Neurocomputing

A review on speech emotion recognition: Does noise affect performance?

--Manuscript Draft--

Manuscript Number:	NEUCOM-D-23-02732
Article Type:	Survey/review study
Section/Category:	Deep Learning
Keywords:	Affective Computing; Speech Emotion Recognition; Noisy Speech Emotion Recognition; Robust SER; Speech enhancement; Noise robust features
Abstract:	Affective Computing systems can measure the affective state of persons as well as the mindset of the individual. Speech Emotion Recognition (SER) is a unimodal affect computing system based on emotional speech data. It is an active area of research, especially in pattern recognition, computer vision, and deep learning. Many works on SER exist in the literature, but only a few consider SER under uncontrolled noisy conditions. A few good surveys exist for SER, but they either do not cover all aspects of SER in noisy environments or do not discuss the specifics in detail. In recent years, a growing interest in researchers using SER in the natural environment and getting improvement in recognition rate. In this review, the methods and approaches used in noisy SER on the literature before 2017 and from 2017 to the beginning of 2023 have been compiled. We give an overview of noisy SER methods, datasets used for SER under noisy conditions, noisy datasets, types of noise used, and toolkits used for noisy SER recognition. This survey also focuses on classifiers, features used, and limitations of existing research in noisy SER systems.

A review on speech emotion recognition: Does noise affect performance?

Swapna Mol George[a], **Muhamed Ilyas P**[a]

[a]*Sullamussalam Science College, Areekode, Kerala, India*

Abstract

Affective Computing systems can measure the affective state of persons as well as the mindset of the individual. Speech Emotion Recognition (SER) is a unimodal affect computing system based on emotional speech data. It is an active area of research, especially in pattern recognition, computer vision, and deep learning. Many works on SER exist in the literature, but only a few consider SER under uncontrolled noisy conditions. A few good surveys exist for SER, but they either do not cover all aspects of SER in noisy environments or do not discuss the specifics in detail. In recent years, a growing interest in researchers using SER in the natural environment and getting improvement in recognition rate. In this review, the methods and approaches used in noisy SER on the literature before 2017 and from 2017 to the beginning of 2023 have been compiled. We give an overview of noisy SER methods, datasets used for SER under noisy conditions, noisy datasets, types of noise used, and toolkits used for noisy SER recognition. This survey also focuses on classifiers, features used, and limitations of existing research in noisy SER systems.

* Corresponding author: Swapna Mol George, email: smgjaison@gmail.com

*Co-author: Muhamed Ilyas P, email: muhamed.ilyas@gmail.com

Keywords: Affective Computing, Speech Emotion Recognition, Noisy Speech Emotion Recognition, Robust SER

1. Introduction

Speech is a stimulus generated from the human brain's response to some sensory responses, which conveys information about the audio message and the mindset of that person, his inner mental states, and the context of the communication. The subtleties of emphasis, tone, and phrasing, the variations in utterance speed and continuity, and the accompanying physical gestures all convey something of the inner life of impulse and feeling [2]. Speech signal processing is one of the complex research areas due to the number of feature signals in speech signals [1]. Must find relations between speech and emotions [9]. Since machines that have affective capacities will need to be skilled and judicious in how they employ such talents, affective computing is a field of study in need of careful and sensitive investigation. In the literature, there are examples of affective computing systems based on image [10], audio, video [88], text [90], EEG [99], physiological signals [100], and multi-model emotion recognition [101][3][89] with machine learning and deep learning models. Speech Emotion Recognition (SER) is one of the research topics that is quickly growing. There are numerous uses for it, including lie detection and criminal investigation[107], diagnostic tool in medicine[109][42], robotic emotion expressions [4], machine-human interaction systems[126], call center answering[19], robotic assistance and helpline systems[108], theatre performance and interaction improvements [87], mental health and fitness analysis in the classroom and online teaching[63][110], emotional state identification of drivers[125], and intelligence assistance[106][3][32] including, digital advertisement, online gaming, and feedback assessment of customers. Table 1 provides a list of applications for voice emotion recognition systems.

1.1. Previous Surveys

Many survey papers provide an overview of the research on speech emotion recognition systems. Several of these [5][6][7][8][136][133][134][135] provide an in-depth summary of the literature on speech emotion recognition tasks. Table 2 contains an outline of the surveys that have been reviewed. We can see that the main topic of current survey studies is speech-emotion recognition systems in clean environments. However, they do not go into great detail about the noisy speech emotion recognition task. Although noise is an integral part of speech, real-time monitoring of speech emotion recognition systems requires the proper evaluation and assessment of speech emotion recognition systems in noisy environments. There is no recent survey paper focusing on all aspects of noisy speech emotion recognition systems.

1.2. Contributions

We provide a current, thorough, and concise overview of the research on noisy speech-based emotion identification systems in the literature, covering the methodologies, emotional speech databases, and assessment metrics for noisy SER. A total of 60 papers have been reviewed, 20 of which ([111–130]) were published before 2017 and 40 ([11–50]) between 2017 and the start of 2023. Our goal is to inform interested new researchers about the main developments in the past and to point out relevant suggestions for the future.

• We provide an overview of speech emotion recognition systems and a summary of noisy speech emotion recognition systems. The purpose of creating a taxonomy is to provide an overview of the methods in the literature for noisy speech emotion Recognition

• We give an up-to-date review of the speech emotional datasets in literature for noisy SER, noise speech dataset sets, and types of noise used for artificial noise addition.

Table1: Application areas of Speech Emotion Recognition

Application Areas	Specific Applications
Investigation	Lie detection
Healthcare	Diagnostic tool, Depression Identification Emotion-aware e-health systems, recognizing the condition of patients through vocal variabilities such as pain, stress, or fear
Machine Human Interaction Systems	Call centre answering Robotic assistance, social robots Customers' feedback assessment Robotic emotion expressions Helpline systems Intelligent conversational systems
Education	Mental health and fitness analysis in the classroom online teaching Affect-aware learning systems
Intelligence Assistance	Digital advertisement Customers' feedback assessment Online gaming Smart home assistants
Tourism	Recommendation systems for tourism
Entertainment	Theatre performance and interaction improvements
Transport	To assess drivers emotional state

Table 2: Overview of previous surveys

Review Paper	**Summary**
Ververidis et al. [8] 2006	Databases, features, methods
Gunawan et al. [136] 2018	Databases, features, Conventional classifiers, Works from DNN
Mustafa et al. [7] 2018	Databases, features, types of emotions, classification techniques, future direction up to 2017
Jahangir et al. [133] 2021	Databases, features, toolkits, deep learning-based models, merits and demerits of deep learning models, evaluation metrics, research challenges
Wani et al. [5] 2021	Databases, features, classifiers, deep neural networks, Challenges
Fahad et al. [6] 2021	Databases, types of databases, features, conventional models, deep learning-based models, Issues in natural environment, Noisy environment
Al-Dujaili et al. [134] 2023	Databases, features, classes, multiple classifiers
De Lope et al. [135] 2023	Databases, features, conventional classifiers, recent deep learning classifiers

• We provide information on features and classifiers used for speech emotion recognition, the toolkits used in existing literature for the smooth execution of these works, and evaluation metrics used in the literature for noisy SER.

The organization of the paper is as follows. In Section 2, we give an overview of the main concepts related to speech emotion recognition. Section 3, provides an overview of noisy speech emotion recognition and its taxonomy. We summarize the datasets, artificial noise addition, noisy datasets,

methods on noisy speech emotion recognition systems, models, toolkits and evaluation metrics respectively. Finally, in Section 4, we give the main discussions, conclusions, and directions for future works.

2. Speech Emotion Recognition

SER, described as the method of inferring human emotions from speech signals, is an area of study for several scholars. The goal of speech emotion recognition systems, often known as classification or regression problems, is to categorize or identify voice input as various emotion categories. Because supervised techniques are successful in machine learning algorithms, researchers use them in their research. We can see numerous studies in the literature related to traditional hand-crafted feature identification [103][104] to automatic feature identification techniques. Literature shows works from utterance-based [105][128], context-aware [121][138], cross-corpus [11], cross-language [125], and cross-cultural [142] speech emotion recognition systems.

2.1. Steps of Speech Emotion Recognition

Speech emotion recognition requires the following steps (Fig:1)

i) Input emotional speech

ii) Pre-processing

iii) Feature extraction

iv)Feature Selection

v) Classifier Selection.

v) Emotion identification

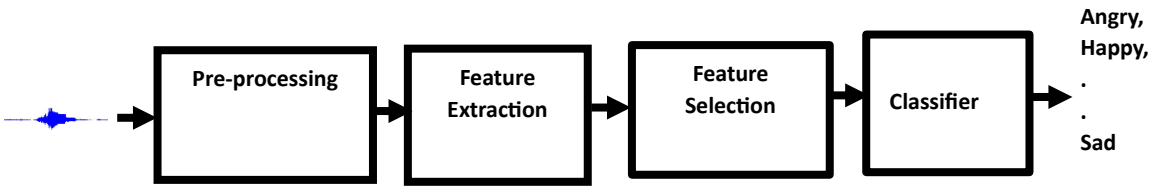

Figure 1: Steps of classical Speech Emotion Recognition

Feature selection selects features from extracted features to reduce processing time. Acoustic speech features can be qualitative, continuous, spectral, and TEO-based. The taxonomy of the acoustic features is given in Fig 2. Most of the early works focused on feature-level speech processing with statistical analysis. Machine learning algorithms like HMM, GMM, SVM, DECISION TREE, etc. can also be seen as part of the literature in SER. Emotion detection from speech with machine-learning approaches becomes challenging when analyzing real-life non-basic emotions, problems like data annotation, and blended emotions [139]. Deep learning is the new state-of-the-art for artificial intelligence, providing state-of-the-art accuracy in many tasks. In the literature, we may find SER systems that use deep learning techniques like CNN, RNN, LSTM, and its variants. The creation of labeled datasets and the lack of standard datasets from various languages are obstacles in speech-emotion recognition systems. Transfer learning overcomes this problem by utilizing knowledge acquired for one task to solve related ones. Transfer learning [130], semi-supervised learning [102], and unsupervised learning [132] based SER systems are also part of the literature. The lack of

standardization techniques for annotation and labeling, feature extraction, feature selection, and model selection are still challenging in speech emotion recognition systems.

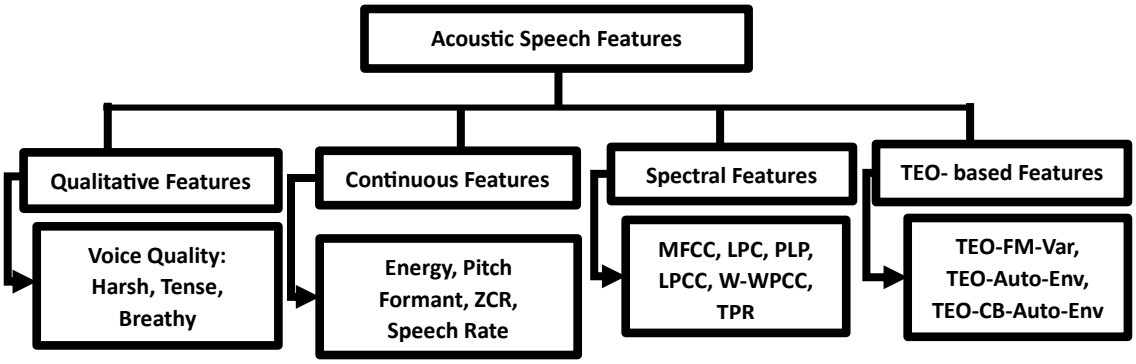

Figure 2: Taxonomy of acoustic speech features [133]

3. Noisy Speech Emotion Recognition

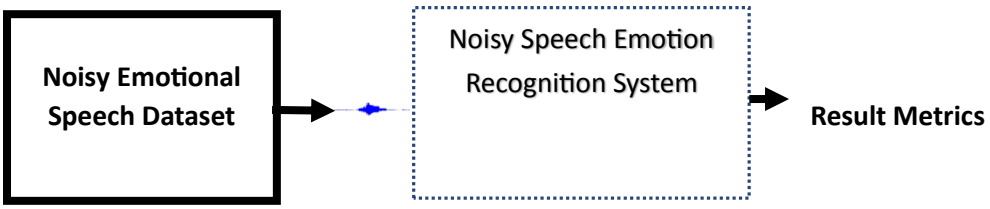

Figure 3: Abstract view of noisy Speech Emotion Recognition

Speech emotion recognition systems fail to get accurate results in audio acquired in a noisy or unconstrained environment. The abstract view of noisy SER is in Fig 3. Below, we give a brief review of speech emotion recognition methods in an uncontrolled environment based on the literature. Accurate and effective methods improve the robustness of SER systems. It is difficult to give a clear-cut taxonomy of all the work on noisy speech emotion tasks based on the literature. Hence, the proposed taxonomy in Figure 4 is a relative grouping of the methods in the literature. The algorithms in some groups may have overlapping properties.

3.1. Datasets

Speech emotion recognition datasets collected in uncontrolled or natural settings have also been part of the literature. Speech emotion datasets can be acted, simulated, and natural or spontaneous datasets based on the type of recording. Based on the type of object that participated, speech emotion databases can be grouped as young datasets, old or elderly datasets, or impaired objects datasets. Classification of speech emotion datasets based on recording and type of objects involved is in Figure 5. We summarize the main speech emotion recognition databases in the literature used for noisy speech emotion in Tables 3, 4, and 5, respectively.

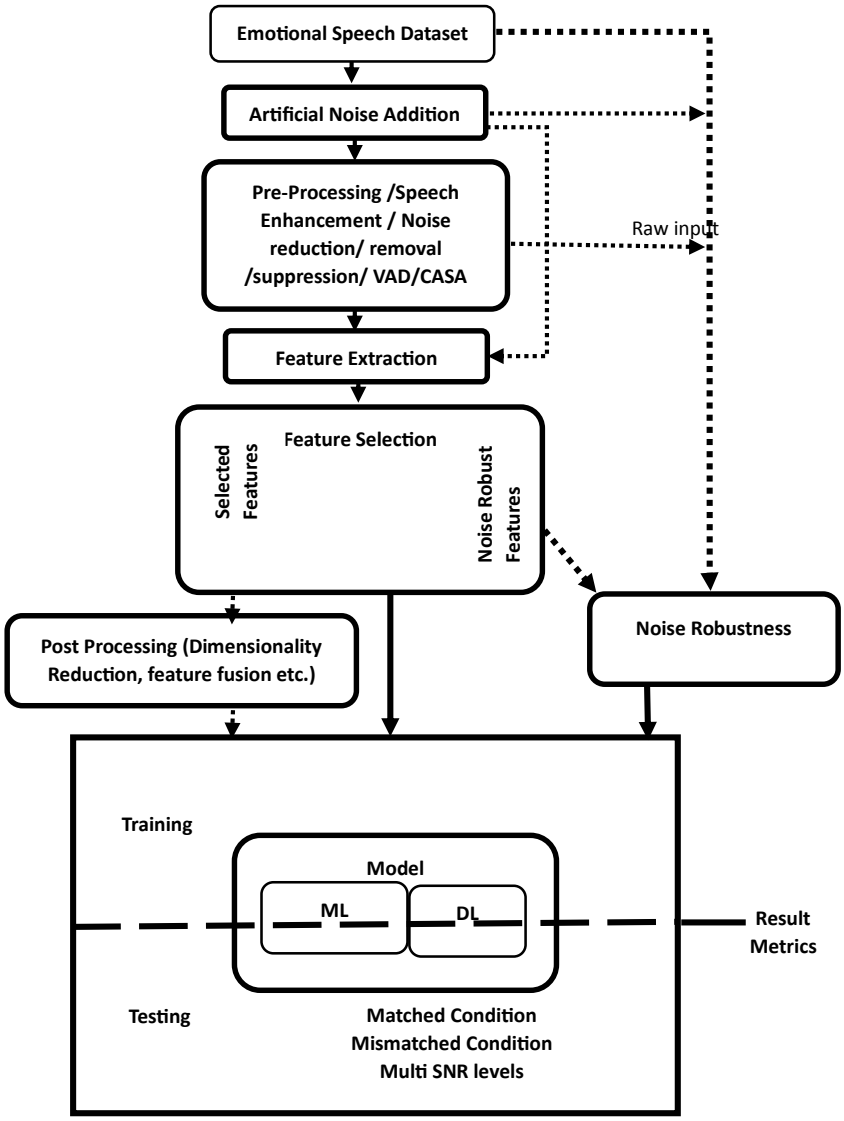

Figure 4: Summary of noisy speech emotion recognition systems based on literature (dotted line indicates some groups may have overlapping properties)

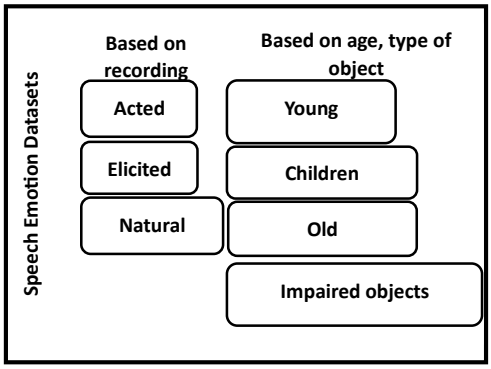

Figure 5: Classification of speech emotion datasets based on recording and type of object involved

Table 3: Emotion speech databases for speech emotion recognition published between 1996-2010 used in literature for noisy speech emotion recognition task.

Sl. No	Emotional Speech Database	Language	Reference	Number of Speakers	Type of database	Emotion Labels
1	Danish Emotional Speech Corpus (DES) [68] 1996	Danish	[113][118][127]	4(2 Males, 2 Females)	Acted	Anger, joy, sadness, surprise and neutrality
2	Speech Under Simulated and Actual Stress (SUSAS) dataset [62] 1997	English	[24][31][113]	32 speakers (19 Males, 13 Females)	Natural	Neutral, angry, slow, loud, soft, Lombard, fast
3	Korean database [69] 2000	Korean	[111][116]	10(5 Males, 5 Females)	Acted	Neutrality, joy, sadness, anger
4	Berlin Database of Emotional Speech (EMO-DB) [56] 2005	German	[17][18][19][23][27][28][30][34][35][36][40][43][44][46][47][112][113][118][119][125][126][127][128]	10 (5 Males, 5 Females)	Acted	Anger, boredom, disgust, fear, happiness, sadness, neutral
5	eNTERFACE [64] 2005	English	[36][33][112][125]	42 (34 Males, 8 Females)	Elicited	Anger, boredom, fear, joy, sadness, neutral, disgust
6	Interactive Emotional Dyadic Motion Capture (IEMOCAP) [51] 2008	English	[11][12][13]16][18][22][23][34][35][41][44][45][46][48][49]	10 (5 Males, 5 Females)	Elicited	Anger, happiness, excitement, sadness, frustration, fear, surprise, neutral state
7	Vera am Mittag (VAM) corpus [65] 2008	German	[43]	47 (11 Males, 36 Females)	Acted	Valence, arousal, dominance
8	Sahand Emotional Speech (SES) database [85] 2008	Persian	[124]	10(5 Males, 5 Females)	Acted	Anger, neutral, happiness, sadness, surprise
9	Polish Emotional Speech [70] 2009	Polish	[123]	8(4 Males ,4 Females	Acted	Anger, joy, sadness, fear and boredom
10	FAU-Aibo [52] 2009	German	[14]	51 children	Natural	Anger, emphatic, neutral, Joy, rest.
11	Mandarin Emotional Speech database (MES) [86] 2010	Chinese	[127]	7 Actors	Acted	Anger, joy, surprise, sadness, disgust

Table 4: Emotion speech databases for speech emotion recognition published between 2013-2017 used in literature for noisy speech emotion recognition task.

Sl. No	Emotional Speech Database	Language	Reference	Number of Speakers	Type of database	Emotion Labels
1	REmote COllaborative and Affective interactions (RECOLA) database [61] 2013	French	[29][36][39][121]	27 (11 Males, 16 Females)	Natural	Arousal, valence
2	Surrey Audio-Visual Expressed Emotion (SAVEE) [60] 2014	English	[23][28][127]	4 Males	Acted	Anger, disgust, fear, happiness, sadness, surprise and neutrality
3	EmotAsS [53] 2017	English	[12]	Speech of cognitively impaired subjects	Natural	Angry, happy, sad and neutral
4	MSP-IMPROV [54] 2017	English	[11]	12 (6 Males, 6 Females}	Natural	Angry, neutral, sad, and happy
5	Chinese natural audio-visual emotion database (CHEAVD 2.0) [55] 2017	Chinese	[16]	238	Natural	Happy, anxious, disgust, sad, worried, neutral, surprise, angry
6	NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME) [59] 2017	Chinese	[26]	44(22 Males, 22 Females)	Natural	Angry, sadness, happiness, frustration, neutral and surprise

Table 5: Emotion speech databases for speech emotion recognition published between 2018-2021 used in literature for noisy speech emotion recognition task.

Sl. No	Emotional Speech Database	Language	Reference	Number of Speakers	Type of database	Emotion Labels
1	Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [57] 2018	English	[19][22][23][28][31] [32] [48]	24 (12 Males, 12 Females)	Acted	Calmness, happiness, sadness, anger, fear, surprise, disgust, and neutral
2	The Chinese Academy of Sciences' Institute of Automation (CASIA) [58] 2018	Chinese	[20][28]	4 Speakers	Natural	Anger, fear, neutral, happy, sad and surprise
3	Low Quality Emotional dataset (LQ Emo Dataset) [66] 2018	English	[47]	WhatsApp audio messages	Natural	Happy, angry, sad, and neutral
4	MSP-Podcast [137] 2019	English	50	Podcast recordings	Natural	Valence (negative versus positive), arousal (calm versus active) and dominance (weak versus strong)
5	DEMoS [140] 2019	Italian	[25]	68 (45 Males ,23 Females)	Natural	Anger, disgust, fear, guilt, happiness, sadness, and surprise.
6	Toronto Emotional Speech Set database (TESS) [67] 2020	English	[32]	2 Females	Acted	Anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral)
7	Novel Sindhi-Speech Emotion dataset (NSSED) [42] 2021	Sindhi	[42]	29 (16 Males, 13 Females)	Natural	Happy, Sad, Neutral, and Angry.
8	ESD [141] 2021	Emirati	[24]	50 (25 Males, 25 Females)	Elicited	Sad, neutral, fear, disgust, angry, and happy

3.2. Artificial Noise Addition

Most of the speech emotion datasets were developed under controlled, noise-free conditions or with specific forms of environmental noise. It can be cleaned or recorded from different environmental conditions. It is necessary to consider the effects of noise in speech recognition systems when considering real-time, end-to-end speech emotion recognition works. To overcome these limits, in literature, we can see the practice of artificially adding different noise types to the existing datasets(See Figure 6), which reduces the cost and effort of creating noisy datasets and also lead to datasets with multi-noise conditions. Noise, reverberation, distance from the speaker to the microphone, and conditions of the recording device affect the quality of the speech signals. Noise means additive Gaussian white noise, stationary noise, non-stationary noise, and environment noise like restaurants, waterfalls, etc. List of noise used in literature as part of noisy SER is in Table 6.

Additive White Gaussian Noise (AWGN): It is necessary to consider the effects of noise on speech emotion recognition. To the signal, this form of noise can be added (by adding arithmetic elements one by one). Additionally, its mean value is zero (randomly sampled from a Gaussian distribution with a mean value of zero; standard deviation can vary). It equally includes each frequency component. AWGN is simpler to model and produce.

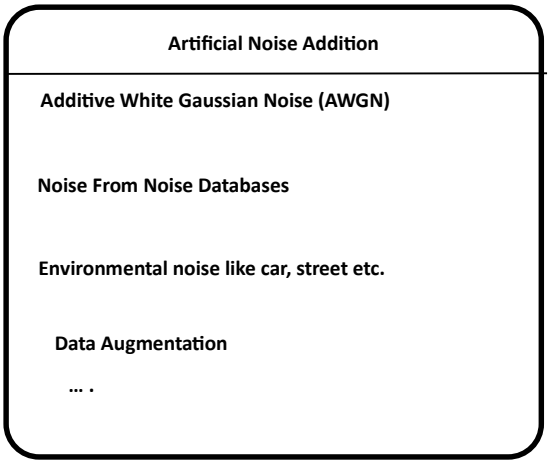

Figure 6: Artificial noise contamination

Reverberation Effect: The reverberation effect depends on the speech signal generation environment. In literature, we can see the simulation of noisy datasets by contaminating existing clean datasets with environment noise like restaurants, cafeterias, and waterfalls at various SNR levels to study the reverberation effect.

Table 6: List of noise used in literature for speech emotion recognition under noisy conditions

Type of Effect	Noise
Reverberation Effect at various SNR levels	Additive Gaussian White Noise (AGWN), sinusoid noise, non-stationary noise, stationary noise, noise from natural datasets, noise from noise datasets, Data augmentation, randomised adversarial data, signal amplification, silence, Laughter, breathing, shout, short pauses, transitions between phonemes, unvoiced phonemes, Under different environmental noise (like car-noise, babble noise, train noise, street noise, talking people, vacuum cleaner, bubble noise, Volvo noise, engine sound, exhibition, restaurant noise, kitchen, park, station, traffic, cafeteria, movie scenes, laboratory, playground, mall, sport event, factory noise, high frequency radio channel noise, fighter jet noise etc.)
Codec Effect	Effect of channels, codec etc.
Micro-Phone distance	Close-talk, 'room microphone', close-talk reverberated

Codec Effect: The effect of different recording devices, channels, bit-rate, and compression techniques also affect the speech quality, hence has a role in speech emotion recognition

Microphone Distance: The distance between the microphone and the speaker also determines the quality of speech signals, which affects the accuracy of speech emotion recognition tasks. To study this effect, the emotion speech database records at different distances from the microphone

Signal-to-Noise Ratio levels: Modified or contaminated speech signals from a clean environment can be generated artificially by adding noisy signals at different SNR levels. It is calculated using the equation (1).

$$SNR = 10 \log_{10} \left(\frac{Ps}{Pn} \right) \quad (1)$$

Ps is the power of clean speech signals and Pn is the power of noise signals

3.3. Noise Datasets

Different noise from various noise datasets mixed up with speech emotion datasets to generate noisy speech emotion datasets. The noise datasets used emotion recognition under noisy conditions based on the literature are in Table 7.

Table 7: Noise dataset used for speech emotion recognition to create noisy speech emotion datasets.

Sl. No	Noise Dataset	Noise Types	Reference
1	Noisex-92[96] 1993	Machine-gun noise STITEL babble, Lynx-helicopter cockpit noise, Fast-jet cockpit noise, Car noise, Factory noise, Office noise etc.	[35]126]
2	Aurora noisy database [93] 2000	Suburban train, Crowd of people (babble), Car, Exhibition Hall, Restaurant, Street, Airport, Train station etc.	[18][124]
3	SPIB [95] 2002	Speech babble, factory floor noise, cockpit noise, military vehicle noise etc	[28]
4	DEMAND (Diverse Environments Multichannel Acoustic Noise Database) [91] 2013	Domestic: kitchen, living room, washing machine, Office: office, meeting room, small office, Public: cafeteria, restaurant, busy subway, Transportation: transit bus, private passenger vehicle, a subway, Nature: sports field, city park, running water, Street: terrace of a café, public town s, a busy traffic etc	[14]
5	ESC-50[94] 2015	50 types of environmental noise and each type of noise contains 40 audio snippets	[22]
6	Muscan Corpus [98] 2015	Consists of three portions: music, speech and noise. The noise portion contain technical noises, such as DTMF tones, dial tones, fax machine noises, as well as ambient sounds, such as car idling, thunder, wind, footsteps, paper rustling, rain, animal noises, etc	[33]
7	Audio Set [97] 2017	Human sound, Animal Sounds, Nature Sound, Music, Sound of things, source-ambiguous sounds, Channel-environment and background sound.	[36]
8	CHiME-4[92] 2017	Five locations (i.e., booth, on the bus, cafe, pedestrian area, and street junction)	[16]

3.4. Noisy Speech Emotion Recognition Methods

Noisy speech emotion recognition methods help to process noisy speech emotion data for the classification or regression analysis of the system. The first approach followed in literature is the method of speech enhancement pre-processing to enhance or remove the noise from raw speech signals or calculated features. The second approach is the identification of noise-robust features. The third approach is to analyze the noise robustness of the SER models by artificially injecting different types of noise at various SNR levels. The summary and comparison of the main noisy speech emotion recognition methods are given in Tables 8 to 14

3.4.1. Pre-Processing/Enhancement

The summary and comparison of the main pre-processing algorithms are given in Figure 8, Tables 8 and 9 respectively. Pre-processing is used to enhance the quality of the speech signals from the noise, to remove, suppress, or reduce the noise effects, or to identify the parts of emotional speech relevant to the SER task. Speech enhancement help to improve speech quality and intelligibility of the corrupted speech signal. Speech enhancement algorithms can be identified from the literature. Speech enhancement algorithms like spectral subtraction, wiener filter, MMSE[126], speech enhancement based on the masking properties and short-time spectral amplitude estimation[130], single-channel spectral enhancement method (SSE)[29], band pass filtering and spectral subtraction[15], simply integrating the enhancement as a pre-processing step in the testing phase[36], MMSE as the enhancement technique[49] and optimally modified log-spectral amplitude estimator (OMLSA) based audio enhancement using spectral subtraction, wiener filter and MMSE[42] reviewed in this paper. Speech enhancement techniques for noise cancellation or reduction from literature can be listed as adaptive noise cancellation based on adaptive thresholding in wavelet domain [119], noise

suppression technique based on adaptive RLS filtering [126], three-level noise reduction algorithm with data downsampling, feature synchronization, and a modified version of graph total variation (GTVR) [39] and pre-processing with filters to remove noise using Butterworth and Chebyshev filters [21].

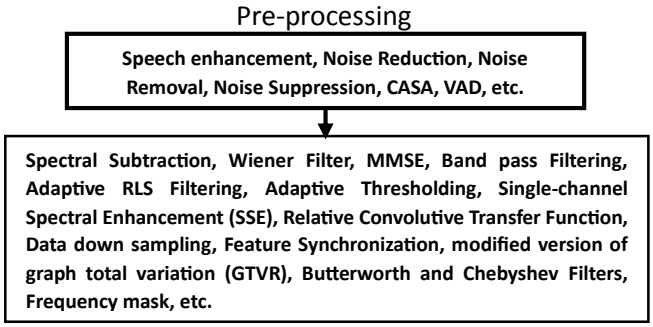

Figure 8: Pre-processing Algorithms

Table 8: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2010-2016(Pre-processing/ Enhancement)

Paper	Database	Emotions	Noisy Dataset/ Types of noise	Features, Method, Classifier, SNR Level	Result
Tawari et al. [119] 2010	EMO-DB, LISA-AVDB	EMO-DB: 7 emotions LISA-AVDB: pos, neg and neu	White Gaussian noise, Car noise (highway, parking, city street)	**Features**: prosodic speech intensity, pitch and speaking rate, and spectral features (MFCC) **Method**: Adaptive noise cancellation based on adaptive thresholding in wavelet domain **Classifie**r: SVM **SNR**:15 ,10 ,5 DB	High recognition accuracy of noisy speech at higher SNR for LISA-AVDB
Revathy et al. [126] 2015	EMO-DB	Happy, angry, anxious, fearful, bored, disgusted, neutral	Noises Volvo, white and F16 from "Noisex-92" database	**Features**: Short-time energy and zero crossing rate **Method**: Noise suppression technique based on adaptive RLS filtering **Classifier**: HMM	Reduces the noise without changing the speech frequencies.
Chenchah et al. [129] 2016	IEMOCAP	Angry, happy, neutral, sad.	Airport, car, babble noise	**Features**: MFCC **Method**: Three speech enhancement algorithms: Spectral subtraction, wiener filter, MMSE **Classifier:** HMM **SNR**: 0,5,10,15 dB	Spectral subtraction and MMSE enhance the recognition rate in airport and babble noise, not efficient to reduce car noise.
Huang et al. [130] 2013	Chinese speech emotion database(custom)	Happy, sad, angry, surprise, fear, anxiety, hesitation, confidence, neutral arousal, valence	Additive Gaussian white Noise	**Features**: max, min, mean, std, range of pitch and dev pitch 10–11 Jitter, Shimmer 12–52 max, min, mean, std, range of F1 to F4 and dev of F1 to F4 52–62 max, min, mean, std, range of intensity and dev intensity 62–192 max, min, mean, std, range of MFCC1 to MFCC13 and dev of MFCC1 to MFCC13 192–372 max, min, mean, std, range of BBE1 to BBE18 and dev of BBE1 to BBE18("BBE" stands for Bark Band Energy) **Method**: speech enhancement based on the masking properties and short-time spectral amplitude estimation. **Classifier**: GMM	Increasing noise level, the overall classification rate dropped

Table 9: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2017-2023(Pre-processing/ Enhancement)

Paper	Dataset	Emotions	Noise	Features, Method, Classifier, SNR Level	Main Result
Avila et al. [29] 2018	RECOLA	Arousal, valence	Background noise and reverberation	**Features**: eGeMAPS **Method**: Single-channel spectral enhancement method (SSE), reverberation is modeled by an STFT domain moving average (MA) model using a relative convolutive transfer function (RCTF) **Classifier**: Quality test **SNR**: 0 dB 5 dB 10 dB 15 dB 20 dB	Severely affected by background noise, especially for lower SNR
Jain et al. [15] 2018	Hindi language speech	Happy, sad, anger, neutral.	YouTube the voice recordings	**Features**: prosodic and spectral features of an acoustic signal and Linear Prediction Cepstral Coefficient (LPCC) **Method**: Band pass Filtering and spectral subtraction **Classifier**: Cubic SVM	Provided an efficiency of 96.30% for male 90.60% for female samples with 6 features
Jing et al. [39] 2018	RECOLA	Arousal, valence	Reverberation noise, CHiME	**Features**: eGeMAPS and MFCCs Method: Three-level noise reduction algorithm: data down sampling, feature synchronization, modified version of graph total variation (GTVR) **Regression**: partial least square regression (PLSR) **SNR:** 0.3.6,9,12 dB	GTVR method not only improves the performance of emotion prediction on noisy data but also yields higher CCC
Koduru et al. [21] 2019	RAVDESS	Angry, happy, sad, neutral	RAVDESS	**Features**: Pitch, energy, zero crossing rates, Mel Frequency Cepstral Coefficients and Discrete Wavelet Transform **Method:** Pre-processing with filters to remove noise (Butterworth and Chebyshev filters) **Classifiers**: SVM, LDA, Decision Tree	Suitable for all kind of signals and gives better speech recognition rate and improves the accuracy and efficiency of the system
Triantafyllopoulos et al. [36] 2019	RECOLA, EMODB, eNTERFACE	RECOLA: arousal Emo-DB :6 emotions eNTERFACE: 7 emotions	Mozilla Common Voice database, Audio Set	**Features**: RECOLA: Raw signal Emo-DB, eINTERFACE: ComParE, eGeMAPS **Method**: Simply integrating the enhancement as a pre-processing step in the testing phase **Classifier**: DNN **SNR**: -5,0,5,10,15,20 dB	Very low and negative SNRs in particular, able to render the SER algorithms usable again.
Alghifari et al [47] 2019	EMO-DB	Anger, boredom, disgust, fear	LQ Emo Dataset	**Features**: MFCC **Method**: Voice activity detection (VAD)pre-processing **Classifier**: Deep feedforward neural network classification	Recognition rate greatly improved
Zhou et al. [16] 2020	IEMOCAP and CHEAVD 2.0	Happy, sad, angry, neutral	CHiME-4	**Features**: Spectograms **Method**: LSTM-PL-MTL architecture with ISPP-based post-processing **Classifier**: LSTM	Achieve performance improvement over unprocessed noisy speech.
Win et al. [49] 2020	IEMOCAP	Anger, boredom, disgust, fear, joy, neutral, sad	Sport event noise	**Features**: MFCC **Method:** Minimum Mean Square Error, MMSE is used as the enhancement technique **Classifier**: Neural Network, SVM **SNR**:0,5,10,15,20 dB	Deep Learning gives the superlatives performance for angry emotion of noisy signal
Laghari et al. [42] 2021	NSSED	Happy, sad, angry, neutral	Stationary noise	**Features**: Mel-spectrogram features **Method:** optimally modified log-spectral amplitude estimator (OMLSA) based audio enhancement using spectral subtraction, wiener filter and MMSE **Classifier**: 1-Dimensional Convolution Neural Network (1DCNN) Model **SNR**: 0.5, 1.5 dB	Enhance recognition rate in airport and babble noise, and not efficient to reduce car noise.
Nassif et al. [31] 2021	SUSAS, ESD, RAVDESS	Neutral, lombard, angry, fast, loud, slow, soft	noise signals in a ratio 2:1	Features: MFCC **Method**: A STFT based frequency mask for speech separation from the noise signal. **Classifier**: Gaussian Mixture Model – Convolutional Neural Network (GMM-CNN)	CASA-based pre-processing module for co-channel noise reduction.
Hamsa et al. [24] 2023	RAVDESS, SUSAS, ESD	Angry, happy, neutral, sad, fearful, disgust	RAVDESS, SUSAS speaking environments	**Features**: Amplitude modulation spectrogram (AMS), RASTA-PLP (Relative Spectral Phonetic Likelihood Probabilities) and MFCC features Method: Wavelet packet transforms (WPT) based filter bank for segregating noise and emotional speech data **Classifier**: Hybrid RBM-DNN-SVM	Results shows superiority of the proposed model in noisy conditions.

Voice activity detection (VAD)pre-processing is an important pre-processing method, which helps to identify human speech from other voices. Adding VAD in SER to identify silent features and to improve the accuracy of SER was performed by [47]. A Computational Auditory Scene Analysis (CASA) is a pre-processing method that segregates the dominant signal from other interference signals before performing the speech recognition task. CASA pre-processing STFT-based frequency mask for speech separation from the noise signal was performed by [31]. Wavelet packet transforms (WPT) based filter bank for segregating noise and emotional speech data [24] can also be identified in the literature as some pre-processing tasks in noisy speech emotion recognition to identify relevant feature information. The post-processing task after feature extraction is also part of the literary works. Post-processing with LSTM-PL-MTL architecture with ISPP [16] helps to extract more accurate feature selection for noisy speech emotion recognition systems

A speech enhancement model with emotional speech is more effective than one that has not with enhancement algorithms. The enhancement algorithms were able to make the SER algorithms usable once more in extremely low and negative SNRs. When compared to noisy audio with high SNRs, enhancement architecture might reduce audio quality and make SER algorithms perform poorly.

Table 10: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2006-2013(Robust Feature Selection)

Paper	Database	Emotions	Noisy Dataset/ Types of noise	Features, Method, Classifier, SNR Level	Result
Hyun et al [111] [2006]	Korean database	Neutral, joy, sad, anger	Signal amplification	**Features**: LFPR (log frequency power ratio)	Performance in noisy environments improved by approximately 10%
Kin et al. [116] 2007	Korean Database	Neutral, joy, sad, anger.	Additive noise	**Features**: eigen-FFT **SNR**: -3, -1.8, 0, 3, 6, 9, 12 and 18 dB	eigen-FFT displayed superior performance over LPC, MFCC, and pitch.
You et al. [117] 2007	Chinese Academy of Sciences database	Neutral, angry, fear, happy, sad, surprise	Gaussian white noise and sinusoid noise	**Features**: 48 prosodic and 16 formant frequency features **Method**: Enhanced Lipschitz Embedding (ELE) **Classifier**: SVM **SNR**: 5,10,15,20,25	Improvement of approximately 10% in emotion recognition accuracy
Schuller et al. [112] 2008	EMO-DB, eNTERFACE	EMO_DB: 7 eNTERFACE: 6 emotion labels	Car-noise, bubble noise, sound of MINI cooper	**Features**: Prosodic, Spectral features **Method**: Noise adaptation and speaker adaptation with noise specific feature selection (NSAFS) by Correlation-based Feature Subset Selection (CFSS) **Classifier**: SVM **SNR**: -30, -20, -10 ,0,10,20,30 dB	Recognition accuracies degraded but can be improved by proposed method
Georgogia nnis et al. [114] 2012	EMO-DB	Angry, happy, neutral, sad disgust, fear, boredom	Pink and white noise to the test set	**Features**: Teager-energy based Mel-frequency cepstral coefficients (TEMFCCs) **Method**: Gaussian mixtures models **SNR**: 0, 10, 20, 30, 40, 50 dB.	TEMFCC features more robust than MFCCs
Han et al. [115]2012	Chinese (Private)	Angry, disgust, fear, joy, neutral, sad, surprise	Multi-channel recording, environment noise	**Feature**s: 37 prosody features and 16 quality features, **Method**: Spectral transformation adaptation method with Canonical Correlation Based on Compensation (CCBC) **Classifier**: Back-propagation Neural Networks	Improves recognition rates and robust performance even in low SNR case
Karimi et al [124] 2013	EMODB, Sahand Emotional Speech (SES) database	EMODB: 7 emotional categories, SES: 5 emotional classes	babble noise from Aurora database with different signal to noise ratios	**Features**: Feature based on filters and wrapper methods **Method**: Voice activity detection (VAD) using SFFS (sequential floating forward selection (SFFS). **Classifier**s: Bayes and SVM **SNR**: −10, −5, 0, 5, 10 dB	VAD improves system performance

Table 11: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2014-2016(Robust Feature Selection)

Paper	Database	Emotions	Noisy Dataset/ Types of noise	Features, Method, Classifier, SNR Level	Result
Mao et al. [127] 2014	SAVEE Emo-DB, DES, MES	Categorical emotion labels	Gaussian noise	**Features**: Affect-salient features **Method**: optimal feature set for SER are learned automatically by CNN through two-stage training: SAE and SDFA **Classifier**: CNN **SNR**:20 dB	Method SDFA achieves the highest and the most stable accuracy
Zao et al. [128] 2014	EMO-DB, Polish database	EMO-DB: 7 emotions Polish Database: 7 emotions	Spontaneous dataset	**Features**: Prosody features, voice quality features, spectral features **Method**: Maximum likelihood estimation (MLE) to formulate a weighted sparse representation **Classifier**: SVM **SNR**: 30, 25, 20, 15, 10 ,5, 0, -5, -10 dB	Enhanced sparse representation SVM obtains the best performance on the clean and noisy emotional speech
Song et al. [125] 2016	EMO-DB eNTERFACE FAU Aibo	case1, case2: 5 emotion labels case3: neutral, non-neutral	Cross-corpus speech emotion recognition	**Features**: Loudness, MFCC [0–14], Log Mel frequency band [0–7], LSP [0–7], F0, F0 envelope Voicing probability, Jitter, Shimmer **Method:** transfer non-negative matrix factorization, with NMF and transfer learning algorithms transfer GNMF (TGNMF) and transfer CNMF (TCNMF) **Classifier**: SVM	Both TGNMF and TCNMF outperform the existing state-of the-art approaches
Aher et al. [122] 2016	EMO-DB	Boredom, angry, happy, disgust, sad, fear, neutral.	cultural and environmental differences, additive noise	**Features**: prosodic feature (energy and pitch) **Method**: Cochlear filter bank coefficients combined with prosodic feature **Classifier**: SVM **SNR**:15, 10, 5,0 dB	Prosodic feature gives more recognition accuracy compared to MFCC in mismatched condition.

3.4.2. Noise Robust Feature Extraction

For speech emotion recognition tasks, various frequency domain characteristics may exhibit variable robustness for noisy input. In the literature, there is ongoing research on finding noise-robust features. Some features are more resistant to noise and reverberation than others. Tables 10, 11, and 12 provide a review and comparison of the primary noise-robust feature selection algorithms.

Table 12: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2017-2023(Noise Robust Feature Selection)

Paper	Dataset	Emotions	Noise	Features, Method, Classifier, SNR Level	Main Result
Chenchah et al. [41] 2017	IEMOCAP	Angry happy, sad, neutral	Car, babble, train, airport	**Features**: MFCC coefficients, PLP coefficients, Power Normalized Cepstral Coefficients **Classifier**: Hidden Markov Models SNR: 0 dB, 5 dB, 10 dB and 15 dB	PNCC enhance robustness of the system
Huang et al. [40] 2017	Emo-DB	Angry, boredom, fear, joy, neutral, sad	White Gaussian noise to test data	**Features**: sub-band spectral centroid weighted wavelet packet cepstral coefficients (W-WPCC) **Method**: WP tree structure is adopted to generate WP filter bank. **Classifier**: Importance weights support vector machine IW-SVM SNR: 30-25,20-15,10-5 dB	Better noise-robustness in noisy environments

Luo et al. [12] 2018	IEMOCAP	Angry, happy, neutral, sad	EmotAsS	**Features**: Spectrograms and hand-crafted features **Method**: Joint representation learning **Classifier**: CRNN	Outperform the baseline plain CRNN SER system.
Liu et al. [20] 2018	CASIA	Angry, fear, neutral, happy, sad, surprise		**Features**: Hyper-prosodic feature and prosodic feature **Method**: CNN with spectrogram input fused with hyper prosodic features used as input vector to DNN **Classifier**: DNN VGG NET **SNR**:35 dB	This method improves the performance of SER system
Flores et al.[43] 2018	Vera am Mittag (VAM) corpus, EMO-DB	Quality assessment	Environmental noise, engine sound	**Features:** Signal power **Method**: By calculating the Signal-to-Noise Ratio and the Compression Error Rate	For EMODB the average signal quality was indicated lower as for VAM.
Huang et al. [45] 2018	IEMOCAP	Angry happy, sad, neutral	Noise is injected into all hidden layers	**Features**: Features of INTERSPEECH 2009 Emotion Challenge **Method**: Semi supervised learning with ladder networks **Classifier**: Deep DAE with last layer of encoder is attached with SVM	Has 2.6% higher performance than denoising auto-encoder, and 5.3% than the static acoustic features.
Huang et al. [27] 2019	EMO-DB	Angry, boredom, disgust, fear, joy, neutral, sad	White Gaussian noise to test data	**Features**: Prosody feature (F0, power), voice quality features (first, second and third formants with their bandwidths) and wavelet packet Cepstral coefficients (WPCC) **Method**: Combining the sub-band energies with sub-band spectral centroids via a weighting. The features fused by deep belief networks (DBNs). **Classifie**r: Support vector machine **SNR**: 10 dB	Deep learning performed better than the conventional systems using SVM as classifiers
Sekkate et al. [46] 2019	EMO-DB, IEMOCAP	EMO-DB:7 IEMOCAP: 7 emotions	Airport, train, babble, street, car, exhibition, restaurant noise	Features: Fusion of MFCCs derived from Discrete Wavelet Transform (DWT) sub-band coefficients (DMFCC), and pitch-based features **Method**: Feature fusion **Classifie**r: NB and SVM, DL	Implementing DL for SER, lack of availability of large datasets.
Latif et al. [11] 2020	IEMOCAP, MSP-IMPROV	Angry, happy, neutral, sad	DEMAND (Kitchen, park, station, traffic, cafeteria)	Features: Spectrograms Method: Data Augmentation (Mix up augmentation, speed perturbation (SP)), Adversarial attacks (Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM) Classifier: DenseNet, LSTM and deep neural network SNR:0, 10, 20 dB	The performance of proposed technique IEMOCAP and MSP-IMPROV more robust compared to existing methods and other state-of-the-art models.
Huang et al. [13] 2020	IEMOCAP	Angry, happy, sad, neural,	Gaussian noise	Features: INTERSPEECH 2014 Computational Paralinguistics Challenge Method: Pooling Method (NetVLAD as trainable discriminative clustering to aggregate frame-level descriptors into a single utterance-level vector), unigram label smoothing Classifier: LSTM	Combination of the NetVLAD and unigram label smoothing boosts the performance
Leem et al. [50] 2021	MSP-Podcast	Arousal Valence Dominance	Noisy version of the MSP-Podcast, non-stationary noise	Feature: Interspeech 2013 Computational Paralinguistics Challenge (ComParE) Method: Decoupled ladder network (DLN) Classifier: Autoencoder SNR: 10, 5, 0 dB	DLN can enhance the prediction of arousal by 11.4%, 8.4%, 10.2%, and dominance by 17.1%, 13.2%, 7.0%,
Abdelhamid et al [23] 2023	IEMOCAP, Emo-DB, RAVDESS, SAVEE	IEMOCAP- 8 EMODB -7 IEMOCAP-6 , SAVEE-7 emotion labels	Data augmentation to training samples	Features: Log-mel spectrogram Method: Data Augmentation, optimization stochastic fractal search-guided whale optimization algorithm (SFS-Guided WOA) Classifier: CNN+LSTM deep neural network	Recognize speech emotions of the adopted four datasets accurately.

Table 13: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2006-2018(Noise Robustness)

Paper	Database	Emotions	Noisy Dataset/ Types of noise	Features, Method, Classifier, SNR Level	Result
Schuller et al. [113] 2006	DES, EMO-DB, and SUSAS	Angry, disgust, fear, happy, neutral, sad, boredom, surprise	White noise addition to DES and EMO-DB	**Features**: intonation, intensity, formants, HNR, MFCC, and VOC19. **Method:** Fast Information-Gain-Ratio filter-selection for selecting from large feature set and **Classifier**: SVM **SNR**: ∞, 20, 10, 0, -5, -10 DB	SUSAS shows noise robustness. Other datasets suffer with noise
Sztahó et al. [120] 2011	Noisy Telephone Speech Database	Neutral, angry/nervous, happy, laughing, sad.	Noisy database	**Features**: F0, intensity, MFCC **Method**: Speech detection, phrase segmenting, **Classifier**: SVM	F0, intensity, MFCC have a role in speech emotion detection
Trigeorgis et al. [121] 2016	RECOLA	Valence, arousal	Spontaneous Database	**Features**: Raw wav **Method**: Convolutional recurrent model	Automatic features more robust than traditional hand-crafted features
Juszkiewicz et al. [123] 2016	Database of Polish Emotional Speech	Neutral, joy, sad, angry	Sounds of talking people, vacuum cleaner etc.	**Features**: Averaged histograms of pitch and MFCC **Method**: Feature normalization with Histogram equalisation technique to reduce the difference between feature vectors in clean and noisy conditions **SNR**: ∞, 30, 20, 10, 0dB	Recognition accuracy highly depends on type of noise, vacuum cleaner noise results worse.
Mansour et al. [30] 2017	Emo-DB	Angry, happy, sad, neutral	Real airport noise using various SNR levels	**Features**: Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) and MFCC-Shifted-Delta-Cepstral (SDC) coefficients **Method**: Robustness check **Classifie**r: HMM **SNR**:0, 5, 10, 15dB.	MFCC-SDC is better performing in both clean and noisy environments
Xiaoqing et al. [38] 2017	EMO-DB	Angry, fear, happy, neutral, sad	Gaussian white noise	**Features:** Pitch, energy, duration, formants, and mel frequency cepstrum coefficients (MFCC) and their statistics parameters **Method**: Reconstruction of samples removes added noise **Classifier:** multiple-kernel learning (MKL) SVM **SNR**: 20,15, 10 dB	The acoustic features extracted from reconstructed speech is robust to noise
Huang et al. [33] 2017	eNTERFACE'05	Angry, disgust, fear, joy, sad, surprise	MUSAN corpus	**Features**: log-Mel filter bank energies **Method**: Convolutional attention module into the vanilla CLDNN between the temporal module and the classifier **Classifie**r: CLDNN	Able to exploit the context information, exhibits noise robustness.
Alghifari et al. [17] 2018	Emo-DB	Angry, happy, sad, neutral	Customized noisy dataset	**Feature**s: MFCC **Method**: DNN	Training with the same dataset, 70.8% in best case performance
Hsiao et al. [14] 2018	FAU-Aibo	Angry, emphatic, neutral, positive, rest.	Spontaneous dataset	**Features**:16 low-level descriptors (LLDs) including 12 mel-frequency cepstral coefficients, root-mean-square energy, zero-crossing rate, harmonics-to-noise ratio, fundamental frequency and the corresponding delta features **Method**: Attention model **Classifier**: LSTM	The performance of 46.3% is among the best performance of FAU-Aibo tasks.
Chakraborty et al. [34] 2018	Emo-DB, IEMOCAP	EMO-DB -6 emotions IEMOCAP- 7 emotions	Voice babble, Factory noise, HF radio channel, F-16 fighter jets, and Volvo 340	**Feature**: MFCC **Method**: Feature compensation technique based on the Vector Taylor Series (VTS) expansion of noisy Mel-Frequency Cepstral Coefficients (MFCCs), **Classifier**: Time Delay Neural Network based Denoising Autoencoder (TDNN-DAE	Outperform NMF-based enhancement or even the NMF-VAD by a significant margin.

Table 14: A summary and comparison of main noisy speech emotion recognition methods in literature published between 2019-2023(Noise Robustness)

Paper	Dataset	Emotions	Noise	Features, Method, Classifier, SNR Level	Main Result
Kwon et al. [48] 2019	IEMOCAP, RAVDESS	Angry, neutral, sad, happy	noise and silent signals	**Features**: Spectrograms **Method**: Dynamic adaptive threshold technique to remove noise and silent signals **Classifier**: Deep Stride CNN (D SCNN)	Method achieves accuracy up to 79.5% using RAVDESS and 81.75% on an IEMOCAP dataset.
Tiwari et al. [35] 2020	Emo-DB, IEMOCAP	Happy, anger, neutral, sad	NOISEX-92	**Features**: 6552-dimensional feature vector that consists of HLDs (mean, standard deviation, skewness, kurtosis, extremes, linear regressions, etc.) and low-level descriptors (LLDs) ((zero crossing rate (ZCR), RMS energy, F0, HNR, MFCCs) **Method**: Mel-Filter Bank Energies (MFBEs) to design the Generative model. **Classifier**: DNN	The results obtained shows robustness to the SER system in unseen noise conditions.
Bandela et al. [18] 2021	EMO-DB, IEMOCAP	Angry, happy, neutral, sad	Aurora noise (airport, babble, car, station, street), Additive White Gaussian Noise	**Feature**s: INTERSPEECH 2010 paralinguistic features, Gammatone Cepstral Coefficients (GTCC) and Power Normalized Cepstral Coefficients (PNCC) **Method**: Non-negative Matrix Factorization (NMF). Unsupervised feature selection below with ordinal locality and FSASL (Feature selection with Adaptive Structure learning) **Classifier**: SVM RBF kernel **SNR**: -5,0,5,10,15,20 DB	Outperforms the baseline works both in clean and noisy environments.
Chatterjee et al. [32] 2021	RAVDESS, TESS(Young, Old)	Angry, disgust, fear, happy, neutral, sad, surprised	Recorded speeches and movie scenes	**Features**: MFCC **Method:** Spectral subtraction for background noise removal **Classifier**: ID CNN	90.48%, 95.79% and 94.47% classification accuracies for the standard datasets RAVDESS, TESS (Young and Old) respectively
Wijayasingha et al. [19] 2021	EMO-DB, RAVEDESS dataset	Calm, happy, sad, angry, fearful, surprised	Freesound Dataset, additive white Gaussian noise (AWGN)	**Features**: magnitude spectrograms, Modified Group Delay (MGD) spectrograms, unwrapped phase spectrograms **Method**: Attension based FCNN SNR:10,15,20,25,30,35, infinity	FCNN architecture with the attention mechanism handle noisy data for SER
Xu et al [22] 2021	IEMOCAP, RAVDESS	Angry, sad, excitement, neutral	ESC-50	**Features**: MFCCs **Method**: Head Fusion based on the multi-head attention-based convolutional neural network (ACNN) model **Classifier**: CNN	Improving Accuracy and Robustness of SER on IEMOCAP and RAVDESS Dataset.
Li et al. [44] 2021	IEMOCAP EMO-DB	Angry, happy, neutral, sad	silent regions, short pauses, transitions between phonemes, unvoiced phonemes	Features: MFCC, spectral roll-off point, spectral flux, spectral centroid, spectral entropy, spectral spread, zero-crossing rate, fundamental frequency, energy, energy entropy and their first-order difference **Method**: self-attention mechanism and LSTM **Classifier**: Bi-directional Long-Short Term Memory with Directional Self-Attention (BLSTM-DSA)	BLSTM-DSA has satisfactory results in terms of anger, neutrality and sadness recognition rate.

Hsu et al. [26] 2021	NNIME	Angry, frustration, sad, surprise, neutral, happy	Silence, Laughter, breathing, shout and background	**Features**: Magnitude spectral features **Method**: Segmentation method is used to decompose the input audio into verbal, non-verbal and silence/background segments. **Classifier**: LSTM based sequence-to-sequence model with attention mechanism.	Helps to infer information from the background present in real-life speech record
Li et al. [28] 2022	RAVDESS, Emo-DB, SAVEE, CASIA	Neutral, happy, sad, anger, boredom, disgust, fear	SPIB dataset, white gaussian noise, laboratory, playground, street, mall	**Features**: Log mel spectrogram **Method**: StarGAN is used to generate Log- Mel spectra **Classifier**: Dense-DCNN **SNR**:0, 4, 8, and 12 dB	Excellent classification performance in a noisy environment, adversarial experiment
Chang et al [25] 2022	DEMoS	Angry, disgust, fear, guilt, happy, sad, surprise	Randomised adversarial data	**Features**: Log Mel Spectrograms **Method**: Federated learning for data privacy, adversarial training at the training stage and randomisation at the testing stage for model robustness. **Classifier**: Deep Neural Network (VGG Architecture)	Achieved significant improvement under adversarial attacks
Mitra et al. [37] 2023	MSP-Podcast	Arousal, dominance, and valence	Noise, reverberation	**Features**: 40-dimensional mel filter bank energies appended with pitch, pitch-delta, and voicing features (MF BF0) **Method**: Bidirectional Encoder Representations from Transformers (BERT) and Hidden units BERT (HuBERT) **Classifier:** TC-GRU **SNR**: 20 to 30, 10 to 20, 0 to 10 dB	Models trained with HuBERT embeddings were relatively robust compared to MF BF0,

3.4.3. Robustness Check

Like other noisy speech emotion recognition methods, noise robustness checks also can be seen in the literature. The aim is to check how much the proposed system robust to work with the selected feature concerning noise. Tables 13 and 14 provide review of various algorithms for robustness improvement. Feature selection algorithms with fast Information-Gain-Ratio filter-selection from large feature set [113] and unsupervised feature selection below with ordinal locality and FSASL (Feature selection with Adaptive Structure learning) with Non-negative Matrix Factorization (NMF) [18], feature set optimization to adaptation to noise conditions [118], feature normalization with Histogram equalization technique to reduce the difference between feature vectors in clean and noisy conditions [123], feature compensation technique based on the Vector Taylor Series (VTS) expansion of noisy Mel-Frequency Cepstral Coefficients (MFCCs) [34] and dynamic adaptive threshold technique to remove noise and silent signals [48] are present in literature.

The attention mechanism has the ability to locate and focus on the salient or reliable parts of the speech signal. The attension models like the Convolutional attention module into the vanilla CLDNN between the temporal module and the classifier [33], Attension based FCNN [19], Head Fusion based on the multi-head attention-based convolutional neural network (ACNN) model [22], and Self-attention mechanism and LSTM to explore the autocorrelation of phonemes in utterance [44] are also part of robust speech emotion recognition systems. Due to the influence of transformer systems, Bidirectional Encoder Representations from Transformers (BERT) and Hidden units BERT (HuBERT) [37] were used to evaluate the robustness of the systems.

3.5. Models/Classifier

A classification or regression issue can be used to approach SER. Features must be provided as model input after being extracted and chosen as pertinent features. Both machine learning and deep learning models can be found in the literature. The literature from before 2016 demonstrates a growing use of machine learning models in noisy SER, with SVM being the most widely used model for generalizing speech emotion recognition. After 2016, the pattern changed as a result of the development of deep learning algorithms. The literature contains examples of deep learning algorithms such as CNN, fully connected CNN, LSTM, Deep stride CNN, Dense-DCNN, autoencoders, TC_GRU, CLDNN, and hybrid models combining deep learning with machine learning. To fully utilize the promise of additional deep learning architectures, more research in noisy SER is necessary. Mainly three strategies adopted by researchers in the case of noisy speech emotion recognition systems for model evaluation. During the testing phase, the noise robustness was evaluated with matched conditions, mismatched conditions, and multi-SNR levels.

Matched Conditions: Integrate the enhancement process on both the training and testing phases of the SER model with the fixed SNR conditions. Compare the performance of trained on enhanced audio and tested on enhanced audio with that of SER algorithms trained on noisy audio and tested on noisy audio.

Mismatched Conditions: Integrate the enhancement process only in the testing phase of the model. The training dataset consists of only clean speech samples while the testing dataset is a combination of noisy and clean datasets

Multi-SNR conditions: Integrate the enhancement process both in the training and testing phase of the model with unknown SNR conditions and compare the performance.

3.6. Toolkits

The feature extraction toolkits offer a broad range of speech-related functionalities concentrating on real-time feature extraction, contamination, segmentation, recognition, classification, and training. The list of toolkits from the literature is listed in Table 15.

Table 15: The Feature extractor toolkit used for emotion recognition under noisy conditions

Sl. No	Name of Toolbox	Purpose	Reference
1	YALMIP Toolbox [78] 2004	To solve semi-definite programming (SDP)	[38]
2	FANT Toolkit [76] 2005	Contamination of noise to clean speech	[34]
3	Hidden Markov Toolkit (HTK) [75] 2006	Training and testing of HMM classifier	[30][120][126]
4	WEKA [81]2009	Classification	[42][43][49][123]
5	openSMILE [71] 2010	Feature Extraction toolkit	[13][18][25][29][35][36] [39][43][44][45][125]
6	VOICEBOX [82]2010	Voice feature extraction toolbox for MATLAB	[47]
7	LIBSVM MATLAB Toolbox [74] 2011	SVM classifier	[27][38][120]
8	PRAAT [83] 2011	Analysing, synthesizing, and manipulating speech	[26][115][120][123][128]
9	Kaldi [84] 2011	Speech recognition toolkit	[34][35]
10	Audio Degradation Toolbox [80] 2013	Mixing clean and noise speech	[39]
11	Keras [77] 2015	Trained a Deep Neural Network (DNN) implemented in Keras	[35]
12	Librosa [73] 2015	Extracting MFCCs	[22]
13	scipy package [72] 2017	For extracting phase spectrograms	[19]
14	Adversarial robustness Toolbox [79] 2018	Toolbox for adversarial robustness research, contains various implementations for attacks, defences and robust training methods	[25]

3.7. Evaluation Metrics

The proper evaluation and analysis of speech emotion recognition under natural environments are essential. To measure the performance of speech emotion recognition under noisy conditions, several evaluation metrics have been suggested by the literature. The most commonly used evaluation metrics are accuracy, confusion matrix, precision, recall, and F1-Score. Table 16 shows the list of evaluation metrics

- UAR (Unweighted Average recall): It is used as a measure to indicate the performance of a conducted classification experiment over its mean average recall of one speaker. Afterward, these UARs per speaker were averaged over all speakers.
- UA (Unweighted Accuracy) refers to the average accuracy of each emotion category
- WA (Weighted Accuracy) refers to the accuracy of all samples
- Precision: Precision is calculated using the formula,

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+ False Positive}} \quad (2)$$

- Recall: Recall is calculated using the formula,

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+ False Negative}} \quad (3)$$

- Confusion Matrix: A table with all the predicted and actual values of a classifier
- F1 Score is calculated using the formula

$$\text{F1 Score} = \frac{\text{Precision}}{\text{Recall + Precision}} \quad (4)$$

- Accuracy: The percentage of accuracy is calculated with the formula

$$\text{Accuracy (\%)} = \frac{\text{Number of utterances that are correctly}}{\text{Total number of utterances}} \text{ X } 100 \quad (5)$$

- EER (Emotion Error Rate) is calculated with the formula

$$\text{ERR} = \frac{\text{Se + De + Ie}}{\text{Ne}} \quad (6)$$

  where Se is the number of emotion substitutions, De is the number of emotion deletions, Ie is the number of emotion insertions, and Ne is the number of emotions in the reference
- Compression Error Rate (CER): The difference between the two spectrograms and measures the absolute difference in dB between the spectrograms of the original and noisy signals.
- Concordance Correlation Coefficient (CCC). As a benchmark metric, it combines the Pearson correlation coefficient (CC) and the square difference between the means of the two samples.
- Mean Square Error (RMSE): The Mean Squared Error calculates how closely a regression line resembles a set of data points. It is a risk function that corresponds to the expected value of the squared error loss.
- Micro-averaging F1-score (Micro-F1): To calculate the Micro-F1 score, calculate the sum of all true positives, false positives, and false negatives over all the labels. Then compute the micro-

precision and micro-recall from the sum of all true positives, false positives, and false negatives over all the labels. Finally, compute the harmonic mean to get the micro F1-score

- Receiver Operating Characteristic Curve (ROC): It is a Deep Learning measure, which computes the performance of a deep learning technique by plotting the TPR (true positive rate), recall, or sensitivity against the FPR (false positive rate) and computing the area under it.

4. Discussion and Conclusion

4.1. Discussion

This survey reviews the literature on noisy SER environments and provides the main experimental results in noisy speech emotion recognition systems. It covers applications, a summary of earlier reviews, and general information about speech-emotion recognition systems in the literature. It then covers the steps of classical speech emotion recognition. This survey also reviewed noisy speech emotion recognition datasets, methods, noise datasets, types of noise, models, toolkits, results metrics used for evaluation, and tables used to compare results.

Table 16: Distribution of articles according to evaluation metrics

Sl. No	**Result Metrics**	**Reference**
1	Accuracy	[15][16][18][19][21][30][32][34][35][38][40][42][46][47][49][113][114] [115]117][118][119][122][123][124][130][125][126][127][129]
2	Confusion Matrix	[14][17][22][23][26][27][38][111][116][124][128]
3	F1-Score	[15][19][24][26][31] [42][48][49]
4	Precision	[15][20][24][26][28][31][42][49]
5	Recall	[15][20][24][26][28][31][42][49]
6	UAR	[11][22][25][36][43]
7	WA	[12][28][44][48]
8	UA	[12][13][33][44][48]
9	CCC	[29][37][39][50]
10	Mean squared Error	[50][121]
11	CER	[43][121]
12	EER	[41]
13	Micro-F1	[28]
14	ROC	[31]

The degree to which noisy SER can accurately identify objects depends on the type of noise used. The model performed better when pre-processing operations like speech enhancement, noise reduction, suppression, voice activity identification, segmentation, and segregation were carried out. Post-processing techniques like dimensionality reduction select main and significant features from large feature sets. It is possible to perform noise-robust feature recognition with or without pre-processing stages. The choice of suitable handcrafted acoustic features for noisy speech emotion recognition (SER) depends on the problem at hand. Performance measures are essential to ensure generalization within databases. The most extensively used databases for the evaluation of noisy speech emotion recognition are EMO-DB, RAVDESS, and RECOLA. The popularity of the datasets is shown in the image (Figure 9). EMO-DB is the database that is more popular in literature in the case of noisy speech emotion recognition systems. Figure 10 shows the language with more emotional speech datasets. English is on top, followed by Chinese, German, Danish, and Korean.

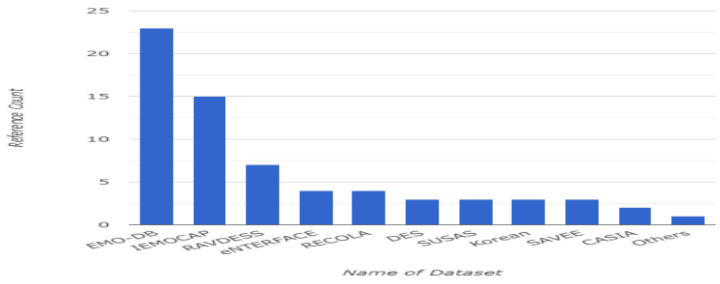

Figure 9: Frequency Distribution of reviewed articles according to database

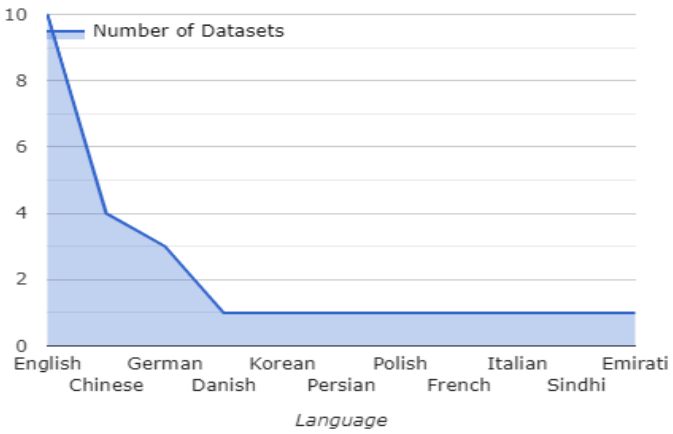

Figure 10: Datasets language trend based on literature reviewed

The majority of the works from the literature use AGWN as the type of noise in noisy SER. The environmental conditions tested by authors vary according to the researcher's interest. Not all environment noise, microphone effects, and codec effects are performed by authors in single work. The performance of one type of noise may be good concerning one model and method adopted. Not satisfactory when considering multiple noise conditions. Only a few datasets and studies based on the codec effect and microphone distance effect in the literature are available for review. The lack of datasets in many languages is still a challenging problem for the researcher. The literature shows only a few works and datasets from children, old, and impaired objects datasets. The majority of the datasets are also related to young speech signals.

Traditional SER systems are not effective for processing large amounts of data, and the SVM is the most widely used machine learning classifier due to its generalization capability. Deep learning architectures lead to automatic feature extraction and are more robust. Figure 11 and Figure 12 shows the frequency of models used before 2016 and after 2016 in the literature. Deep learning algorithms became more prominent due to their automatic feature extraction capabilities and noise robustness.

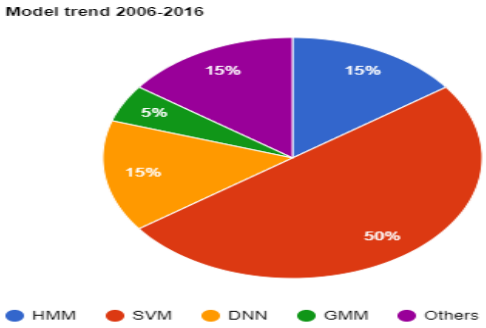

Figure 11: Classifier/Model trend before 2017 based on literature

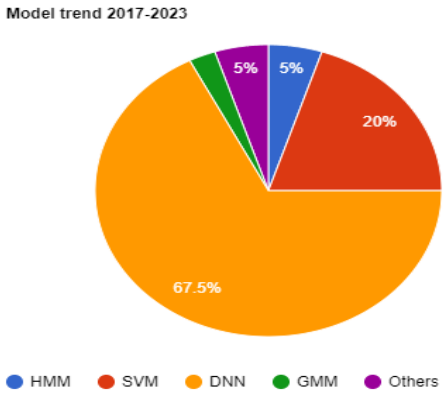

Figure 12: Classifier/Model trend after 2017 based on the literature

The most extensively used results metrics for noisy speech emotion recognition are accuracy, confusion matrix, F1-score, precision, and recall. The popularity of the result metrics is shown in Figure 13. Accuracy is the more popular result metric in literature in noisy speech emotion recognition systems.

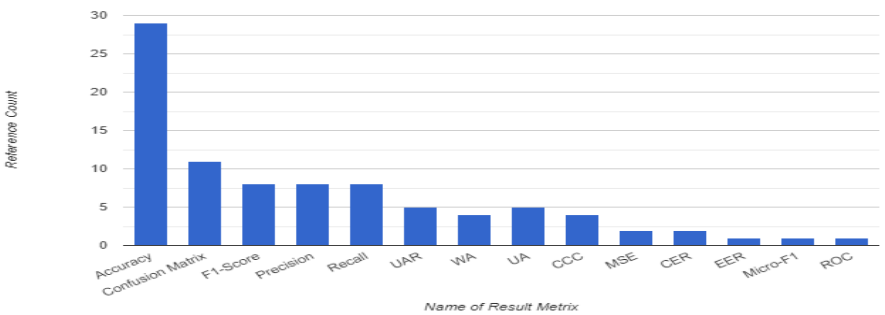

Fig 13: Popular evaluation Metrics based on the literature

4.2. Conclusion

This article has examined and contrasted the prior work on noisy SER. The field of research will continue to be busy. It was discovered that existing noisy SER research addresses a wide range of issues, including different types of noise, the generation of noisy datasets, a lack of large datasets,

cross-cultural, cross-corpus, cross-language, context-aware, automatic speech emotion recognition systems, more robust feature selection, and the implementation of new algorithms using deep learning techniques, which are likely to be the focus of noisy SER research in the future.

Future research in noisy SER can be conducted in various directions. Powerful and noise-resistant acoustic features for emotion classification have not received significant research. Faster speed enhancement techniques, precise model identification, and semi-supervised algorithms are required to extract more discriminative features. When there are insufficient high-quality data for SER research, adding noisy data should be used three times more frequently than adding clean data to avoid overfitting. The role of context, cross-language studies, incorporating more noise, adding cultural information, and speaking style become challenging due to the lack of large datasets and the cost of constructing labeled datasets for deep learning architectures. Observations presented in this article could be helpful for further architectural developments in the field of SER systems.

References

[1] Picard, Rosalind W. *Affective computing*. MIT press, 2000.

[2] Sapir, Edward, An introduction to the study of speech, *Language.* 1 (1921).

[3] Hossain, M. Shamim, Ghulam Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, *Information Fusion.* 49 (2019): 69-78.

[4] Han, Kun, Dong Yu, et al, Speech emotion recognition using deep neural network and extreme learning machine, *Interspeech 2014*. 2014.

[5] Wani, Taiba Majid, Teddy Surya Gunawan, et al, A comprehensive review of speech emotion recognition systems, *IEEE Access* 9 (2021). 47795-47814.

[6] Fahad, Md Shah, Ashish Ranjan, et al, A survey of speech emotion recognition in natural environment, *Digital signal processing*.110 (2021): 102951.

[7] Mustafa, Mumtaz Begum, Mansoor AM Yusoof, et al, Speech emotion recognition research: an analysis of research focus, *International Journal of Speech Technology.* 21 (2018): 137-156.

[8] Ververidis, Dimitrios, Constantine Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech *communication.* 48, no. 9 (2006): 1162-1181.

[9] Liscombe, Jackson, Giuseppe Riccardi, et al, Using context to improve emotion detection in spoken dialog systems. (2005)

[10] Tripathi, Suraj, Abhay Kumar, et al, Deep learning based emotion recognition system using speech features and transcriptions, *arXiv preprint. arXiv:1906.05681* (2019).

[11] Latif, Siddique, Rajib Rana, et al, Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition, *Proceedings of the 21st Annual Conference of the International Speech Communication Association. (INTERSPEECH 2020)*, vol. 4, pp. 2327-2331. University of Southern Queensland, 2020.

[12] Luo, Danqing, Yuexian Zou, et al, Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition, *Interspeech*. pp. 152-156. 2018.

[13] Huang, Jian, Jianhua Tao, et al, Learning Utterance-Level Representations with Label Smoothing for Speech Emotion Recognition, *INTERSPEECH.* pp. 4079-4083. 2020.

[14] Hsiao, Po-Wei, Chia-Ping Chen, Effective attention mechanism in dynamic models for speech emotion recognition, *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. pp. 2526-2530. IEEE, 2018.

[15] Jain, Udit, Karan Nathani, et al, Cubic SVM classifier based feature extraction and emotion detection from speech signals, *2018 international conference on sensor networks and signal processing (SNSP)*. pp. 386-391. IEEE, 2018.

[16] Zhou, Hengshun, Jun Du, et AL, Using Speech Enhancement Preprocessing for Speech Emotion Recognition in Realistic Noisy Conditions, *INTERSPEECH.* pp. 4098-4102. 2020.

[17] Alghifari, Muhammad Fahreza, Teddy Surya Gunawan, et al, Speech emotion recognition using deep feedforward neural network, *Indonesian Journal of Electrical Engineering and Computer Science* 10. no. 2 (2018): 554-561

[18] Bandela, Surekha Reddy, T. Kishore Kumar, Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition, *Applied Acoustics.* 172 (2021): 107645.

[19] Wijayasingha, Lahiru, John A. Stankovic, Robustness to noise for speech emotion classification using CNNs and attention mechanisms, *Smart Health.* 19 (2021): 100165.

[20] Liu, Gang, Wei He, et al, Feature fusion of speech emotion recognition based on deep learning, I *2018 International conference on network infrastructure and digital content (IC-NIDC)*. pp. 193-197. IEEE, 2018.

[21] Koduru, Anusha, Hima Bindu Valiveti, et al, Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology.* 23, no. 1 (2020): 45-55.

[22] Xu, Mingke, Fan Zhang, et al, Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset, *IEEE Access* 9. (2021): 74539-74549.

[23] Abdelhamid, Abdelaziz A., El-Sayed M. El-Kenawy, et al, Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm, *IEEE Access.* 10 (2022): 49265-49284.

[24] Hamsa, Shibani, Ismail Shahin, et al, Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG, *Expert Systems with Applications.* (2023): 119871.

[25] Chang, Yi, Sofiane Laridi, Zhao Ren, et al, Robust federated learning against adversarial attacks for speech emotion recognition, *arXiv preprint. arXiv:2203.04696* (2022).

[26] Hsu, Jia-Hao, Ming-Hsiang Su, et al, Speech emotion recognition considering nonverbal vocalization in affective conversations, *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 29 (2021): 1675-1686.

[27] Huang, Yongming, Kexin Tian, et al, Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition, *Journal of ambient intelligence and humanized computing.* 10 (2019): 1787-1798.

[28] Li, Lu-Qiao, Kai Xie, et al, Emotion recognition from speech with StarGAN and Dense-DCNN, *IET Signal Processing.* 16, no. 1 (2022): 62-79.

[29] Avila, Anderson R., Md Jahangir Alam, et al, Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition, *Interspeech*. pp. 3663-3667. 2018.

[30] Mansour, Asma, Zied Lachiri, A comparative study in emotional speaker recognition in noisy environment, *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*. pp. 980-986. IEEE, 2017.

[31] Nassif, Ali Bou, Ismail Shahin, et al, CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions, *Applied Soft Computing.* 103 (2021): 107141.

[32] Chatterjee, Rajdeep, Saptarshi Mazumdal, et al, Real-time speech emotion analysis for smart home assistants, *IEEE Transactions on Consumer Electronics.* 67, no. 1 (2021): 68-76.

[33] Huang, Che-Wei, Shrikanth Shri Narayanan, Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition, *2017 IEEE international conference on multimedia and expo (ICME)*. pp. 583-588. IEEE, 2017.

[34] Chakraborty, Rupayan, Ashish Panda, et al, Front-End Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition, *INTERSPEECH*. pp. 3257-3261. 2019

[35] Tiwari, Upasana, Meet Soni, et al, Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* pp. 7194-7198. IEEE, 2020.

[36] Triantafyllopoulos, Andreas, Gil Keren, et al, Towards robust speech emotion recognition using deep residual networks for speech enhancement. (2019).

[37] Mitra, Vikramjit, Vasudha Kowtha, et al, Pre-trained Model Representations and their Robustness against Noise for Speech Emotion Analysis, *arXiv preprint. arXiv:2303.03177* (2023).

[38] Xiaoqing, Jiang, Xia Kewen, et al, Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning, *The Journal of China Universities of Posts and Telecommunications.* 24, no. 2 (2017): 1-17.

[39] Jing, Shaoling, Xia Mao, et al, A closed-form solution to the graph total variation problem for continuous emotion profiling in noisy environment, *Speech Communication.* 104 (2018): 66-72.

[40] Huang, Yongming, Wu Ao, et al, Novel sub-band spectral centroid weighted wavelet packet features with importance-weighted support vector machines for robust speech emotion recognition, *Wireless Personal Communications.* 95 (2017): 2223-2238.

[41] Chenchah, Farah, Zied Lachiri, A bio-inspired emotion recognition system under real-life conditions, *Applied Acoustics.* 115 (2017): 6-14.

[42] Laghari, Muddasar, Muhammad Junaid Tahir, et al, Robust speech emotion recognition for sindhi language based on deep convolutional neural network, *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*. pp. 543-548. IEEE, 2021.

[43] Lotz, Alicia Flores, Fabian Faller, et al, Emotion recognition from disturbed speech-towards affective computing in real-world in-car environments, *Elektronische Sprachsignalverarbeitung.* (2018).

[44] Li, Dongdong, Jinlin Liu, et al, Speech emotion recognition using recurrent neural networks with directional self-attention, *Expert Systems with Applications.* 173 (2021): 114683.

[45] Huang, Jian, Ya Li, et al, Speech emotion recognition using semi-supervised learning with ladder networks, *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).* pp. 1-5. IEEE, 2018.

[46] Sekkate, Sara, Mohammed Khalil, et al, An investigation of a feature-level fusion for noisy speech emotion recognition, *Computers.* 8, no. 4 (2019): 91.

[47] Alghifari, Muhammad Fahreza, Teddy Surya Gunawan, et al, On the use of voice activity detection in speech emotion recognition, *Bulletin of Electrical Engineering and Informatics.* 8, no. 4 (2019): 1324-1332

[48] Kwon, Soonil, A CNN-assisted enhanced audio signal processing for speech emotion recognition, *Sensors.* 20, no. 1 (2019): 183.

[49] Win, Htwe Pa Pa, Phyo Thu Thu Khine, Emotion recognition system of noisy speech in real world environment, *International Journal of Image, Graphics and Signal Processing (IJIGSP).* 12, no. 2 (2020): 1-8.

[50] Leem, Seong-Gyun, Daniel Fulford, et al, Separation of Emotional and Reconstruction Embeddings on Ladder Network to Improve Speech Emotion Recognition Robustness in Noisy Conditions, *Interspeech 2021.* (2021): 2871-2875.

[51] C. Busso, M. Bulut, C.-C. e. a. Lee, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation. vol. 42, no. 4, p. 335, 2008.

[52] Stefan Steidl, Automatic classification of emotion related user states in spontaneous children's speech, University of Erlangen-Nuremberg Erlangen. Germany, 2009

[53] Hantke, Simone, et al. Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings, *Interspeech 2017*. (2017): 3137-3141

[54] C. Busso, S. Parthasarathy, A. Burmania, et al, Msp-improv: An acted corpus of dyadic interactions to study emotion perception, *IEEE Transactions on Affective Computing*. vol. 8, no. 1, pp. 67–80, 2017.

[55] Li, Ya, Jianhua Tao, Linlin Chao, et al, CHEAVD: a Chinese natural emotional audio–visual database, *Journal of Ambient Intelligence and Humanized Computing.* 8 (2017): 913-924.

[56] Burkhardt, F Paeschke, A Rolfes, et al, A database of German emotional speech, *Interspeech 2005*.1517-1520, doi: 10.21437/Interspeech.2005-446

[57] S. R. Livingstone, F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, *PloS one*. 13 (5) (2018) e0196391.

[58] Choi, Y., et al, StarGAN: unified generative adversarial networks for multi-domain image-to-image translation, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT. pp. 8789-8797 (2018)

[59] H.-C. Chou, W.-C. Lin, L.-C. Chang, et al, NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus, *7th International Conference Affect. Comput. Intell. Interaction*. 2017, pp. 292–298

[60] Jackson P., Haq S, Surrey Audio-Visual Expressed Emotion (SAVEE) Database, University of Surrey, Guildford, UK (2014)

[61] F. Ringeval, et al, Introducing the recola multimodal corpus of remote collaborative and affective interactions, *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition.* 2013, pp. 1–8.

[62] J.H.L. Hansen, S.E. Bou-Ghazale, Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, *EUROSPEECH* .1997, pp. 1–4

[63] Cen, Ling, Fei Wu, et al, A real-time speech emotion recognition system and its application in online learning, *Emotions, technology, design, and learning*. pp. 27-46. Academic Press, 2016.

[64][O. Martin, I. Kotsia, B. Macq, et al, The eNTERFACE'05 audio-visual emotion database, *International Conference on Data Engineering Workshops*. 2006. DOI: 10.1109/ ICDEW.2006.145

[65] GRIMM M., K. KROSCHEL, S. NARAYANAN, The vera am mittag german audiovisual emotional speech database, *Proc. of the IEEE ICME-2008*. pp. 865–868. Hannover, Germany, 2008.

[66] M. F. Alghifari, T. S. Gunawan, M. Kartiwi, Speech Emotion Recognition Using Deep Feedforward Neural Network, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, 2018.

[67]Pichora-Fuller, M. Kathleen, Dupuis, et al, Toronto emotional speech set(TESS), https://doi.org/10.5683/SP2/E8H2MF, Borealis,v1,2020

[68]Engberg I. S, Hansen A. V, Documentation of the Danish Emotional Speech Database DES, Aalborg, Denmark, 1996

[69] Kang, Bong-Seok, Text Independent Emotion Recognition Using Speech Signals, Yonsei Univ(2000).

[70]Staroniewicz P, Majewski W, Polish Emotional Speech Database – Recording and Preliminary Validation, Esposito, A., Vích, R. (eds) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. Lecture Notes in Computer Science (), vol 5641. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03320-9_5, (2009).

[71] F. Eyben, F. Weninger, F. e. a. Gross, Recent developments in opensmile, the munich open-source multimedia feature extractor, *Proceedings of the 21st ACM international conference on Multimedia. ACM*, 2013, pp. 835–838.

[72]L. Wyse, Audio Spectrogram Representations for Processing with Convolutional Neural Networks, *arXiv preprint* arXiv:1706.09559 (2017). arXiv:1706.09559.

[73] B. McFee, C. Raffel, D. Liang, et al, Librosa: Audio and music signal analysis in Python, *Proc. 14th Python Sci. Conf*. vol. 8, Jul. 2015, pp. 18–25.

[74] Chang CC, Lin CJ, LIBSVM: a library for support vector machines, *ACM Trans Intell Syst Technol (TIST)*. 2(3), pp 1–27 (2011)

[75]S.J. Young, G. Evermann, M.J. Gales, et al, The HTK Book, version 3.4, 2006.

[76] Hirsch, H. Guenter, Fant-filtering and noise adding tool, *Niederrhein University of Applied Sciences,* http://dnt.-kr. hsnr. de/download. html (2005).

[77] Chollet F, Keras: The python deep learning library, Keras., *IoKeras. io* (2015)

[78] Löfberg J, YALMIP: A toolbox for modeling and optimization in MATLAB, *Proceedings of the 2004 International Symposium on Computer Aided Control Systems Design*, Sept 2−4, 2004, Taipei, China. Piscataway, NJ, USA: IEEE, 2004: 284−289

[79] M.-I. Nicolae, M. Sinn, M. N. Tran, et al, Adversarial robustness toolbox v1.0.0, Jul. 2018, 34 pages

[80] Mauch M, Ewert S, The Audio Degradation Toolbox and its Application to Robustness Evaluation, *Proc. of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013).* pp. 83–88. Curitiba, Brazil 2013.

[81] HALL M, E. FRANK, G. HOLMES, et al, The weka data mining software: An update, *SIGKDD Explor. Newsl.*11(1), pp. 10–18, 2009.

[82] D. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. (2010, 14/2/2019).

[83] P. Boersma, Praat: Doing phonetics by computer [Computer Program], 2011. [Online]. Available: http://www.praat.org/

[84] D. Povey, A. Ghoshal, G. Boulianne, et al, The Kaldi speech recognition toolkit*, IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011

[85] Sedaaghi M. H, Documentation of the sahand emotional speech database (SES) (Technical Report), Department of Electrical Eng., Sahand Univ. of Tech, Iran (2008)

[86] X. Mao, L. Chen, Speech emotion recognition based on parametric filter and fractal dimension, *IEICE Trans. Inf. Syst.* vol. E93–D, no. 8, pp. 2324–2326, 2010.

[87] Vryzas, Nikolaos, Rigas Kotsakis, et al, Speech emotion recognition for performance interaction, *Journal of the Audio Engineering Society.* 66, no. 6 (2018): 457-467.

[88] Fan, Yin, Xiangju Lu, et al, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, *Proceedings of the 18th ACM international conference on multimodal interaction*. pp. 445-450. 2016.

[89] Yoon, Seunghyun, Seokhyun Byun et al, Multimodal speech emotion recognition using audio and text, *IEEE Spoken Language Technology Workshop (SLT)*, pp. 112-118. IEEE, 2018.

[90] Wu, Chung-Hsien, Ze-Jing Chuang, et al, Emotion recognition from text using semantic labels and separable mixture models, *ACM transactions on Asian language information processing (TALIP)* 5. no. 2 (2006): 165-183.

[91] J.Thiemann, N. Ito, E. Vincent, The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings, *Proceedings of Meetings on Acoustics ICA2013.* vol. 19, no. 1. Acoustical Society of America, 2013, p. 035081.

[92] E. Vincent, S. Watanabe, A. A. Nugraha, et al, An analysis of environment, microphone and data simulation mismatches in robust speech recognition, *Computer Speech & Language*. vol. 46, pp. 535–557, 2017.

[93] Pearce D, Hirsch H. G, The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy, ICSLP*'00 proceedings.* Beijing: ICSLP (2000).

[94] K. J. Piczak, ESC: Dataset for environmental sound classification, *Proc. 23rd ACM Int. Conf. Multimedia.* Brisbane, QLD, Australia: ACM, 2015, pp. 1015–1018.

[95] JohnsonD.H, Shami P.N, The signal processing information base. *IEEE Signal Process. Mag.* 10(4), 36–42 (2002)

[96] Varga, Andrew, Herman JM Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems*, Speech communication 12*. no. 3 (1993): 247-251.

[97] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, et al, Audio set: An ontology and human-labeled dataset for audio events*, IEEE International Conference on Acoustics, Speech and Signal Processing.* 2017, pp. 776–780. DOI: 10. 1109/ ICASSP. 2017. 7952261

[98] David Snyder, Guoguo Chen, Daniel Povey, MUSAN: A Music, Speech, and Noise Corpus, 2015, *arXiv*:1510.08484v1

[99] Qing, Chunmei, Rui Qiao, et al, Interpretable emotion recognition using EEG signals, Ieee Access 7. 2019: 94160-94170

[100] Jerritta S, M. Murugappan, R. Nagarajan t al, Physiological signals based human emotion recognition: a review, *2011 IEEE 7th international colloquium on signal processing and its applications*. pp. 410-415. IEEE, 2011.

[101] Chen, Tian, Hongfang Yin, et al, Emotion recognition based on fusion of long short-term memory networks and SVMs, *Digital Signal Processing.* 117 (2021): 103153.

[102] Pourebrahim, Yousef, Farbod Razzazi, et al, Semi-supervised parallel shared encoders for speech emotion recognition, *Digital Signal Processing.* 118 (2021): 103205.

[103] Lin, Yi-Lin, Gang Wei. Speech emotion recognition based on HMM and SVM, *International conference on machine learning and cybernetics*. vol. 8, pp. 4898-4901. IEEE, 2005.

[104] Jin, Qin, Chengxin Li, et al, Speech emotion recognition with acoustic and lexical features, *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. pp. 4749-4753. IEEE, 2015.

[105] Ingale, Ashish B, D. S. Chaudhari, Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)* 2. no. 1 (2012): 235-238.

[106] K. Wang, N. An, B. N. Li, et al, Speech emotion recognition using fourier parameters, *IEEE Transactions on Affective Computing*. vol. 6, no. 1, pp. 69 - 75, January 2015.

[107] S. Ntalampiras, Potamitis, N. Fakotakis, An adaptive framework for acoustic monitoring of potential hazards, *EURASIP 1. Audio, Speech, Music Process, no. 13*. 2009.

[108] P. Chandrasekar, S. Chapaneri, D. Jayaswal, Automatic speech emotion recognition: A survey, *IEEE International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA).* pp. 341 - 346, April 2014.

[109] D.J. France, R.G. Shivavi, S. Silverman, et al., Acoustical properties of speech as indicators of depression and suicidal risk, *IEEE Trans. Biomed. Eng. 7.* pp 829-837, 2000.

[110] Kerkeni, Leila, Youssef Serrestou, A review on speech emotion recognition: Case of pedagogical interaction in classroom, *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. pp. 1-7. IEEE, 2017.

[111] Hyun, Kyung-Hak, Eun-Ho Kim, et al, Robust speech emotion recognition using log frequency power ratio, *SICE-ICASE International Joint Conference*. pp. 2586-2589. IEEE, 2006.

[112] Schuller, Bjoern W, Speaker, noise, and acoustic space adaptation for emotion recognition in the automotive environment, *ITG Conference on Voice Communication [8. ITG-Fachtagung]*. pp. 1-4. VDE, 2008.

[113] Schuller, Björn, Dejan Arsic, et al, Emotion recognition in the noise applying large acoustic feature sets.(2006).

[114] Georgogiannis, Alexandros, Vassilis Digalakis, Speech emotion recognition using non-linear teager energy based features in noisy environments, *2012 proceedings of the 20th European signal processing conference (EUSIPCO)*. pp. 2045-2049. IEEE, 2012.

[115] Han, Zhiyan, Shuxian Lun, et al, A study on speech emotion recognition based on CCBC and neural network, *International Conference on Computer Science and Electronics Engineering*. vol. 2, pp. 144-147. IEEE, 2012.

[116] Kim, Eun Ho, Kyung Hak Hyun, et al, Speech emotion recognition using eigen-fft in clean and noisy environments, *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. pp. 689-694. IEEE, 2007.

[117] You, Mingyu, Chun Chen, et al, Manifolds based emotion recognition in speech, *International Journal of Computational Linguistics & Chinese Language Processing. Volume 12, Number 1, March 2007: Special Issue on Affective Speech Processing*, pp. 49-64. 2007.

[118] Schuller, Bjorn, Dino Seppi, et al, Towards more reality in the recognition of emotional speech, *IEEE international conference on acoustics, speech and signal processing-ICASSP'07*. vol. 4, pp. IV-941. IEEE, 2007.

[119] Tawari, Ashish, Mohan M. Trivedi, Speech emotion analysis in noisy real-world environment, *20th International Conference on Pattern Recognition*. pp. 4605-4608. IEEE, 2010.

[120] Sztahó, Dávid, Viktor Imre, Automatic classification of emotions in spontaneous speech, *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues: COST 2102 International Conference, Budapest, Hungary, September 7-10, 2010, Revised Selected Papers*, pp. 229-239. Springer Berlin Heidelberg, 2011.

[121] Trigeorgis, George, Fabien Ringeval, et al, Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network, *IEEE international conference on acoustics, speech and signal processing (ICASSP).* pp. 5200-5204. IEEE, 2016.

[122] Aher, Prashant K., Swapnil D. Daphal, et al, Analysis of feature extraction techniques for improved emotion recognition in presence of additive noise, *International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. pp. 350-354. IEEE, 2016.

[123] Juszkiewicz, Łukasz, Improving noise robustness of speech emotion recognition system, Intelligent *Distributed Computing VII: Proceedings of the 7th International Symposium on Intelligent Distributed Computing-IDC 2013. Prague, Czech Republic, September 2013*, pp. 223-232. Springer International Publishing, 2014.

[124] Karimi, Salman, Mohammad Hossein Sedaaghi, Robust emotional speech classification in the presence of babble noise, *International Journal of Speech Technology.* 16 2013: 215-227.

[125] Song, Peng, Wenming Zheng, et al, Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization, *Speech Communication* 83. (2016): 34-41.

[126] Revathy A, P. Shanmugapriya, V. Mohan, Performance comparison of speaker and emotion recognition, *3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*. pp. 1-6. IEEE, 2015.

[127] Mao, Qirong, Ming Dong, et al, Learning salient features for speech emotion recognition using convolutional neural networks, *IEEE transactions on multimedia* 16. no. 8 (2014): 2203-2213.

[128] Zhao, Xiaoming, Shiqing Zhang, et al, Robust emotion recognition in noisy speech via sparse representation, *Neural Computing and Applications* 24. (2014): 1539-1553.

[129] Chenchah, Farah, Zied Lachiri, Speech emotion recognition in noisy environment, *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP).* pp. 788-792. IEEE, 2016.

[130] Huang, Chengwei, C. H. E. N. Guoming, et al, Speech emotion recognition under white noise, *Archives of Acoustics* 38. no. 4 (2013): 457-463.

[131] Song, Peng, Yun Jin, et al, Speech emotion recognition using transfer learning, *IEICE TRANSACTIONS on Information and Systems* 97. no. 9 (2014): 2530-2532.

[132] Eskimez, Sefik Emre, Zhiyao Duan, et al, Unsupervised learning approach to feature analysis for automatic speech emotion recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 5099-5103. IEEE, 2018.

[133] Jahangir, Rashid, Ying Wah Teh, et al, Deep learning approaches for speech emotion recognition: state of the art and research challenges, Multimedia *Tools and Applications.* (2021): 1-68.

[134] Al-Dujaili M.J., Ebrahimi-Moghadam, A. Speech Emotion Recognition: A Comprehensive Survey. Wireless Pers Commun 129. 2525–2561 (2023). https://doi.org/10.1007/s11277-023-10244-3

[135] de Lope, Javier, and Manuel Graña, An ongoing review of speech emotion recognition, *Neurocomputing* (2023).

[136] Gunawan, Teddy Surya, Muhammad Fahreza Alghifari, et al, A review on emotion recognition algorithms using speech analysis, *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)* 6. no. 1 (2018): 12-20.

[137] R. Lotfian and C. Busso, Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings, *IEEE Transactions on Affective Computing.* vol. 10, no. 4, pp. 471–483, October-December 2019.

[138] Kakuba, Samuel, Dong Seog Han, Speech Emotion Recognition using Context-Aware Dilated Convolution Network, *27th Asia Pacific Conference on Communications (APCC)*. pp. 601-604. IEEE, 2022.

[139] Laurence Devillers et al, Challenges in real-life emotion annotation and machine learning based detection*, Science Direct, Neural Networks 18.* (2005) 407–422

[140] E. Parada-Cabaleiro, G. Costantini, A. Batliner, et al, Demos: An italian emotional speech corpus, *Language Resources and Evaluation*, vol. 54, pp. 341–383, Feb. 2019

[141] Shahin, Ismail, Ali Bou Nassif, et al, Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments, *Neural Computing and Applications* 32. (2020): 2575-2587.

[142] Kamaruddin, Norhaslinda, Abdul Wahab, et al, Cultural dependency analysis for understanding speech emotion, *Expert Systems with Applications* 39. no. 5 (2012): 5115-5133.

Declaration of interests

☒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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: