et al. [114] compare concepts for robust fusion of prosodic and verbal cues in speech emotion recognition. They use a bag-of-words representation for linguistic content analysis. Then, they proceed to an systematic feature selection by using a SVM-SFFS method. The classification is based on an ensemble approach, where base classifiers are SVM, DT, and NBC. The experiment are carried out over the EMODB [18] database.

Kadiri et al. [60] present an interesting alternative to the features extraction problem. They use a different approach to the usually employed in the state-of-art by capturing the deviations between the emotional speech and the neutral, non-emotional one. They report that their method based on a hierarchical binary decision tree is comparable or better than the existing prosody and spectral features. Wu et al. [148] apply three types of classifiers (GMM, SVM and MLP) and a meta decision tree (MDT) to fusion the confidence results. They use acoustic-prosodic information and semantic labels as features. Although the number of emotion labels is quite limited they are able to validate the proposal about MDT by improving the results achieved by the classifiers separately. Sreenivasa Rao et al. [127] propose a GMM as classifier and MFCC as features to recognize emotion in speech. They evaluate the proposal for two databases in Telugu language, acted and real emotion speech, respectively. Mao et al. [85] combine the use of conventional machine learning techniques such as SVM and LDA to classify emotions in a Chinese database. The use of LDA improves the results when applied with all tested classifiers. Bozkurt et al. [16] use spectrally weighted MFCC as features to detect emotions by HMM based classifiers. The spectral weighting is derived from the normalized inverse harmonic mean of the line spectral frequency features. The proposal tries to obtain a data fusion of spectral content and formant location information. Kaya and Karpov [62] propose normalization strategies that apply on several corpus emotion recognition. An extreme learning machine (ELM) is used as classifier. They conclude that significant improvements can be achieved by applying these techniques.

Iqbal and Barua [55] present a real-time recognition system based on Gradient Boosting (GB) and MFCC and other spectral features to classify emotions into the four basic emotions: anger, happiness, sadness and neutral. They report an accuracy equivalent to other works in the state-of-art. Dimitrova-Grekow and Konopko [31] use several machine learning algorithms to detect the emotions and they obtain the best result with random forest (RF). As features they use fundamental frequency series (FFS). Rong et al. [112] also use RF and other decision trees algorithms with MFCC as features to classify emotions in two Chinese databases with a reduced number of examples and emotions. Issa et al. [56] introduce a new architecture based on one-dimensional CNN for identification of emotions using MFCC and other spectral features for training. Caponetti et al. [22] propose the use of LSTM recurrent neural networks to recognize emotions. They use two different features: MFCC and biologically inspired Lyon cochlear model. They report that the latter give better recognition results in comparison to the MFCC features. Ramakrishnan and El Emary [110] present a wide range of features for SER and analyze their performance with the help of HMM and SVM classifiers. Palo and Mohanty [100] develop a reduced set of combinational features for speech emotion recognition. They validate the use of LPCC and MFCC derived from wavelets to train RFB neural networks reporting that their proposal outperforms state-of-art. Cao et al. [21] utilize a SVM-based system to recognize emotions in speech. Their proposal considers a ranking approach which incorporates information about the general expressivity of speakers.

Lotjidereshgi and Gournay [81] propose a biologically inspired approach to classify the emotion. The method is based on liquid state machines (LSM), a classifier inspired on the spiking neural network (SNN) model giving accuracy rates comparable to the state-of-the-art. Ooi et al. [98] propose an architecture in which prosodic and spectral features are used to recognize the emotions. The prosodic features are used to differentiate emotions. The spectral features, particularly MFCC, are processed by applying PDA and LDA analysis and the results are used to train a RBF neural network. Origlia et al. [99] propose the use of phonetic syllables rather than other approaches based on prosody. They evaluate the model in the valence, activation and dominance space and show that it is competitive with other contemporary works in terms of performance. Vlasenko et al. [140] present a model to classify emotion based on HMM and MFCC features. They use two different databases to check out the proposal: while one is used to train the model, the other is used for testing.

Some works reportingcomparisons of classifiers, like [149,46,148,68,126,40,157,155,55], reduce the number of emotions in order to report improved performances. Thirumuru et al. [137] propose a novel feature representation using single frequency filtering and nonlinear energy operator to be used with machine learning classifiers (SVM).

Multimodal approaches benefit from the fusion of diverse kinds of signals. Chen et al. [25] use a K-means clustering-based method to improve the multimodal emotions classifying in HRI contexts. Zhang et al. [157] propose several models to recognize emotion in a multimodal context. In this case they consider speech but also song and video. They use a combination of low level descriptors that includes energy and spectral features, MFCCs and voicing-related low level descriptors. Then, they calculate several statistics resulting in 1,170 features. A directed acyclic graph (DAG) SVM is used as classifier. Mower et al. [95] propose a multimodal framework based one emotion profiles and SVM.

4. Deep learning approaches

Fig. 1 shows a typical structure combining several deep learning architectures, i.e. CNNs and LSTM, processing as input the MFCC image and providing the emotion identification.Many early approaches, such as Huang et al. [53] used CNN to learn affective salient features for speech emotion recognition. Lim et al. [75] propose a method based on concatenated CNN and RNN without using any traditional hand-crafted features, i.e., a kind of spectrogram images. They report a better accuracy than they achieved using conventional methods at the time of publication. Wani et al. [147] also define a deep neural network with CNN layers for classification. Ando et al. [8] present a speech emotion recognition (SER) scheme to extend models generated from multiple listener data to listener-dependent (LD) emotion recognition models. They propose and compare three procedures to adapt deep neural networks with CNN and LSTM layers, where the inputs are acoustic features, and the output is the emotion class. Meng et al. [91] and Zhao et al. [164] also combine CNN with LSTM to introduce a novel architecture for emotion detection. They compute the spectrograms from audio recordings and pick up the values of the 3D Log-Mel spectrum to feed their model. Pandey et al. [102] propose yet another deep neural network with convolutional layers and LSTM to recognize emotions.

Fuentes et al. [38] propose the use of a culturally adapted CNN model to recognize speech emotions to be applied as input for a recommendation system. They combine of MFCC and LFE to create an image-based feature dataset. Anvarjon et al. [9] suggest a convolutional neural network (CNN) fed with spectrograms generated from the audio recordings. They propose a model composed by three blocks with several convolutional layers followed by a pooling layer to extract the features plus two additional dense layers and a softmax layer as output. Jiang et al. [59] propose the use of parallelized convolutional recurrent neural networks for emotion recognition. First, they extract features from each utterance, which are learned by using a long short-term memory. Also, they compute the Mel spectrograms and use a CNN to extract features for the images. These two sets of high-level features are learned and normalized and injected over a softmax classifier to determine the emotion in the utterance. Zhang et al. [159] propose a model of multi-CNN to detect emotion in utterances: a 1D CNN for the raw audio recording, a 2D CNN for the Mel-spectrogram generated from the original waveform, and a 3D CNN for temporal-spatial dynamic. The outputs are integrated by an average-pooling layer to produce the classification result. The experiments are performed on Chinese language databases and lightly improved other state-of-art solutions. De Pinto et al. [104] propose a CNN model based on 1D convolutional layers to classify emotions. Shahin et al. [117] compare the results of their model based on CNN and LSTM to the results achieved with conventional classifiers in a wide range of corpora. Nagase et al. [96] apply a deep neural network composed by convolutional and LSTM layers for emotion recognition. They propose label smoothing in order to reduce the overfitting produced by mislabeled information. They use the IEMOCAP [19] and a Japanase dataset. Andayani et al. [7,6] propose a neural architecture based on LSTM and transformers. Mocanu et al. [94] propose a 2D CNN with deep metric learning to emotion recognition. Gokilavani et al. [43] propose a model with five CNN layers.

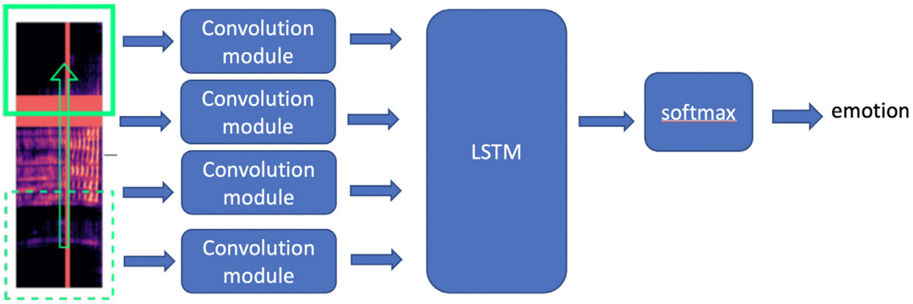

Fig. 1. A typical deep learning architecture for SER based on MFCC input treated as a sequence of images [29].

They also apply some techniques for data augmentation such as adding noise to the original audio signals. Sultana et al. [131] present a system based on CNN and LSTM networks to classify emotions. They use the RAVDESS database [79] and a second dataset in Bangla language. They use transfer learning between both datasets, i.e. they train on RADVESS and test on the Bangla dataset. Guo et al. [45] use the phase information as well as the usual magnitude information in data to improve the emotion recognizing results.

Slimi et al. [124] propose the use of a DL network to be used with small-sized databases. The method does not apply conventional data augmentation techniques. Firstly, the spectrogram associated to each recording is generated by using conventional techniques, and resized while preserving the aspect ratio. They report that the bigger the final image dimensions, the higher the accuracy, although it also implies greater computational complexity, and increasing thenumber of neural network parameters. The neural network consists of just one hidden layer and they report better accuracy than the state of the art.

Praseetha and Vadivel [107] propose a deep learning and recurrent neural networks to classify emotions. They feed the networks with features based on MFCC and spectrograms generated from the audio recordings. The DL networks is composed by two hidden layers apart from the input and output ones. All of them are fully connected. The recurrent model is based in a set of gated recurrent units (GRU). They report better accuracy with their proposal than the state-of-the-art. Hasan and Islam [48] use a recurrent neural network (RNN) to categorize emotions in Bengali speech. The features are based on MFCC and MS and they try to recognize six emotions: happiness, sadness, anger, surprise, fear and disgust. They report quite discrete accuracy in these preliminary works on a new dataset.

Many DL approaches to SER rely on transfer learning of well known DL architectures, often applied to spectrum images. Stolar et al. [130] and Badshah et al. [12] present methods based on CNN applied to spectrogram images. Their transfer learning approach uses a pre-trained AlexNet model [69]. Huang and Bao [52] use an AlexNet [69] deep learning architecture with MFCCs as features. They justify the application of 2D convolutional layers with such features because they combine the coefficient information in an axis and its value in the other axis. Zhang et al. [163] consider transfer learning because the limited data. They propose an attention mechanism based on a fully convolutional network that detects which time–frequency region of a speech spectrogram is more relevant for the emotion detection. Gerczuk et al. [42] use a transfer learning approach with a multi-corpus database composed by 26 free available corpora. The final dataset (EmoSet) contains 84,181 multi-lingual audio recordings and it has a duration over 65 h. They use several convolutional neural network architectures and spectrogram generated from the original audio recordings. Popova et al. [106] propose a DL transfer approach based on a VGG-16 [120] and mel spectrograms as features. Wang et al. [143] propose a deep neural network approach based on VGG architectures [120] with bidirectional LSTM.

Tripathi et al. [138] propose a ResNet [49] based neural network on speech features and trained under focal loss to recognize emotion in speech. Focal loss reshapes the cross-entropy loss function by giving less importance during training to easy examples, usually the majority of the dataset, and focusing more on the hard ones. The method tries to improve the accuracy when there exists a significant class imbalance among various classes. Li et al. [73] also apply ResNet-like deep neural networks [49] to classify emotions in the IEMOCAP [19] database.

Park et al. [103] present a data augmentation method for speech recognition that is applied directly to the feature inputs of neural networks. It consists of warping the features, masking blocks of frequency channels and masking blocks of time steps. They report the method improves the performance of the tested networks and they are able to obtain state of art results by augmenting the training set using these policies even without the aid of language models. Yi et al. [151] use a DL network approach with a generative adversial network for data augmentation. Shilandari et al. [119] and Latif et al. [70] also propose the use of GAN for data augmentation. Bakhshi et al. [13] generate textured images from the audio recordings to be used by CNN for emotion classifying.

Zeng et al. [156] propose a deep learning approach based on a ResNet [49] architecture extended with a gated mechanism similar to the used in LSTM. They use spectrograms generated from the audio files as features. Jannat et al. [58] use an Inception-v3 [135] deep learning architecture in a multimodal approach. In this preliminary work they consider only two emotions (happiness and sadness). They report audio-only performance with an accuracy of 66.41% (cross-validation). They use plots of the raw audio signal as features.

Sanchez-Gutierrez and Gonzalez-Perez [113] apply several discriminative measures (i.e, Anova, Fisher score, etc.) to identify useful neural nodes in deep learning networks in order to prune them and to diminish the resulting error rate. The analysis is performed on several datasets as for example the EMODB [18] speech emotion database. Manohar and Logashanmugam [84] propose a feature selection method to increase the performance in deep neural networks for emotion classification.

Wang et al. [145] present a multimodal system to recognize emotions from images and audio recordings. Both subsystems are based on deep learning models. In particular the audio subsystem uses CNN and LSTM networks, which are fed with spectrogram images generated from the recordings. Heredia et al. [50] propose a multimodal (video, audio, and text) DL architecture to detect emotions in social robots. Middya et al. [92] propose another multimodal deep learning approach for emotion classification. Dong et al. [32] follow yet another multimodal approach based on deep learning to detect emotions (audio and video). Braunschweiler et al. [17] propose a model to classify emotions in a multimodal context (audio and text). Zhou et al. [166] propose a method based on multi-classifier interactive learning to improve the classification accuracy.

5. Conclusions

Nowadays, one of the pillars of science is the availability of data to sustain new developments and the comparison among predictive models.In the field of SER there are many local small datasets. A few datasets are larger, but still not very extensive for the upcoming needs. Some of the most recent datasets have not been exploited, while only two of the older datasets have been exploited reaching conclusions that may become obsolete as new datasets are proposed. In some respects the field of SER is becoming mature regarding the kind of signal features that are more relevant, and the pervasive application of DL architectures is already flourishing in SER literature. However, these approaches need further validation and analysis of their sensitivity to data sampling and the assessment of overfitting to the actual dataset used for validation.

The most inmediate road for future work is the joint exploitation of available corpora, such as reviewed in [162]. These works would allow to overcome the shortage of data and learm how to build more general recognizers, that are able to recognice emotions across the diversity of idiosyncratic features of the various databases. Additionally, the field of deep learning is bringing new architectures and features to light at high speed, hence it is to be expercted that a number of big jumps in performance will be appearing in the near future, because of the high value of emotional interaction with the users.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been carried out with the support of project PID2020-116346GB-I00 funded by the Spanish MICIN.

Appendix A. Abbreviations

ACFC	Auto Correlation Function Coefficients
ANN	Artificial Neural Network
ANFIS	Adaptive Neuro Fuzzy Inference Systems
ASD	Autism Spectrum Disorder
ASR	Automatic Speech Recognition
NBC	Naïve Bayesian Classifier
BLG	Bayesian Logistic Regression
CNN	Convolutional Neural Networks
DT	Decision Trees
ELM	Extreme Learning Machine
F0	Fundamental Frequency
FFS	Fundamental Frequency Series
GAN	Generative Adversarial Networks
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HMI	Human–Machine Interaction
HMM	Hidden Markov Model
kNN	k Nearest Neighbor
LDA	Linear Discriminant Analysis
LFE	Log Mel-Filterbank Energies
LMT	Logistic Model Trees
LSM	Liquid State Machines
LSP	Line Spectral Pairs
LSTM	Long-Short Term Memory
MDT	Meta Decision Tree
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MS	Modulation Spectral
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PNCC	Power Normalized Cepstral Coefficients
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Networks
SBC	Subband based Cepstral Parameter
SER	Speech Emotion Recognition
SFFS	Sequential Floating Forward Selection
SNN	Spiking Neural Networks
SVM	Support Vector Machine
SVR	Support Vector Regression
WPC	Wavelet Packet Coefficient
ZCR	Zero Crossing Rate

References

[1] B.J. Abbaschian, D. Sierra-Sosa, A. Elmaghraby, Deep learning techniques for speech emotion recognition, from databases to models, Sensors 21 (2021) 1249.

[2] L. Abdel-Hamid, Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features, Speech Communication 122 (2020) 19–30.

[3] S. Akash, K. Aschana, M. Abhijith, M. Shuvalila, Speech based emotion recognition system, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 5 (6) (2016) 39–42.

[4] M.B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76.

[5] C.N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011, Artificial Intelligence Review 43 (2015) 155–177.

[6] F. Andayani, L.B. Theng, M.T. Tsun, C. Chua, Hybrid lstm-transformer model for emotion recognition from speech audio files, IEEE Access 10 (2022) 36018–36027, https://doi.org/10.1109/ACCESS.2022.3163856.

[7] F. Andayani, L.B. Theng, M.T. Tsun, C. Chua, Recognition of emotion in speech-related audio files with lstm-transformer, in: 2022 5th International Conference on Computing and Informatics (ICCI), 2022, pp. 087–091, https://doi.org/10.1109/ICCI54321.2022.9756100.

[8] A. Ando, T. Mori, S. Kobashikawa, T. Toda, Speech emotion recognition based on listener-dependent emotion perception models, APSIPA Transactions on Signal and Information Processing 10 (2021).

[9] T. Anrarjon, Kwon Mustaqeem, S.: Deep-net: A lightweight CNN-based speech emotion recognition system using deep system using deep, Sensors 20 (2020) 5212.

[10] J.P. Arias, C. Busso, N. Becerra, Shape-based modeling of the fundamental frequency contour for emotion detection in speech, Computer Speech and Language 28 (2014) 278–294.

[11] H. Atassi, A. Esposito, A speaker independent approach to the classification of emotional vocal expressions, in: 20th IEEE International Conference on Tools with Artificial Intelligence, 2008, pp. 147–152.

[12] Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network (2017).

[13] A. Bakhshi, A. Harimi, S. Chalup, Cytex: Transforming speech to textured images for speech emotion recognition, Speech Communication 139 (2022) 62–75, https://doi.org/10.1016/j.specom.2022.02.007.

[14] Y. Bhavani, S.B. Swathi, R.R. Aileni, M.R. Gaddam, A survey on various speech emotion recognition techniques, in: 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 1099–1104, https://doi.org/10.1109/ICAIS53314.2022.9742874.

[15] Bhutekar, S.D., Chandak, M.B.: Designing and recording emotional speech databases. In: National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012). pp. 6–10 (2012).

[16] E. Bozkurt, E. Erzin, Ç.E. Erlem, A.T. Erdem, Formant position based weighted spectral features for emotion recognition, Speech Communication 53 (2011) 1186–1197.

[17] N. Braunschweiler, R. Doddipatla, S. Keizer, S. Stoyanchev, Factors in emotion recognition with deep learning models using speech and text on multiple corpora, IEEE Signal Processing Letters 29 (2022) 722–726, https://doi.org/10.1109/LSP.2022.3151551.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proc. 9th European Conf. Speech Communication and Technology, 2005, pp. 1517–1520.

[19] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (4) (2008) 335–359.

[20] J. Cai, R. Xiao, W. Cui, S. Zhang, G. Liu, Application of electroencephalography-based machine learning in emotion recognition: A review, Frontiers in Systems Neuroscience 15 (2021), https://doi.org/10.3389/fnsys.2021.729707.

[21] H. Cao, R. Verma, A. Nenkova, Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech, Computer Speech and Language 29 (2015) 186–202.

[22] L. Caponetti, C.A. Buscicchio, G. Castellano, Biologically inspired emotion recognition from speech, EURASIP Journal on Advances in Signal Processing 2011 (2011) 24.

[23] V.M. Chavan, V.V. Gohokar, Speech emotion recognition by using SVM-classifier, Int. J. Engineering and Advanced Technology 1 (5) (2012) 11–15.

[24] L. Chen, X. Mao, Y. Xue, L.L. Cheng, Speech emotion recognition: Features and classification models, Digital Signal Processing 22 (2012) 1154–1160.

[25] L. Chen, K. Wang, M. Li, M. Wu, W. Pedrycz, K. Hirota, K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction, IEEE Transactions on Industrial Electronics (2022), https://doi.org/10.1109/TIE.2022.3150097, 1–1.

[26] F. Chenchah, Z. Lachiri, A bio-inspired emotion recognition system under real-life conditions, Applied Acoustics 115 (2017) 6–14.

[27] P.P. Dahake, K. Shaw, P. Malathi, Speaker dependent speech emotion recognition using MFCC and support vector machine, in: 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), 2016, pp. 1080–1084.

[28] F. Daneshfar, S.J. Kabudian, A. Neekabadi, Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier, Applied Acoustics 166 (2020).

[29] J. De Lope, M. Graña, A hybrid time-distributed deep neural architecture for speech emotion recognition, International Journal of Neural Systems 32 (06) (2022) 2250024, https://doi.org/10.1142/S0129065722500241, pMID: 35575003.

[30] L. Deng, Deep learning: from speech recognition to language and multimodal processing, APSIPA Transactions on Signal and Information Processing 5 (2016).

[31] T. Dimitrova-Grekow, P. Konopko, New parameters for improving emotion recognition in human voice, in: Proc. IEEE 2019 International Conference Systems, Man and Cybernetics, 2019, pp. 4205–4210.

[32] G.N. Dong, C.M. Pun, Z. Zhang, Temporal relation inference network for multi-modal speech emotion recognition, in: IEEE Transactions on Circuits and Systems for Video Technology, 2022, https://doi.org/10.1109/TCSVT.2022.3163445, 1–1.

[33] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: Towards a new generation of databases, Speech Communication 40 (2003) 33–60.

[34] K. Duouis, M.K. Pichora-Fuller, Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set, Canadian Acoustics - Acoustique Canadienne 39 (3) (2011) 182–183.

[35] P. Ekman, W. Friesen, Constants across cultures in face and emotions, J. Personality and Social Psychology 17 (2) (1971) 124–129.

[36] Engberg, I.S., Hansen, A.V.: Documentation of the Danish emotional speech database. Tech. rep., Center for Person Kommunilation, Denmark (1996).

[37] I.S. Engberg, A.V. Hansen, O. Andersen, P. Dalsgaard, Design, recording and verification of a Danish emotional speech database, in: Proc. 5th European Conf. Speech Communication and Technology, 1997, pp. 1695–1698.

[38] J. Fuentes, J. Taverner, J.A. Rincon, Towards a classifier to recognize emotions using voice to improve recommendations, in: F. De La Prieta et al. (Eds.), Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness, Springer, Cham, 2020, pp. 218–225.

[39] P. Gangamohan, S.R. Kadiri, B. Yegnanarayama, Analysis of emotional speech—a review, in: A. Esposito, L.C. Jain (Eds.), Toward Robotic Socially Believable Behaving Systems -, vol. I, Springer, Cham, 2016, pp. 205–238.

[40] Y. Gao, B. Li, N. Wang, T. Zhu, Speech emotion recognition using local and global features, Int. Conf. Brain Informatics (2017) 3–13.

[41] Garg, V., Kumar, H., Sinha, R.: Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers (2013).

[42] Gerczuk, M., Amiriparian, S., Otti, S., Schuller, B.W.: EmoNet: A transfer learning framework for multi-corpus speech emotion recognition. arXiv p. 2103.08310v1 (2021).

[43] M. Gokilavani, H. Katakam, S.A. Basheer, P. Srinivas, Ravdness, crema-d, tess based algorithm for emotion recognition using speech, in: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 1625–1631, https://doi.org/10.1109/ICSSIT53264.2022.9716313.

[44] M. Grimm, K. Kroschel, S. Narayanan, The Vera am Mittag German audio-visual emotional speech database, in: Proceedings of the IEEE International Conference on Multimedia and Expo ICME 2008, 2008.

[45] L. Guo, L. Wang, J. Dang, E.S. Chng, S. Nakagawa, Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition, Speech Communication 136 (2022) 118–127, https://doi.org/10.1016/j.specom.2021.11.005.

[46] Z. Han, J. Wang, Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine, in: 2017 Chinese Automation Congress (CAC), 2017, pp. 2513–2516.

[47] Haq, S., Jackson, P.J.B.: Multimodal emotion recognition. In: Wang, W. (ed.) Machine audition: Principles, algorithms and systems, pp. 398–423. IGI Global (2010).

[48] M. Hasan, A. Islam, Emotion recognition from Bengali speech using RNN modulation-based categorization, in: Proc. IEEE Third International Conference on Smart Systems and Inventive Technology, 2017, pp. 1131–1136.

[49] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv, 1512.03385 (2015)

[50] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, A. Aguilera, Adaptive multimodal emotion detection architecture for social robots, IEEE Access 10 (2022) 20727–20744, https://doi.org/10.1109/ACCESS.2022.3149214.

[51] M. Hou, J. Li, G. Lu, A supervised non-negative matrix factorization model for speech emotion recognition, Speech Communication 124 (2020) 13–20.

[52] Huang, A., Bao, P.: Human vocal sentiment analysis. arXiv, 1905.08632 (2019)

[53] Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. pp. 80–804 (2013).

[54] T. Iliou, C.N. Anagnostopoulos, Comparison of different classifiers for emotion recognition, in: 13th Panhellenic Conference on Informatics Comparison Of Different Classifiers for Emotion, 2009, pp. 102–106.

[55] Iqbal, A., Barua, K.: A real-time emotion recognition from speech using gradient boosting. In: Proc. Int. Conf. Electrical, Computer and Communication Engineering. pp. 1–5 (2019).

[56] D. Issa, M. Faith-Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (2020).

[57] J. James, L. Tian, C.I. Watson, An open source emotional speech corpus for human robot interaction, Proc. Interspeech (2018) 2768–2772.

[58] Jannat, R., Tynes, I., LaLime, L., Adorno, J., Canavan, S.: Ubiquitous emotion recognition using audio and video data. In: UbiComp/ISWC 2018. pp. 956–959 (2018)

[59] P. Jiang, H. Fu, H. Tao, P. Lei, L. Zhao, Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition, IEEE Access 7 (2019) 90368–90377.

[60] S.R. Kadiri, P. Gangamohan, S.V. Gangashetty, P. Alku, B. Yegnanarayana, Excitation features of speech for emotion recognition using neutral speech as reference, Circuits, Systems, and Signal Processing 39 (2020) 4459–4481.

[61] U. Kamath, J. Liu, J. Whitaker, Deep Learning for NLP and Speech Recognition, Springer Nature, Cham, 2019.

[62] H. Kaya, A.A. Karpov, Efficient and effective strategies for cross-corpus acoustic emotion recognition, Neurocomputing 275 (2018) 1028–1034.

[63] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M.A. Mahjoub, C. Cleder, in: A. Cano (Ed.), Automatic speech emotion recognition using machine learning, Social Media and Machine Learning. IntechOpen, 2019.

[64] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M.A. Mahjoub, C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO, Speech Communication 114 (2019) 22–35.

[65] R.A. Khalil, E. Jonesa, M.I. Babar, T. Jan, M.H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review, IEEE Access 7 (2019) 117327–117345.

[66] K.V.K. Kishore, P.K. Satish, Emotion recognition in speech using MFCC and wavelet features, in: 3rd IEEE International Advance Computing Conference (IACC), 2013, pp. 842–847.

[67] M. Kotti, C. Kotropoulos, Gender classification in two emotional speech databases, in: Proc. 19th Int. Conf. on Pattern Recognition, 2008, pp. 1–4.

[68] K.V. Krishna, N. Sainath, A.M. Posonia, Speech emotion recognition using machine learning, in: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 1014–1018, https://doi.org/10.1109/ICCMC53470.2022.9753976.

[69] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012).

[70] S. Latif, R. Rana, S. Khalifa, R. Jurdak, B.W. Schuller, Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition, IEEE Transactions on Affective Computing (2022), https://doi.org/10.1109/TAFFC.2022.3167013, 1–1.

[71] X. Li, M. Akagi, Multilingual speech emotion recognition using a three-layer model, in: Proceedings Interspeech, 2016, pp. 3608–3612.

[72] X. Li, M. Akagi, Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model, Speech Communication 110 (2019) 1–12.

[73] X. Li, Z. Zhang, C. Gan, Y. Xiang, Multi-label speech emotion recognition via inter-class difference loss under response residual network, IEEE Transactions on Multimedia (2022), https://doi.org/10.1109/TMM.2022.3157485, 1–1.

[74] E. Lieskovska, M. Jakubec, R. Jarina, M. Chmulik, A review on speech emotion recognition using deep learning and attention mechanism, Electronics 10 (2021) 1163.

[75] Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. pp. 1–4 (2016).

[76] Y.L. Lin, G. Wei, Speech emotion recognition based on HMM and SVM, Proc. Fourth IEEE Int. Conf. on Machine Learning and Cybernetics. (2005) 4898–4901.

[77] Liu, Z., Hu, B., Li, X., Liu, F., Wang, G., Yang, J.: Detecting depression in speech under different speaking styles and emotional valences. pp. 261–271. Springer (2017).

[78] Z.T. Liu, M. Wu, W.H. Cao, J.W. Mao, J.P. Xu, G.Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.

[79] S.R. Livingstone, F.A. Russo, The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facieal and vocal expressions in North American English, PLoS ONE 13 (5) (2018).

[80] R. Lotfian, C. Busso, Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings, IEEE Trans. Affective Computing 10 (4) (2019) 471–483.

[81] Lotfidereshgi, R., Gournay, P.: Biologically inspired speech emotion recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. pp. 5135–5139 (2017).

[82] Luger, M., Yang, B.: An incremental analysis of different feature groups in speaker independent emotion recognition (2007).

[83] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, Prabal Datta Barua, M. Murugappan, Y. Chakole, U.R. Acharya, Automated emotion recognition: Current trends and future perspectives, Computer Methods and Programs in Biomedicine 215 (2022), https://doi.org/10.1016/j.cmpb.2022.106646.

[84] K. Manohar, E. Logashanmugam, Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm, Knowledge-Based Systems 246 (2022), https://doi.org/10.1016/j.knosys.2022.108659.

[85] J.W. Mao, Y. He, Z.T. Liu, Speech emotion recognition based on linear discriminant analysis and support vector machine decision tree, in: 2018 37th Chinese Control Conference (CCC), 2018, pp. 5529–5533.

[86] Mao, X., Chen, L., Fu, L.: Multi-level speech emotion recognition based on hmm and ann. In: IEEE World Congress on Computer Science and Information Engineering. pp. 225–229 (2009).

[87] V. Mapelli, Inter1sp: Spanish emotional speech synthesis database, European Language Resources Association (2011).

[88] S. Mariooryard, C. Busso, Compensating for speaker or lexical variabilities in speech for emotion recognition, Speech Communication 57 (2014) 1–12.

[89] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops, 2006, pp. 1–8.

[90] R. Matin, D. Valles, A speech emotion recognition solution-based on support vector machine for children with autism spectrum disorder to help identify human emotions, in: 2020 Intermountain Engineering, Technology and Computing (IETC), 2020.

[91] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network, IEEE Access 7 (2019) 125868–125881.

[92] A.I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities, Knowledge-Based Systems 244 (2022), https://doi.org/10.1016/j.knosys.2022.108580.

[93] A. Milton, S.S. Roy, S.T. Selvi, SVM scheme for speech emotion recognition using MFCC feature, International Journal of Computer Applications 69 (9) (2013) 34–39.

[94] B. Mocanu, R. Tapu, Emotion recognition from raw speech signals using 2d cnn with deep metric learning, in: 2022 IEEE International Conference on Consumer Electronics (ICCE), 2022, pp. 1–5, https://doi.org/10.1109/ICCE53296.2022.9730534.

[95] E. Mower, M. Mataric, S. Narayanan, A framework for automatic human emotion classification using emotion profiles, IEEE Trans. on Audio, Speech, and Language Processing 19 (5) (2011) 1057–1070.

[96] R. Nagase, T. Fukumori, Y. Yamashita, Speech emotion recognition using label smoothing based on neutral and anger characteristics, in: 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), 2022, pp. 626–627, https://doi.org/10.1109/LifeTech53646.2022.9754749.

[97] Neiberg, D., Elenius, K., Laskowski, K.: Emotion recognition in spontaneous speech using GMMs (2006).

[98] C.S. Ooi, K.P. Seng, L.M. Ang, L.W. Chew, A new approach of audio emotion recognition, Expert Systems with Applications 41 (2014) 5858–5869.

[99] A. Origlia, F. Cutugno, V. Galatà, Continuous emotion recognition with phonetic syllables, Speech Communication 57 (2014) 155–169.

[100] H.K. Palo, M.N. Mohanty, Wavelet based feature combination for recognition of emotion, Ain Shams Engineering Journal 9 (4) (2018) 1799–1806.

[101] Panda, S.P.: Automated speech recognition system in advancement of human-computer interaction. In: Proc. IEEE 2017 International Conference on Computing Methodologies and Communication. pp. 302–306 (2017).

[102] S.K. Pandey, H.S. Shekhawat, S. Prasanna, Attention gated tensor neural network architectures for speech emotion recognition, Biomedical Signal Processing and Control 71 (2022), https://doi.org/10.1016/j.bspc.2021.103173.

[103] Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A simple data augmentation method for automatic speech recognition. In: Proc. Interspeech 2019. pp. 2613–2617 (2019)

[104] Pinto, M.D., Polignano, M., Lops, P., Semeraro, G.: Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In: IEEE (2020).

[105] J. Pittermann, A. Pittermann, W. Minker, Handling Emotions in Human-Computer Dialogues, Springer, Netherlands, 2010.

[106] A. Popova, A. Rassadin, Ponomarenko A.: Emotion recognition in sound, in: B. Kryzhanovsky (Ed.), Advances in Neural Computation, Machine Learning, and Cognitive Research, Springer, 2018, pp. 117–124.

[107] V.M. Praseetha, S. Vadivel, Deep learning models for speech emotion recognition, J. Computer Science 14 (11) (2018) 1577–1587.

[108] Rajasekhar, A., Hota, M.K.: A study of speech, speaker and emotion recognition using mel frequency cepstrum coefficients and support vector machines. In: 2018 International Conference on Communication and Signal Processing (ICCSP). pp. 114–118 (2018).

[109] T.M. Rajisha, A.P. Sunija, K.S. Riyas, Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM, Procedia Technology 24 (2016) 1097–1104.

[110] S. Ramakrishnan, I.M. El-Emary, Speech emotion recognition approaches in human computer interaction, Telecommunication Systems 52 (3) (2013) 1467–1478.

[111] Rieger, S.A., Muraleedharan, R., Ramachandran, R.P.: Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In: The 9th International Symposium on Chinese Spoken Language Processing. pp. 589–593 (2014).

[112] Rong, J., Chen, Y.P.P., Chowdhury, M., Li, G.: Acoustic features extraction for emotion recognition. In: 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007) (2007).

[113] M.E. Sánchez-Gutiérrez, P.P. González-Pérez, Discriminative neural network pruning in a multiclass environment: A case study in spoken emotion recognition, Speech Communication 120 (2020) 20–30.

[114] B. Schuller, R. Müller, M. Lang, G. Rigoll, Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles, Interspeech (2005) 805–808.

[115] Schuller, B., Rigoll, G., Lang, M.: Hidden markov model-based speech emotion recognition. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing (2003).

[116] Seehapoch, T., Wongthanavasu, S.: Speech emotion recognition using support vector machines. In: Int. Conf. Knowledge and Smart Technology. pp. 86–91 (2013).

[117] I. Shahin, N. Hindawi, A.B. Nassif, A. Alhudhaif, K. Polat, Novel dual-channel long short-term memory compressed capsule networks for emotion recognition, Expert Systems with Applications 188 (2022), https://doi.org/10.1016/j.eswa.2021.116080.

[118] P. Shegokar, Sircar P.: Contnuous wavelet transform based speech emotion recognition, in: International Conference on Signal Processing And Communication Systems, Gold Coast, Australia, 2016, pp. 1–8.

[119] A. Shilandari, H. Marvi, H. Khosravi, W. Wang, Speech emotion recognition using data augmentation method by cycle-generative adversarial networks, Signal, Image and Video Processing (Feb 2022), https://doi.org/10.1007/s11760-022-02156-9.

[120] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv, 1409.1556 (2014).

[121] Singh, Y.B., Goel, S.: Survey on human emotion recognition: Speech database, features and classification. In: Proc. IEEE Int. Conf. Advances in Computing, Communication Control and Networking. pp. 298–301 (2018).

[122] Y.B. Singh, S. Goel, A systematic literature review of speech emotion recognition approaches, Neurocomputing 492 (2022) 245–263, https://doi.org/10.1016/j.neucom.2022.04.028.

[123] Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S.: Emotion recognition from audio signals using support vector machine. In: 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS). pp. 139–144 (2015).

[124] Slimi, A., Hamroun, M., Zrigui, M., Nicolas, H.: Emotion recognition from speech using spectrograms and shallow neural networks. In: ACM Int. Conf. Advances in Mobile Computing & Multimedia. pp. 298–301 (2020).

[125] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, Y. Yu, Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization, Speech Communication 83 (2016) 34–41.

[126] Sowmya, G., Naresh, K., Sri, J.D., Sai, K.P., Indira, D.V.: Speech2emotion: Intensifying emotion detection using mlp through ravdess dataset. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS). pp. 1–3 (2022). DOI: 10.1109/ICEARS53579.2022.9752022.

[127] K. Sreenivasa Rao, T. Pavan Kumar, K. Anusha, B. Leela, I. Bhavana, S.V. Gowtham, Emotion recognition from speech, International Journal of Computer Science and Information Technologies 3 (2) (2012) 3603–3607.

[128] Stanković, T., Karnjanadecha, M., Delić, V.: Improvement of Thai speech emotion recognition by using face feature analysis. In: Int. Symposium Intelligent Signal an Communication Systems. pp. 1–5 (2011).

[129] R. Stock-Homburg, Survey of emotions in human–robot interactions: Perspectives from robotic psychology on 20 years of research, International Journal of Social Robotics 14 (2) (Mar 2022) 389–411, https://doi.org/10.1007/s12369-021-00778-6.

[130] Stolar, M.N., Lech, M., Bolia, R.S., Skinner, M.: Real time speech emotion recognition using RGB image classifcation and transfer learning. In: Proc. 11th IEEE Int. Conf. Signal Processing and Communication Systems. pp. 1–8 (2017).

[131] S. Sultana, M.Z. Iqbal, M.R. Selim, M.M. Rashid, M.S. Rahman, Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks, IEEE Access 10 (2022) 564–578, https://doi.org/10.1109/ACCESS.2021.3136251.

[132] Y. Sun, G. Wen, J. Wang, Weighted spectral features based on local Hu moments for speech emotion recognition, Biomedical Signal Processing and Control 18 (2015) 80–90.

[133] Sunitha-Ram, C., Ponnusamy, R.: An effective automatic speech emotion recognition for Tamil language using support vector machine. In: 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). pp. 19–23 (2014).

[134] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: A review, Int. J. Speech Technology 21 (2018) 93–120.

[135] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J.: Rethinking the Inception architecture for computer vision. arXiv, 1512.00567v3 (2015).

[136] A. Tellegen, D. Watson, T. Hofmann, On the dimensional and hierarchical structure of affect, Psychological Science 10 (4) (1999) 297–303.

[137] R. Thirumuru, K. Gurugubelli, A.K. Vuppala, Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition, Digital Signal Processing 120 (2022), https://doi.org/10.1016/j.dsp.2021.103293.

[138] Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P.: Focal loss based residual convolutional neural network for speech emotion recognition. arXiv, 1906.05682 (2019)

[139] Ververidis, D., Kotropoulos, C.: Automatic speech classification to five emotional states based on gender information. In: 12th IEEE European Signal Processing Conf. pp. 341–344 (2004).

[140] B. Vlasenko, D. Prylipko, R. Böck, A. Wendemuth, Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications, Computer Speech and Language 28 (2014) 48–500.

[141] B. Vlasenko, B. Schuller, A. Wendemuth, G. Rigoll, Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing, in: A. Paiva, R. Prada, R.W. Picard (Eds.), ACII 2007, LNCS 4738, Springer, Berlin Heidelberg, 2007, pp. 139–147.

[142] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M.B. Ali, M. Adan, M. Mujtaba, Generative adversarial networks for speech processing: A review, Computer Speech & Language 72 (2022), https://doi.org/10.1016/j.csl.2021.101308.

[143] C. Wang, Y. Ren, N. Zhang, F. Cui, S. Luo, Speech emotion recognition based on multi-feature and multi-lingual fusion, Multimedia Tools and Applications 81 (4) (2022) 4897–4907, https://doi.org/10.1007/s11042-021-10553-4.

[144] K. Wang, G. Su, L. Liu, S. Wang, Wavelet packet analysis for speaker-independent emotion recognition, Neurocomputing 398 (2020) 257–264.

[145] X. Wang, X. Chen, C. Cao, Human emotion recognition by optimally fusing facial expression and speech feature, Signal Processing: Image Communication 84 (2020).

[146] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, W. Zhang, A systematic review on affective computing: emotion models, databases, and recent advances, Information Fusion 83–84 (2022) 19–52, https://doi.org/10.1016/j.inffus.2022.03.009.

[147] T.M. Wani, T.S. Gunawan, S.A.A. Qadri, M. Kartiwi, E. Ambikairajah, A comprehensive review of speech emotion recognition systems, IEEE Access (2021).

[148] Wu, C.H., Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels (extended abstract). In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 477–483 (2015).

[149] Xiao, Z., Dellandera, E., Dou, W., Chen, L.: Features extraction and selection for emotional speech classification. pp. 411–416 (2005).

[150] Yang, C., Ji, L., Liu, G.: Study to speech emotion recognition based on TWINsSVM. In: 2009 Fifth International Conference on Natural Computation. pp. 312–316 (2009).

[151] L. Yi, M.W. Mak, Improving speech emotion recognition with adversarial data augmentation network, IEEE Transactions on Neural Networks and Learning Systems 33 (1) (2022) 172–184, https://doi.org/10.1109/TNNLS.2020.3027600.

[152] D. Yu, L. Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer-Verlag, 2015.

[153] Yu, W.: Research and implementation of emotional feature classification and recognition in speech signal. In: International Symposium on Intelligent Information Technology Application Workshops. pp. 471–474 (2008).

[154] Yun, S., Yoo, C.D.: Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model. In: IEEE ICASSP. pp. 4169–4172 (2009).

[155] A.A.A. Zamil, S. Hasan, S.M.J. Baki, J.M. Adam, Zaman I.: Emotion detection from speech signals using voting mechanism on classified frames, in: 2019 International Conference on Robotics, Electrical and Signal Processing Technique, Bangladesh, 2019, pp. 281–285.

[156] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools and Applications 78 (2017) 3705–3722.

[157] B. Zhang, G. Essl, Provost E.M.: Recognizing emotion from singing and speaking using shared models, in: 2015 International Conference on Affective Computing and Intelligent Interaction, Xi'an, China, 2015, pp. 139–145.

[158] Zhang, Q., An, N., Wang, K., Ren, F., Li, L.: Speech emotion recognition using combination of features. In: 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP). pp. 523–528 (2013).

[159] S. Zhang, X. Tao, Y. Chuang, X. Zhao, Learning deep multimodal affective features for spontaneous speech emotion recognition, Speech Communication 127 (2021) 73–81.

[160] Zhang, S.: Speech emotion recognition based on fuzzy least squares support vector machines. In: 2008 7th World Congress on Intelligent Control and Automation. pp. 1299–1302 (2008).

[161] Zhang, S., Lei, B., Chen, A., Chen, C., Chen, Y.: KIsomap-based feature extraction for spoken emotion recognition. In: Proc. IEEE 10th International Conference on Signal Processing. pp. 1374–1377 (2010).

[162] S. Zhang, R. Liu, X. Tao, X. Zhao, Deep cross-corpus speech emotion recognition: Recent advances and perspectives, Frontiers in Neurorobotics 15 (2021), https://doi.org/10.3389/fnbot.2021.784514.

[163] Zhang, Y., Du, J., Wang, Z., Zhang, J., Tu, Y.: Attention based fully convolutional network for speech emotion recognition. In: Proc. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1771–1775 (2018).

[164] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomedical Signal Processing and Control 47 (2019) 312–323.

[165] X. Zhao, S. Zhang, B. Lei, Robust emotion recognition in noisy speech via sparse representation, Neural Computing & Applications 24 (2014) 1539–1553.

[166] Y. Zhou, X. Liang, Y. Gu, Y. Yin, L. Yao, Multi-classifier interactive learning for ambiguous speech emotion recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 695–705, https://doi.org/10.1109/TASLP.2022.3145287.

[167] F. Zhu-Zhou, R. Gil-Pita, J. Garcia-Gomez, M. Rosa-Zurera, Robust multi-scenario speech-based emotion recognition system, Sensors 22 (6) (2022), https://doi.org/10.3390/s22062343.