Journal Pre-proof

Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers

Mehmet Berkehan Akçay, Kaya Oğuz

PII: S0167-6393(19)30226-2
DOI: https://doi.org/10.1016/j.specom.2019.12.001
Reference: SPECOM 2682

To appear in: *Speech Communication*

Received date: 16 June 2019
Revised date: 24 October 2019
Accepted date: 12 December 2019

Please cite this article as: Mehmet Berkehan Akçay, Kaya Oğuz, Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers, *Speech Communication* (2019), doi: https://doi.org/10.1016/j.specom.2019.12.001

Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers

Mehmet Berkehan Akçay[a,*], Kaya Oğuz[b]

[a]*Department of Software Engineering, Izmir University of Economics, Izmir, Turkey*
[b]*Department of Computer Engineering, Izmir University of Economics, Izmir, Turkey*

Abstract

Speech is the most natural way of expressing ourselves as humans. It is only natural then to extend this communication medium to computer applications. We define speech emotion recognition (SER) systems as a collection of methodologies that process and classify speech signals to detect the embedded emotions. SER is not a new field, it has been around for over two decades, and has regained attention thanks to the recent advancements. These novel studies make use of the advances in all fields of computing and technology, making it necessary to have an update on the current methodologies and techniques that make SER possible. We have identified and discussed distinct areas of SER, provided a detailed survey of current literature of each, and also listed the current challenges.

Keywords: Speech emotion recognition, Survey, Speech features, classification, speech databases

1. Introduction

As humans we find speech to be the most natural way to express ourselves. We depend so much on it that we recognize its importance when we have to use other ways of communication, such as emails or text messages. It is no surprise that emojis have become common in text messages, because these text messages could be misunderstood, and we would like to pass the emotion along with the text as we do in speech.

Since emotions help us to understand each other better, a natural outcome is to extend this understanding to computers. Speech recognition is already in our everyday life, thanks to the smart mobile devices that are able to accept and reply to voice commands with synthesized speech. The speech emotion recognition (SER) could be used to enable them to detect our emotions, as well.

SER has been around for more than two decades [1] and it has applications in human-computer interaction [2], as well as robots [3], mobile services [4], call centers [5], computer games [6], and psychological assessment [7, 8]. Although it has many applications, emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them. They are evaluated by their perception in other humans, and at times, even we are known to misinterpret them.

We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. When we take a bird's eye view, we can separate it into several

*Corresponding author
Email addresses: berkehan.akcay@ieu.edu.tr (Mehmet Berkehan Akçay), kaya.oguz@ieu.edu.tr (Kaya Oğuz)

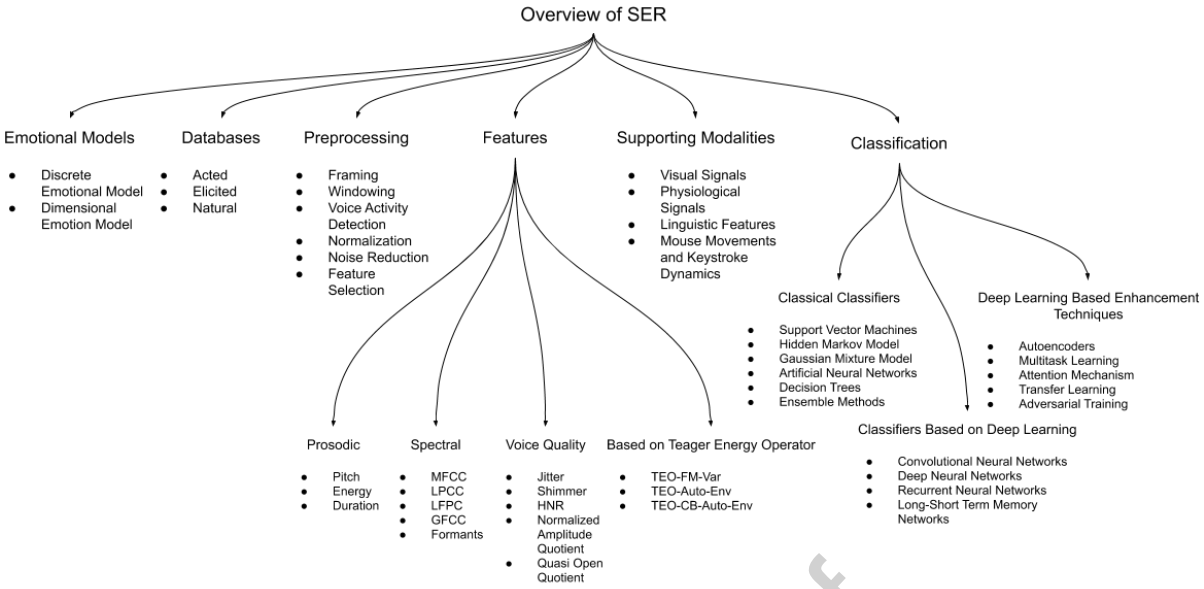

Figure 1: An overview of speech emotion recognition systems. The recognition requirements flow from left to right. The emotions are embedded in the databases on the far left, and they are extracted at the far right end of the figure.

distinct areas, as shown in Figure 1. It would be beneficial to understand emotions better so that the classification process can be improved. There are various approaches to model the emotions, and it is still an open problem; however, the discrete and the dimensional models are commonly used. Therefore, we first review the emotional models. A SER system requires a classifier, a supervised learning construct, that will be trained to recognize emotions in new speech signals. Such a supervised system brings the necessity of labeled data that have emotions embedded in them. The data requires preprocessing before their features can be extracted. Features are essential to a classification process. They reduce the original data to its most important characteristics. For speech signals they can be categorized under four groups; prosodic, spectral, voice quality, and features based on Teager energy operator. The classifier can be strengthened by incorporating additional features from other modalities, such as visual or linguistic depending on the application and availability. All these features are then passed to the classification system which has a wide range of classifiers available to them. More recently, classifiers that incorporate deep learning have also become common.

All areas provided in Figure 1 are surveyed from left to right, with an up-to-date literature. The following section discusses the existing surveys and the areas that they cover. Section 3 surveys emotions, and Section 4 surveys the databases. The methodologies for preprocessing, feature extraction, supporting modalities, and classification are grouped under speech emotion recognition, and are detailed in Section 5. The paper concludes with the listing of current challenges in Section 6 and concluding remarks in Section 7.

2. Related Work

There are other publications that survey existing studies on speech emotion recognition. A list of such surveys that are published relatively recently are listed in Table 1 and compared according to the

areas they cover. Naturally, earlier publications do not include recent advances and trends such as deep neural networks in their sections on classifiers.

In 2006, Ververidis and Kotropoulos specifically focused on speech data collections, while also reviewing acoustic features and classifiers in their survey of speech emotion recognition [9]. Ayadi et al. have presented their survey with an updated literature and included the combination of speech features with supporting modalities, such as linguistic, discourse, and video information [10]. Koolagudi and Rao have also relied on the classification of databases, features, and classifiers for their survey [11].

Anagnostopoulos and Giannoukos have provided a comprehensive survey of publications between 2000 and 2011 [12]. Their survey is one of the first ones to include studies that have applications of deep neural networks to SER. They also highlight the studies that use hybrid classifiers, ensembles, and voting schemes.

The study by Ramakrishnan includes not only the databases, features and classifiers in the SER systems, but also mentions the normalization of signals, which is preprocessing stage that is performed before the extraction of the features [13]. He also suggests application areas for SER systems, which are not part of the SER technologies but affect them in their requirements and design.

A recent but brief survey by Basu et al. highlights publications that involve databases, noise reduction techniques for preprocessing signals, features, and classifiers including recent advances such as Convolutional and Recurrent Neural Networks [14].

A more recent survey by Sailunaz et al. focus on emotion detection from text and speech, where publications that incorporate text information as well as speech signals to determine speech are discussed. Unlike other surveys, they also discuss the emotional models [15]. Their survey also discusses the recent classifiers.

In comparison to other surveys, this study provides a thorough survey of all areas in the SER; the databases, features, preprocessing techniques, supporting modalities, classifiers, and emotional models.

3. Emotions

To successfully implement a speech emotion recognition system, we need to define and model emotion carefully. However, there is no consensus about the definition of emotion, and it is still an open problem in psychology. According to Plutchik, more than ninety definitions of emotion were proposed in the twentieth century [16]. Emotions are convoluted psychological states that are composed of several components such as personal experience, physiological, behavioral, and communicative reactions. Based on these definitions, two models have become common in speech emotion recognition: discrete emotional model, and dimensional emotional model.

Discrete emotion theory is based on the six categories of basic emotions; sadness, happiness, fear, anger, disgust, and surprise, as described by Ekman [17, 18]. These inborn and culturally independent emotions are experienced for a short period [19]. Other emotions are obtained by the combination of the basic ones. Most of the existing SER systems focus on these basic emotional categories. In daily life, people use this model to define their observed emotions, hence labeling scheme based on emotional

Table 1: Recent surveys on Speech Emotion Recognition and the areas they cover, compared by the areas this study covers. The comparison is done by their inclusion of databases, features, preprocessing methods, supporting modalities, classifiers, and emotional models.

Publication	Date	DB	Feat.	Prep.	Supp. Mod.	Classf.	Em. Mod.
Emotional speech recognition: Resources, features, and methods [9]	2006	✓	✓	x	x	Partial[3]	x
Survey on speech emotion recognition: Features, classification schemes, and databases [10]	2011	✓	✓	x	✓	Partial[3]	x
Emotion recognition from speech: a review [11]	2012	✓	✓	x	x	Partial[3]	x
Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 [12]	2012	✓	✓	x	x	✓	x
Recognition of Emotion from Speech: A Review [13]	2012	✓	✓	Partial[1]	x	Partial[3]	x
A review on emotion recognition using speech [14]	2017	✓	✓	Partial[1]	x	✓	x
Emotion detection from text and speech: a survey [15]	2018	✓	✓	x	Partial[2]	✓	✓
This study	2019	✓	✓	✓	✓	✓	✓

[1] Only noise reduction or normalization

[2] Only textual features

[3] Does not include recent advances, such as deep neural networks

categories are intuitive. Nonetheless, these discrete categories of emotions are not able to define some of the complex emotional states observed in daily communication.

Dimensional emotional model is an alternative model that uses a small number of latent dimensions to characterize emotions such as valence, arousal, control, power [20, 21]. These dimensions are definitive and generic aspects of emotion. In the dimensional approach, emotions are not independent of each other; instead, they are analogous to each other in a systematic way. One of the most preferred dimensional models is a two-dimensional model that uses arousal, activation, or excitation on one dimension, versus valence, appraisal, or evaluation on the other. Valence dimension describes whether an emotion is positive or negative, and it ranges between unpleasant and pleasant. Arousal dimension defines the strength of the felt emotion. It may be excited or apathetic, and it ranges from boredom to frantic excitement [22]. The three-dimensional model includes a dimension of dominance or power, which refers to the seeming strength of the person that is between weak and strong. For instance, the third dimension differentiates anger from fear by considering the strength or weakness of the person, respectively [23].

There are several disadvantages for the dimensional representation. It is not intuitive enough and

special training may be needed to label each emotion [24]. In addition, some of the emotions become identical, such as fear and anger, and some emotions like surprise cannot be categorized and lie outside of the dimensional space since surprise emotion may have positive or negative valence depending on the context.

4. Databases

Databases are an essential part of speech emotion recognition since classification process relies on the labeled data. Quality of the data affects the success of the recognition process. Incomplete, low-quality, or faulty data may lead to incorrect predictions; hence, data should be carefully designed and collected. Databases for speech emotion recognition can be investigated in three parts:

- Acted (Simulated) speech emotion databases
- Elicited (Induced) speech emotion databases
- Natural speech emotion databases

Utterances in acted speech databases are recorded by professional or semi-professional actors in sound-proof studios. It is relatively easier to create such a database compared to the other methods; however, it is stated by the researchers that acted speech cannot convey the real-life emotions adequately, and even may be exaggerated. This lowers the recognition rates for real-life emotions.

Elicited speech databases are created by placing speakers in a simulated emotional situation that can stimulate various emotions. Although the emotions are not fully-elicited, they are close to real ones.

Natural speech databases are mostly obtained from talk shows, call-center recordings, radio talks, and similar sources. Sometimes, these real-world speeches are referred to as spontaneous speech. It is harder to obtain the data since ethical and legal problems arise when processing and distributing them.

Once the method of creating a database is decided, other design issues are considered, such as age and gender. Most databases contain adult speakers, but databases of children and elders also do exist. Other considerations include repeating utterances with different actors, different emotions, and different genders.

For example, the commonly used Berlin dataset contains seven emotions uttered by ten professional actors, half male, half female [25]. Each utterance is repeated with different actors and different emotions. A list of prominent datasets are summarized in Table 2.

Table 2: There are several data sets used for emotion recognition. This table contains the prominent ones, along with unique data sets for various languages and special cases such as the ones that contain utterances by elders and children.

Database	Language	Size	Access Type	Emotions	Type	Modalities
Berlin Emotional Database (EmoDB) [25]	German	7 Emotions x 10 speakers (5 male, 5 female) x 10 utterances	Open access	Anger, boredom, disgust, fear, happiness, sadness, neutral	Acted	Audio
Chinese Emotional Speech Corpus (CASIA) [26]	Mandarin	6 Emotions x 4 Speakers (2 male, 2 female) x 500 utterances (300 parallel, 200 non-parallel texts)	Commercially available	Surprise, happiness, sadness, anger, fear, neutral	Acted	Audio
The Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [27]	English	10 speakers(5 male, 5 female) 1150 utterances	Available with license	Happiness, anger, sadness, frustration, neutral	Acted	Audio/Visual

Table 2: There are several data sets used for emotion recognition. This table contains the prominent ones, along with unique data sets for various languages and special cases such as the ones that contain utterances by elders and children.

Database	Language	Size	Access Type	Emotions	Type	Modalities
Surrey Audio-Visual Expressed Emotion (SAVEE)[28]	English	14 speakers (male) x 120 utterances	Free	Anger, disgust fear, happiness, sadness, surprise, neutral, common	Acted	Audio/Visual
Toronto Emotional Speech Database (TESS)[29]	English	2 speakers (female), 2800 utterances	Free	Anger, disgust, neutral fear, happiness, sadness pleasant, surprise	Acted	Audio
Beihang University Database of Emotional Speech (BHUDES)[30]	Mandarin	5 speakers (2 male, 3 female), 323 utterances		Anger, happiness, fear, disgust, surprise	Acted	Audio
Chinese Annotated Spontaneous Speech corpus (CASS)[31]	Mandarin	7 speakers (2 male, 5 female), 6 hours of speech	Commercially available	Anger, fear, happiness, sadness, surprise, neutral	Natural	Audio
Chinese Natural Emotional Audio–Visual Database (CHEAVD)[32]	Mandarin	238 speakers (child to elderly) 140 minute emotional segments from movies, TV-shows.	Free to research use	Anger, anxious, disgust, happiness, neutral, sadness, surprise and worried	Acted Natural	Audio/Visual
Danish Emotional Speech Database (DES)[33]	Danish	4 speakers (2 male, 2 female) 10 minutes of speech	Free	Neutral, surprise, anger, happiness, sadness	Acted	Audio
Chinese Elderly Emotional Speech Database (EESDB)[34]	Mandarin	16 speakers (8 male, 8 female), 400 utterances from teleplay	Free to research use	Anger, disgust, fear, happiness, neutral, sadness, surprise	Acted	Audio
Electromagnetic Articulography Database (EMA)[35]	English	3 speakers (1 male, 2 female) 14 sentences for male, 10 sentences for female	Free to research use	Anger, happiness, sadness, neutral	Acted	Audio/ Articulatory movement data
Italian Emotional Speech Database (EMOVO)[36]	Italian	6 speakers (3 male, 3 female) x 14 sentences x 7 emotions= 588 utterances	Free	Disgust, happiness, fear, anger, surprise, sadness, neutral	Acted	Audio
eNTERFACE'05 Audio-Visual Emotion Database[37]	English	42 speakers (34 male, 8 female) from 14 nationalities 1116 video sequences	Free	Anger, disgust, fear, happiness, sadness, surprise	Elicited	Audio/Visual
Keio University Japanese Emotional Speech Database (Keio-ESD)[38]	Japanese	71 speaker (male) 940 utterances	Free	Anger, happiness, disgusting, downgrading, funny, worried, gentle, relief, indignation, shameful, etc.(47emotions)	Acted	Audio
LDC Emotional Speech Database[39]	English	7 speakers (4 male, 3 female), 470 utterances	Commercially available	Hot anger, cold anger, disgust, fear, contempt, happiness, sadness, neutral, panic, pride, despair, elation, interest, shame, boredom	Acted	Audio
RECOLA Speech Database[40]	French	46 speakers (19 males, 27 females) 7 hour of speech	Free	Five social behaviors (agreement, dominance, engagement, performance, rapport); arousal and valence	Natural	Audio/Visual
SAMAINE Database[41]	English Greek Hebrew	150 speakers, 959 conversation	Free	Valence, activation, power, expectation, overall emotional intensity	Natural	Audio/Visual
Speech Under Simulated and Actual Stress Database (SUSAS)[42]	English	32 speakers (19 male, 13 female), 16000 utterances also include speech of Apache Helicopter pilots	Commercially available	Four states of speech under stress: Neutral, Angry, Loud, and Lombard	Natural Acted	Audio
Vera Am Mittag Database (VAM)[43]	German	47 speakers from talk-show, 947 utterances	Free	Valence, activation, and dominance	Natural	Audio/Visual
FAU Aibo Emotion Corpus[44]	German	51 children talking to robot dog Aibo, 9 hours of speech	Commercially available	Anger, bored, emphatic, helpless, joyful, motherese, neutral, reprimanding, rest, surprised, touchy	Natural	Audio
TUM AVIC Database[45]	English	21 speakers (11 male, 10 female), 3901 utterances	Free	Five level of interest; 5 non-linguistic vocalizations (breathing, consent, garbage, hesitation, laughter)	Natural	Audio/Visual
AFEW Database[46]	English	330 speakers, 1426 utterances from movies, TV-shows	Free	Anger, disgust, surprise, fear, happiness, neutral, sadness	Natural	Audio/Visual

Table 2: There are several data sets used for emotion recognition. This table contains the prominent ones, along with unique data sets for various languages and special cases such as the ones that contain utterances by elders and children.

Database	Language	Size	Access Type	Emotions	Type	Modalities
Turkish Emotional Speech Database (TURES)[47]	Turkish	582 speakers (394 male, 188 female) from movies, 5100 utterances	Free to research use	Happiness, surprised, sadness, anger, fear, neutral, valence, activation, and dominance	Acted	Audio
BAUM-1 Speech Database[48]	Turkish	31 speakers (18 male, 13 female) 288 acted, 1222 spontaneous video clip	Free to research use	Happiness, anger, sadness, disgust, fear, surprise, bothered, boredom, contempt unsure, being thoughtful, concentration, interest	Acted Natural	Audio/Visual

5. Speech Emotion Recognition

5.1. Preprocessing

Preprocessing is the very first step after collecting data that will be used to train the classifier in a SER system. Some of these preprocessing techniques are used for feature extraction, while others are used to normalize the features so that variations of speakers and recordings would not affect the recognition process.

5.1.1. Framing

Signal framing, also known as speech segmentation, is the process of partitioning continuous speech signals into fixed length segments to overcome several challenges in SER.

Emotion can change in the course of speech since the signals are non-stationary. However, speech remains invariant for a sufficiently short period, such as 20 to 30 milliseconds. By framing the speech signal, this quasi-stationary state can be approximated, and local features can be obtained. Additionally, the relation and information between the frames can be retained by deliberately overlapping 30% to 50% of these segments. Continuous speech signals restrain the usage of processing techniques such as Discrete Fourier Transform (DFT) for feature extraction in applications such as SER. Consequently, fixed size frames are suitable for classifiers, such as Artificial Neural Networks, while retaining the emotion information in speech.

5.1.2. Windowing

After framing the speech signal, the next phase is generally applying a window function to frames. The windowing function is used to reduce the effects of leakages that occurs during Fast Fourier Transform (FFT) of data caused by discontinuities at the edge of the signals. Typically a Hamming window is used, as given in Equation 1, where the window size is M for the frame $w(n)$.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \qquad 0 \leq n \leq M-1 \tag{1}$$

5.1.3. Voice Activity Detection

An utterance consists of three parts; voiced speech, unvoiced speech, and silence. Voiced speech is generated with the vibration of vocal folds that creates periodic excitation to the vocal tract during the

pronunciation of phonemes which are perceptually distinct units of sound that distinguish one word from another; such as bag, tag, tab.

On the other hand, unvoiced speech is the result of air passing through a constriction in the vocal tract, producing transient and turbulent noises that are aperiodic excitations of the vocal tract. Due to its periodic nature, voiced speech can be identified and extracted. The detection of the presence of voiced speech among various unvoiced speech and silence is called endpoint detection, speech detection or voice activity detection.

The performance of the endpoint detection algorithm affects the accuracy of the system. It's hard to model silence and noise accurately in a dynamic environment; if voice and noise frames are removed, it will be easier to model speech. In addition, speech consists of many silent and noisy frames which increase the computational complexity. Removal of these frames decreases the complexity and increases accuracy. Most widely used methods for voice activity detection are zero crossing rate, short time energy, and auto-correlation method.

Zero crossing rate is the rate at which a signal changes its sign from positive to negative or vice versa within a given time frame. In voiced speech, the zero crossing count is low whereas it has a high count in unvoiced speech [49]. The voiced speech has high energy due to its periodicity while low energy is observed in the unvoiced speech. The auto-correlation method provides a measure of similarity between a signal and itself as a function of delay. It is used to find repeating patterns. Because of its periodic nature, voiced signals can be detected using the auto-correlation method.

5.1.4. Normalization

Feature normalization is an important step which is used to reduce speaker and recording variability without losing the discriminative strength of the features. By using feature normalization, the generalization ability of features are increased. Normalization can be done at different levels, such as function level and corpus level. Most widely used normalization method is z-normalization (standard score). If mean μ and standard deviation σ of the data is known, z-normalization is calculated as $z = \frac{x-\mu}{\sigma}$.

5.1.5. Noise Reduction

In real life, the noise present in the environment is captured along with the speech signal. This affects the recognition rate, hence some noise reduction techniques must be used to eliminate or reduce the noise. Minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) estimators are most successfully applied methods for noise reduction [50].

In MMSE, the clean signal is estimated from a given sample function of the noisy signal. It needs apriori information of speech and noise spectrum. It is based on the assumption that the additive noise spectrum and estimate of the speech spectrum is available. The aim of the method is minimizing the expected distortion measure between clean and estimated speech signal.

There are also single-channel noise reduction techniques such as spectral subtraction that can be used for noise reduction.

5.1.6. Feature Selection and Dimension Reduction

Feature selection and dimension reduction are important steps in emotion recognition. There is a need to use a feature selection algorithm because there are many features and there is no certain set of features to model the emotions. Otherwise, with so many features, the classifiers are faced with the curse of dimensionality, increased training time and over-fitting that highly affect the prediction rate.

Feature selection is the process of choosing a relevant and useful subset of the given set of features. The unneeded, redundant or irrelevant attributes are identified and removed to provide a more accurate predictive model. Luengo et al. used a Forward 3-Backward 1 wrapper method which selects features that maximize the accuracy in each step [51]. After these three steps, the least useful feature is eliminated. A 93.50% of recognition rate has been obtained without feature selection using prosodic features with a SVM classifier, whereas, a 92.38% recognition rate has been obtained with the selected six features. They state that the slight reduction in the recognition rate is compensated by the lower computational cost of extracting the features and training. They also report that using GMM, a recognition rate of 84.79% has been achieved with the complete set of 86 prosodic features, while the rate has increased to 86.71% with the selection of six best features.

Schuller et al. used an SVM based Sequential Floating Forward Search (SFFS) algorithm to decrease the number of features [52]. With the original 276 features, they obtained 76.23% recognition rate, while SFFS yielded the top 75 features with an 80.53% recognition rate.

Rong et al. proposed a selection algorithm called Ensemble Random Forest to Trees (ERFTrees) that can be used on a small number of data sets with a large number of features [53]. ERFTrees consists of two parts: feature selection and voting strategy. First, using the C4.5 decision tree and Random Forest algorithm, two subsets of candidate features are selected among original features. The majority voting method combines these two subsets to obtain the final feature set. They have achieved a 66.24% recognition rate with original 84-dimensional features, and a rate of 61.18% with selected 16-features by Multi-Dimensional Scaling, a rate of 60.40% with ISOMAP, and a rate of 69.32% with the proposed algorithm.

In his study, Schuller used correlation-based feature subset (CFS) selection [54]. In CFS, useful features are uncorrelated with each other while they are highly correlated with the target class. From 760 acoustic features, for each of valence, activation, dominance dimensions, the number of features is reduced to 238, 109, and 88 for each, while the correlation coefficients increased from 0.789 to 0.810, 0.403 to 0.451, and 0.745 to 0.788 for each, respectively.

5.2. Features

Features are an important aspect of speech emotion recognition. Carefully crafted set of features that successfully characterize each emotion increases the recognition rate. Various features have been used for SER systems; however, there is no generally accepted set of features for precise and distinctive classification. The existing studies have all been experimental so far.

Speech is a continuous signal of varying length that carries both information and emotion. Therefore, global or local features can be extracted depending on the required approach. Global features, also

called long-term or supra-segmental features, represent the gross statistics such as mean, minimum and maximum values, and standard deviation. Local features, also known as short-term or segmental features, represent the temporal dynamics, where the purpose is to approximate a stationary state. These stationary states are important because emotional features are not uniformly distributed over all positions of the speech signal [55]. For example, emotions such as anger are predominant at the beginning of utterances, whereas, the surprise is overwhelmingly conveyed at the end of it. Hence, to capture the temporal information from the speech, local features are used.

These local and global features of SER systems are analyzed in the following four categories.

- Prosodic Features
- Spectral Features
- Voice Quality Features
- Teager Energy Operator (TEO) Based Features

Prosodic and spectral features are used more commonly in SER systems. Some of the features are listed under different categories by various studies depending on their approach. TEO features are specifically designed for recognizing stress and anger. These features are detailed individually; however, in practice, they are commonly combined to obtain better results.

5.2.1. Prosodic Features

Prosodic features are those that can be perceived by humans, such as intonation and rhythm. A typical example is rising the intonation in a sentence that is meant as a question: "You are coming tonight?," where in this case, the intonation rises on the word "tonight," hinting that this is meant as a question. They are also known as para-linguistic features as they deal with the elements of speech that are properties of large units as in syllables, words, phrases, and sentences. Since they are extracted from these large units, they are long-term features. Prosodic features have been discovered to convey the most distinctive properties of emotional content for speech emotion recognition [24].

The most widely used prosodic features are based on fundamental frequency, energy, and duration. The fundamental frequency, F_0, is created by the vibrations in the vocal cord. It yields rhythmic and tonal characteristics of the speech. The change of the fundamental frequency over the course of an utterance yields its fundamental frequency contour whose statistical properties can be used as features. The energy of the speech signal, sometimes referred as volume or the intensity, provides a representation which reflects amplitude variation of speech signals over time. Researchers suggest that high arousal emotions such as anger happiness or surprise yields increased energy while disgust and sadness result with decreased energy [56]. Duration is the amount of time to build vowels, words and similar constructs that are present in speech. Speech rate, duration of silence regions, rate of duration of voiced and unvoiced regions, duration of longest voiced speech are among the most widely used duration related features.

There are correlations between prosodic features and emotional states. Prosodic features expose the changes during the course of emotional speech. For instance, throughout the production of the high-level arousal emotions such as anger, fear, anxiety, and joy, mean F_0, F_0 variability, and vocal intensity

increases. F_0 contour decreases over time during the expression of anger. In contrast, it increases over time during the expression of joy. Low-level arousal such as sadness yields lower mean F_0, F_0 variability, and vocal intensity compared to natural speech, while also F_0 decreases over time [57, 58]. Duration to express anger is shorter than duration to express sadness [55].

There are many studies which focus on different aspects of the prosodic features. Prosodic features and their correlation with emotional states are inspected in [57, 59]. Some studies show that SER systems get similar results or perform better compared to human judges when prosodic features are used [60, 51].

As previously mentioned, the fundamental frequency is an important prosodic feature for SER. Many features can be derived from the F_0 contour, yet it is unknown which fundamental frequency related feature represents the emotions better.

Busso et al. analyzed various expressive F_0 contour statistics to find the emotionally salient aspects of the F_0 contour [61]. Gross statistics such as the mean, maximum and minimum values, and the range of the F_0 are found to be the most salient aspects of F_0 contour. They also conduct their experiment by extracting features on the sentence and voiced regions levels. The results showed that features from the sentence level surpass the features from the voiced region level.

The performance of the prosodic features based on their granularity is also analyzed in several studies. Schuller et al. compare gross statistics of pitch and energy contours, to instantaneous pitch and energy features using continuous Hidden Markov Model [62]. They obtained 86.6% recognition rate using global features, 77.6% by local ones while human judges have a recognition rate 79.8%. Rao et. al compared the local and global prosodic features, and their combination [55]. The global features are computed from gross statistics of prosodic features. The local prosodic features are gathered from the sequence of syllable duration, frame level pitch and energy values. Compared to the performance of the local features, when the local and global prosodic features are combined, performance is slightly increased. It is also observed that from the word and syllable level prosodic analysis, final words of sentences and syllables involve more information to distinguish emotions compared to other parts of words and syllables.

5.2.2. Spectral Features

When sound is produced by a person, it is filtered by the shape of the vocal tract. The sound that comes out is determined by this shape. An accurately simulated shape may result in an accurate representation of the vocal tract and the sound produced. Characteristics of the vocal tract are well represented in the frequency domain [11]. Spectral features are obtained by transforming the time domain signal into the frequency domain signal using the Fourier transform. They are extracted from speech segments of length 20 to 30 milliseconds that is partitioned by a windowing method.

Mel Frequency Cepstral Coefficients (MFCC) feature represents the short term power spectrum of the speech signal. To obtain MFCC, utterances are divided into segments, then each segment is converted into the frequency domain using short time discrete Fourier transform. A number of sub-band energies are calculated using a Mel filter bank. Then, the logarithm of those sub-bands is calculated. Finally, inverse Fourier transform is applied to obtain MFCC. It is the most widely used spectral feature [63].

Linear Prediction Cepstral Coefficients(LPCC) also embodies vocal tract characteristics of speakers.

Those characteristics show differences with particular emotions. LPCC can be directly obtained with a recursive method from Linear Prediction Coefficient(LPC). LPC is basically the coefficients of all-pole filters and is equivalent to the smoothed envelope of the log spectrum of the speech [64].

Another feature, Log-Frequency Power Coefficients (LFPC), mimics logarithmic filtering characteristics of the human auditory system by measuring spectral band energies using Fast Fourier Transform [65].

Gammatone Frequency Cepstral Coefficients (GFCC) is also a spectral feature obtained by a similar technique of MFCC extraction. Instead of applying Mel filter bank to the power spectrum, Gammatone filter-bank is applied.

Formants are the frequencies of the acoustic resonance of the vocal tract. They are computed as amplitude peaks in the frequency spectrum of the sound. They determine the phonetic quality of a vowel, hence used for vowel recognition.

Sato et al. use segmental MFCC features for speech emotion recognition [66]. They labeled each frame using multi-template MFCC clustering. They compared the performance with prosody based algorithms using k-nearest neighbors and compared with conventional MFCC based algorithms using HMM. They achieved better performance using the new method.

Bitouk et al. introduced a new set of spectral features which are statistics of MFCC calculated over three phoneme type classes of interest – stressed and unstressed vowels, and consonants in the utterance [67]. Compared to prosodic features or utterance level spectral features, they yielded results that have higher accuracy using the proposed features. In addition, combination of these features with prosodic features also increase accuracy. It has been also found that compared to stressed and unstressed vowel features, the consonant regions of the utterance involve more emotional information.

5.2.3. Voice Quality Features

Voice quality is determined by the physical properties of the vocal tract. Involuntary changes may produce a speech signal that might differentiate emotions using properties such as the jitter, shimmer, and harmonics to noise ratio (HNR). There is a strong correlation between voice quality and emotional content of the speech [68].

Jitter is the variability of fundamental frequency between successive vibratory cycles, while shimmer is the variable of the amplitude. Jitter is a measure of frequency instability, whereas shimmer is the amplitude instability. Harmonics to Noise Ratio is the measurement of the relative level of noise in the frequency spectrum of vowels. It is the ratio between periodic to aperiodic component in voiced speech signals. These variations are perceived as changes in voice quality.

Other quality measurements used in literature are Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), Maxima Dispersion Quotient (MDQ), spectral tilt or slope of wavelet responses (peak-slope), Parabolic Spectral Parameter (PSP), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd) [69].

Lugger et al. use prosodic and voice quality features, namely Open Quotient Gradient, Glottal Opening Gradient, Skewness Gradient, Rate of Closure Gradient, and Incompleteness of Closure [70].

They use a two-level classification, where in the first level they classify two different activation levels (high and low), then each of these is classified using a Bayesian classifier. While prosodic features have 66.7% recognition rate alone, using prosodic and voice quality parameters increase the recognition rate to 74.5%. In a follow-up study, they use a three-level classifier and increase recognition rate to 88.3% [71].

Li et al. used shimmer and jitter as quality features and added them to spectral baseline features for classification of emotions [72]. HMM and GMM is used for classification and a higher recognition rate is achieved by adding jitter and shimmer.

Zhang used prosodic and voice quality features jitter, shimmer, HNR, and the first three formants [73]. They used the prosodic and voice quality features and achieved a 10% higher recognition rate when compared to the usage of prosodic features alone.

Borchert and Düsterhöft used quality features such as formants, HNR, jitter shimmer to utilize for valence dimension and prosody features for arousal axis [74]. They used several classifiers for emotion recognition including Bayesian Networks, SMO, Neuraral Networks, J48 Decision Tree. 70% average recognition rate is obtained for speaker-independent emotion recognition.

There are also qualitative speech features which describe speech signals by voice quality labels. These features are harsh, tense, breathy, modal, whisper, creaky and lax-creaky voices. These features have a high correlation with the perceived emotions. However, it's hard to extract and have a relativistic interpretation based on the researchers understanding [75].

Laver [76] has been associated breathy voice with intimacy, harsh voice with anger, whispering voice with confidentiality, and creaky voice with boredom. Scherer et al. have associated tense voice with anger, fear and joy; and lax voice with sadness [77]. In addition, Murray et al. have suggested associating breathy voice to anger and happiness; and associating sadness to resonant voice quality [78].

It can be said that the voice quality features are more supplemental than primary features for a speech emotion recognition system. Some of the studies list jitter, shimmer, and HNR under prosodic features [51, 67, 79].

5.2.4. Teager Energy Operator Based Features

There are features that depend on the Teager Energy Operator (TEO). It is used to detect stress in speech and has been introduced by Teager [80] and Kaiser [81, 82]. According to Teager, speech is formed by a non-linear vortex-airflow interaction in the human vocal system. A stressful situation affects the muscle tension of the speaker that results in an alteration of the airflow during the production of the sound. The operator developed by Teager to measure the energy from a speech by this non-linear process was documented by Kaiser as follows where $\Psi[]$ is Teager Energy Operator and $x(n)$ is the sampled speech signal.

$$\Psi[X(n)] = x^2(n) - x(n+1)x(n-1) \tag{2}$$

Zhou et al. proposed three new TEO-based features which are TEO-decomposed FM (frequency modulation) variation (TEO-FM-Var), normalized TEO auto-correlation envelope area (TEO-Auto-Env), and critical band based TEO auto-correlation envelope area (TEO-CB-Auto-Env) [83]. The variation

in energy of airflow characteristics in the vocal tract for voiced speech spoken under stress is explored among these features. They compared these with pitch and MFCC features with text-dependent and text-independent pairwise stress classifications using the SUSAS dataset. TEO-FM-VAR and TEO-AUTO-ENV did not perform well compared to classical pitch and MFCC features. However, TEO-CB-Auto-Env outperforms both pitch and MFCC under stress condition. Similar results have been shown by Low et al. in their study on the detection of clinical depression in adolescents [79]. They used prosodic, spectral, voice quality, as well as Teo-Based features. TEO-based features, specifically TEO-CB-Auto-Env, outperformed all other features including their combination.

5.3. Supporting Modalities

Several technologies are available that can be used for emotion recognition systems. These systems use modalities such as visual signals, physiological signals, word recognition, brain signals to classify emotions. Although as standalone systems, these technologies are used to recognize emotions, they are not yet successful enough to recognize emotions fully. However, they can be used as supporting methods to enhance the power of speech emotion recognition systems.

Systems that use numerous modalities to classify emotions are called multimodal or multi-cue fusion emotion recognition systems. In multimodal systems, a fusion of multiple modalities can be analyzed in four different classes: feature level, decision level, model level, and hybrid fusion [84].

In feature level fusion, feature vectors of different modalities are combined and a new feature vector is constructed before they are used in the classification. However, the new high dimensional feature set suffers from the curse of dimensionality, and cause data sparseness. To overcome this problem feature selection algorithms explained in Section 5.1.6 can be used.

In decision level fusion, each feature set from different modalities classified with domain-specific classifiers, and recognition results are combined by some criteria to obtain the final result. However, by using feature sets of each modality on separate classifiers, information correlation among feature sets are lost [85].

To overcome this correlation problem hybrid feature fusion and model level features fusions are proposed. These methods combine feature level and decision level fusion methods.

Model level fusion emphasizes the mutual correlation among the streams of multiple signals. Hidden Markov Model and Bayesian Network based systems are used for model-level fusion [86].

Hybrid fusion combines different levels of the fusion schemes to increase recognition rate [84]. In most cases of the hybrid fusion, features from different modalities first fused in feature level and classified using a classification algorithm. Then, each modality is classified with separate classifiers and decision level fusion is applied. Finally results from both classification are fused again, and the final result is obtained.

Visual signals and audio signals are complementary to each other[87]. Therefore, visual signals are the most used modality alongside speech signals to classify emotions, furthermore, they are easy to collect which can be acquired alongside speech using a single camera. A large number of audio-visual databases are available to use for multimodal classification. Furthermore, most of the research on multimodal emotion recognition is focused on audio-visual methods [86],[88],[89],[90],[91],[92],[93]. Some

of the audio-visual databases are shown in Table 2. They are mostly acquired from movies, reality shows or talk shows. Facial expressions, gestures, posture, body movements are the visual cues used alongside speech signals. Most of the research on emotion recognition system is focused on acted data. However, in recent years there is increasing attention on spontaneous data.

The research on physiological signals or biosignals with speech is scarce due to the need for a device to collect physiological signals. Currently, data collection is mostly done as elicited in the laboratory environment. While it is challenging to collect biosignals for both training and classification, they have the advantage of being uncontrolled; for example, people may hide their emotions in their speech; however, it is harder to alter biosignals [94].

Biosignals that can be used are Blood Volume Pulse (BVP), EMG (electromyogram), skin conductivity, skin temperature, respiration, heart rate, EDA (electrodermal activity), ECG (electrocardiogram), and PPG (photoplethysmography). In his study, Kim used biosignals and speech to classify emotions [95]. He used sequential backward selection for feature selection, and LDA for classification. For the fusion of the modalities, feature level, decision level, and hybrid fusion methods are used. Best results are obtained using feature-level fusion. It has been found that biosignal-audio fusion is not as complementary as audio-visual recognition systems.

Word recognition technology can also be used to enhance the performance of speech emotion recognition. In their study, Eyben et al. used low-level speech features and binary linguistic features for emotion detection using dimensional model [96]. They have used an interpreter component that assigns function related attributes to words like agree and disagree, or positive and negative. They performed classification separately using speech and linguistic features and also using feature level fusion combined with the modalities and Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) is used for classification. They showed that acoustic features performed better than linguistic features. However, best results are obtained when both modalities are combined. Wu et al. used Meta-Decision tree to classify speech data, and maximum entropy level to characterize the relationship between emotional states and the Emotion Association Rules that are extracted from a Chinese knowledge base [97]. Then, a weighted product fusion method is used to combine both to produce the final result. Speech-based recognition and semantic label-based methods achieved a recognition rate of 80% and 80.92%, respectively. Multimodal recognition achieved an 83.55% recognition rate.

Keystroke dynamics, mouse movement, and touch behavior are other technologies that can be used alongside speech. In an empirical study, Tsihrintzis et al. compared facial expression and keystroke as modalities to detect emotions [98]. They found out that anger and sadness can be recognized better using keystrokes dynamics, while facial expression based methods performed better on surprise and disgust. In their study, Khanna et al.[99] observed that approximately 70% of people's typing speed decreased when they are in a sad state, and 80% of people type faster when they in a happy state. There is no current study on multimodal emotion recognition using speech and keystroke dynamics. However, it can be studied especially for emotion recognition for people playing video games.

Table 3: Classifiers and features used in the literature

	Prosodic Features	Spectral Features	Voice Quality Features	Teo-Based Features	Other
HMM	Schuller et.al[62], Kwon et al.[100], Nogueiras et al.[60] Ververidis et al.[101]	Kwon et al.[100], Nogueiras et al.[60] Nwe et al.[102], Sato et al.[66]		Zhou et al.[83]	
GMM	Busso et al.[61], Kwon et al.[100] Luengo et al.[51], Low et al. [79]	Kwon et al.[100], Luengo et al.[51] Low et al.[79]	Low et al.[79]	Low et al.[79]	
SVM	Borchert et al.[74], Luengo et al.[51] Rao et al.[55], Schuller et al.[103, 52], Shen et al.[104] Low et al.[79]	Borchert et al.[74], Bitouk et al.[67] Hu et al.[105], Luengo et al.[51] Schuller et al.[52], Shen et al.[104] Low et al.[79]	Borchert et al.[74], Schuller et al.[52] Low et al.[79]	Low et al. [79]	
MLP	Nakatsu et al.[106], Schuller et al.[103, 52] Petrushin et al. [107] Nicholson et al. [108]	Nakatsu et al., Schuller et al.[52] Nicholson et al. [108]	Schuller et al.[52]		
kNN	Rong et al., Schuller et al.[103]	Rong et al.[53]			
Decision Tree	Borchert et al.[74], Lee et al.[109], Schuller et al.[103]	Borchert et al.[74], Lee et al.[109]	Borchert et al.[74]		
Rule Based Fuzzy Estimator	Grim et al.[23]	Grim et al.[23]			
Denoising Autoencoder	Deng et al.[110, 111]	Deng et al.[110, 111]	Deng et al.[110, 111]		
DNN	Han et al.[112]	Han et al.[112]			
CNN					Mao et al.[113], Trigeorgis et al.[114], Kim et al.[115], [116] Lim et al.[117]
RNN	Wöllmer et al.[118], Mirsamadi et al.[119], Eyben et al.[96] Lee et al.[120]	Wöllmer et al.[118], Mirsamadi et al.[119], Eyben et al.[96], Lee et al.[120]	Mirsamadi et al.[119], Lee et al.[120]		
Ensemble	Albornoz et al.[121], Schuller et al.[103] Wu et al.[97] Lee et al.[109]	Albornoz et al.[121], Wu et al.[97] Lee et al.[109]	Wu et al.[97], Lee et al.[109]		

5.4. Classifiers

Speech emotion recognition systems classify underlying emotions for a given utterance. Including traditional classifiers and deep learning algorithms, many machine learning algorithms are used to carry out the speech emotion recognition task. However, just as with any complicated problem, there is no generally accepted machine learning algorithm that can be used; current studies are generally empirical. In Table 4, studies are summarized including databases, features, classifiers, and result of the experiments. In addition, Table 3 presents the studies with features and classifiers used in them.

5.4.1. Traditional Classifiers

SER systems typically make use of classification algorithms. A classification algorithm requires an input X, an output Y, and a function that maps X to Y as in $f(X) = Y$. The learning algorithm approximates the mapping function, which helps predict the class of new input. The learning algorithm

needs labeled data which identifies the samples and their classes. Once the training is over, data that has not been used during training is used to test the performance of the classifier.

Most preferred algorithms are Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). There are also classification methods based on Decision Trees (DT), k-Nearest Neighbor (k-NN), k-means, and Naive Bayes Classifiers. In addition to usage of single classifiers, ensemble methods are also used for SER that combines several classifiers to obtain better results.

5.4.1.1. Hidden Markov Model. Hidden Markov Model is a commonly used method for speech recognition and has been successfully extended to recognize emotions, as well. As the name suggests, HMM relies on the Markov property which says that the current state of a system at a time t only depends on the previous state at time $t-1$. The term "hidden" denotes the inability of seeing the process that generates the state at time t. It is then possible to use probability to predict the next state by making observations of the current state of the system.

Nogueiras et al. used low-level pitch and energy features, and their contours using hidden semi-continuous Markov models [60]. They obtained a recognition rate of over 70% for 6 emotion class including happiness, anger, joy, fear, disgust, sadness.

Schuller et al. compared two methods [62]. In the first method, utterances are classified by GMMs using global statistics of features derived from the raw pitch and energy contour of the speech signal. In the second one, continuous HMM is applied using low-level instantaneous features rather than global statistics. The average recognition accuracy of seven discrete emotion classes exceeded 86% using global statistics, whereas the recognition rate of the human deciders for the same corpus is 79.8%.

Nwe et al. showed that LFPC feature on Hidden Markov Model (HMM) yields better performance than MFCC and LPCC [65]. They achieved a recognition rate of 77.1% and 89% for average and best recognition rate, respectively, while human recognition was 65.8%.

Lin et al. used HMM and SVM to classify five emotions, which are anger, happiness, sadness, surprise, and a neutral state [122]. For HMM, 39 candidate features are extracted, then SFS is applied for feature selection. Classification performance with selected features is compared to classification using MFCC. From the difference between Mel frequency scale sub-bands energies, a new vector is built and the performance of the K-nearest neighbor is tested using this newly built vector. For HMM, 99.5% accuracy is obtained for the speaker-dependent case, whereas the accuracy for SVM is 88.9%.

5.4.1.2. Gaussian Mixture Model. Gaussian Mixture Model is a probabilistic method which can be viewed as a special case of continuous HMM that contains only one state. The idea behind the mixture models is modeling the data in terms of a mixture of several components, where each component has a simple parametric form, such as a Gaussian. It is assumed that each data point belongs to one of the components, and it is tried to infer the distribution for each component separately.

Neiberg et al. compared MFCC and MFCC-low features obtained by placing filter banks in the 20 - 300 Hz region to model pitch feature as MFCC and plain pitch features with GMM as the classifier [123]. For classification, a root GMM is trained using Expectation Maximization (EM) algorithm with a max-

imum likelihood criterion. Later, from the root model using the maximum a posteriori (MAP) criterion, one GMM per class is constructed. For the training and test data, average log-likelihoods of n-grams using manual orthographic transcriptions are also used in addition to acoustic features. They created different GMMs for each of MFCC, MFCC-low, and pitch, and they also combined all three classifiers. They tested the classifiers with data from a Swedish company Voice Provider (VP), and The ISL Meeting Corpus (ISL). Best results are obtained by the combination of the classifiers using acoustic features using priors.

5.4.1.3. Artificial Neural Networks. Artificial Neural Networks are a commonly used method for several kinds of classification problems. It is basically constructed with an input layer, one or more hidden layers, and an output layer. The layers are made up of nodes; while the number of nodes in the input and output layers depends on the representation of the data and the labeled classes, the hidden layers can have as many nodes as required. Each layer is connected to the next using weights that are initially randomly chosen. When a sample is chosen from the training data, its values are loaded to the input layer, and then forwarded to the next layer. At the output layer, the weights are updated using the backpropagation algorithm. Once the training is complete, the weights are expected to be able to classify new data.

Nicholson et al. used ANN for a speaker and context independent SER system [108]. They selected speech power, pitch, LPC and delta LPC as parameters. For each emotion in the database, they created a sub neural network. Each neural network output a value representing the likelihood that the utterance corresponds to an emotion. Based on this value, the best prediction is selected by the decision logic. They obtained a 50% average recognition rate using this one-class-in-one neural network.

Petrushin et al. developed an application to be used in call-centers [107]. They tested their system using ANN, and an ensemble of ANN and k-nearest neighbor (kNN) classifiers. They selected some statistics of the pitch, energy, the speaking rate and the first and second formants as a feature set. 55%, 65%, and 70% of average accuracies are obtained for kNN, ANN, and the ensemble of ANNs, respectively.

5.4.1.4. Support Vector Machine. SVMs are supervised classifiers which find an optimal hyperplane for linearly separable patterns. Given the training data the objective of an SVM classifier is to find the hyperplane that has the maximum margin, between data points of both classes. If these patterns are not linearly separable, using a kernel function original data points are mapped to a new space.

Kwon et al. extracted pitch, log energy, formant, Mel-band energies, and MFCCs as base features [100]. To consider the speaking rate and model the dynamics of the corresponding temporal change of pitch and spectrum, velocity and acceleration information are included for pitch and MFCCs, respectively. They gathered these features from utterances and calculated the statistics of the features. Finally classified the signals using LDA, QDA, binary Gaussian kernel SVM (GSVM), linear SVM (LSVM), HMM. Pair-wise classification and Multi-class classification is tested. GSVM and HMM give the best results with accuracies 42.3% and 40.8%, respectively.

Shen et al. using SVM classified the emotions in Berlin Database [104]. They compared energy, pitch, LPCMCC features. 66.02%, 70.7%, and 82.5% classification accuracies are obtained for using only energy and pitch, for using LPCMCC, and for the combination of prosodic and spectral features,

respectively.

Hu et al. used the GMM supervector based SVM with spectral features for classification [105]. For each emotional utterance in the database, a GMM is trained, later the corresponding GMM supervector is used as the input for SVM. They also compared result using ordinary GMM. While the accuracy is 82.5% for GMM supervector based SVM, 77.9% accuracy is obtained using GMM.

Pan et al. used SVM with pitch, energy, MFCC, LPCC, MEDC features to classify Berlin database and their self-built Chinese emotional database [124]. They trained different SVM classifiers for combinations of the features. For the Chinese database, best results are obtained using MFCC, MEDC, Energy features with an accuracy of % 91.3043 while it's %95.087 for Berlin database using the same combination of the features.

Schuller et al. used a Multilayer SVM for speech emotion recognition [125]. Multilayer SVM has several input layers, then has a hidden layer which may contain a number of SVM layers, and finally, an output layer similar to neural networks but trains Support Vector Coefficients and the biases of all SVMs in the architecture, instead of weights. In this study, a layer-wise binary classification is repeatedly made until only one class remains. As a result of the experiments, it is found out that classes which are hardly separable should be divided at last. They also tested regular SVM, MLP, GMM, kNN, and k-nearest neighbor algorithms. Best results are obtained by ML-SVM by an 18.71% error rate. In a second approach, they added the spoken content as supporting modality for emotional key-phrases by applying Belief Network based spotting. They further improved recognition rate by combining acoustical and linguistic features using MLP as soft decision fusion. The error rate is decreased to 8%.

Truong et al. used Support Vector Regression (SVR) with a RBF kernel to compared self-reported emotion ratings to observed emotion ratings and tried to see how differences between two ratings affect development of emotional speech recognizers in a two-dimensional arousal-valence space [126]. They used acoustic and lexical features to be used for the classification. They found that observed emotions are easier to recognize than self-reported emotions and that averaging ratings from multiple observers increases the performance of the emotion recognizers.

5.4.1.5. Ensemble of Classifiers. In ensemble learning, a number of machine learning algorithms are combined to increase predictive performance. Each algorithm in ensemble classifier is combined in some way, typically by a voting procedure, to obtain a final result. Performances of the ensembles are often higher than the individual classifiers. Different types of architectures are available in ensemble classifiers. One of the ways is feeding the same data to each classifier by comparing the results obtaining a final decision. Another approach is using the hierarchical classifier. In this approach, input data is fed to one algorithm, then the result is fed to another type of classifier in a hierarchical approach, then the final decision is given.

Xiao et al. proposed an hierarchical two-staged classifier based on a "Dimensional Emotion Classifier" [127]. At stage 1 emotional states classified according to arousal dimension in two sub-stages into three classes as active, median, and passive. At stage 2, member of these three classes further classified based on appraisal dimension. At each classification stage, most relevant features are selected

using SFS algorithm. They used back-propagation neural network as classification algorithm. Berlin and DES databases are used as dataset which are based on categorical emotional model. These categories are mapped into the dimensional space. For Berlin dataset 68.60% recognition rate is obtained and the result is increased to 71.52% when gender classification is applied prior. With DES dataset 81% recognition rate is obtained.

Lee et al. proposed and hierarchical classifier for speech emotion recognition task [109]. The proposed method classifies an input utterance through sequential layers of binary classifications. They used Bayesian Logistic Regression as a binary classifier. They obtained an improvement of 7.44% over a baseline SVM classifier.

Wu et al. used acoustic features and semantic labels for emotion classification [97]. They used GMM, SVM, and MLP as base level classifiers. A Meta Decision Tree is used for fusion of base classifier to get recognition confidence for acoustic features. For semantic labels, Emotion Association Rules (EARS) are extracted and using maximum entropy model to identify the relationship between emotional states and EARs. Finally, the weighted product fusion method is used to combine acoustic features and semantic labels. They achieved 85% average recognition rate by combining acoustic and linguistic information.

Albornoz et al. used two-stage hierarchical classifier for a SER system [121]. When the number of emotional classes to recognize is decreased, higher performance is obtained. Furthermore, different emotions can be represented better with different features and different classifiers. Hence hierarchical classifier is chosen for this task. In each stage of the two-stage hierarchical classifier, separate feature extraction and classification steps are applied. They created two variants of the hierarchical classifier in the first one, at the first stage emotions are partitioned into three groups namely disgust, BNS (boredom, neutral, surprise), and JAF (joy, anger, fear), whereas in the second one at the first stage into two groups: BNS and JAFD. In the second stage of both variations, specific features are extracted for each group and these are fed into group-specific classifiers. Finally, individual emotions are obtained. They tested different combinations of features and classifiers. Experimental results showed that HMM with 30 Gaussians in mixtures are best for stage I, and for Stage II HMM is better to discriminating Boredom, Neutral, Sadness whereas MLP is better for Joy, Anger, Fear. They also showed that 12 mean MFCC features and their deltas and acceleration coefficients are best features for Stage I and BNS group for Stage II, 12 mean MFCC, 30 mean log-spectrum, mean and standard deviation of the fundamental frequency and pitch are the best features for Stage II JAF group.

5.4.2. Deep Learning Based Classifiers

Most of the deep learning algorithms are based on Artificial Neural Networks hence are commonly referred to deep neural networks. The term “deep” comes from the number of hidden layers as it can reach to hundreds of layers, whereas a traditional neural network contains two or three hidden layers. In recent years, performance of the deep learning algorithms surpass the traditional machine learning algorithms, hence the focus on research changed direction towards them and the current trend in SER research is no different. The advantage of some of these algorithms is that there is no need for feature extraction and feature selection steps. All features are automatically selected with deep learning algorithms. Most widely

used deep learning algorithms in SER domain are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN).

5.4.2.1. Recurrent Neural Networks. RNNs are a family of neural networks which are specialized in processing sequential data. By the usage of internal memory, they can remember the received input data and make a precise prediction about what is coming next. Because of their nature, RNNs are successfully used for sequential data such as time series, speech, text, video.

When a unit of RNN produces an output, it forwards data to the next unit and also loops the output back itself. As a result, it has two types of input: present input and input from the recent past. The input from the recent past is important because the sequence of the data contains important information about what is coming next.

RNNs have a short time memory, however, by using Long-Short Time Memory architecture, RNN can gain access to long term memory. LSTM-RNNs are a kind of gated RNN which are the most effective models used in practical applications that solves the long term dependency problem of the RNN. LSTM-RNNs have special "LSTM cells" that have internal recurrence besides the outer recurrence of RNN. In addition to standard input and output of the RNN, it has more parameters and gating units with sigmoid nonlinearity that control the flow of information. LSTM has three types of gates: input gate, forget gate and remember gate. By opening and closing these gates, LSTM cell makes decisions about what to store, and when to allow inputs, outputs, and deletions.

Eyben et al. proposed an online SER system using LSTM-RNN [96]. They added the time dimension to 2D activation-valence emotional model to create a new 3D model. The motivation behind adding the third time dimension to evaluating emotions incrementally in real-time. Since LSTM-RNN does not require any segmentation besides the low-level framing, it is suitable to use for real-time applications. Another factor to use LSTM-RNN is the fact that its suitability for the connected time series. Prosodic, spectral and voice quality features are used as acoustical features. In addition, linguistic features are extracted, and acoustic and linguistic features are combined by the feature level fusion.

Tian et al. used hierarchical fusion to combine acoustic and lexical features for emotion recognition with dimensional emotional model [128]. They used LSTM for classification task with LLD, eGeMAPS, and global prosodic features are used as acoustic features, and Disfluency and Non-verbal Vocalization (DIS-NV) Features, Point-wise Mutual Information (PMI) Features, and Crowd-Sourced Emotion Annotation (CSA) features are used as lexical features. In hierarchical fusion, an LSTM network with three hidden layers is proposed. In the first layer LLD and eGeMAPS features are used in first hidden layer, later GP and DIS-NV features are added, finally PMI and CSA features are used in third layer. Also a network with two hidden layer with GP, DIS-NV, and CSA features are proposed. AVEC 2012 and IEMOCAP databases are used for experiments. Better results are obtained with hierarchical fusion compared to feature level and decision level fusions.

Wöllmer et al. used LSTM-RNN for continuous emotion recognition in a 3D space spanned by activation, valence, and time [118]. For experiments Belfast Sensitive Artificial Listener data which is part of the HUMAINE database is used. Best results for activation (MSE 0.08) is obtained by LSTM-

RNN. However, for the valence dimension LSTM-RNN and SVR performed equally (MSE 0.18) They also observed that classification performance of the valence was relatively low as the detection of valence from only acoustic features is known to be hard. Hence they suggest additional modalities to be used, such as linguistic features for valence classification. Similar results are obtained by other studies which indicated that valence estimation is higher with semantic features while arousal elicitation is better with acoustic features [129, 130].

Kaya et al. investigated LSTM-RNN for cross-corpus and cross-task acoustic emotion recognition using dimensional emotional model [131]. They employed an approach to utilize the frame level valence and arousal predictions of LSTM models for utterance level emotion classification. They combined discretized predictions of LSTM models with the components of the baseline system. Baseline system used SVM and learner performance is increased further using least squares based weighted kernel classifiers. Results from LSTM and baseline system is combined with Weighted score level fusion. For cross-corpus experiment RECOLA, SEMAINE and CreativeIT datasets are used [132]. Their results showed the suitability of the proposed method for both time-continuous and utterance level cross-corpus acoustic emotion recognition tasks.

5.4.2.2. Convolutional Neural Networks. Convolutional Neural Networks (CNNs) are particular types of neural networks which are designed to process data that has a grid-like topology, such as images. Through applications of several relevant filters, CNN can successfully capture temporal and spatial dependencies from an input source. The inputs are reduced into a form without loss of feature so that computational complexity decreases and the success rate of algorithm is increased. A CNN is composed of several layers: convolution layer, polling layer, and Fully-Connected layer.

A convolution layer is used to extract high-level features from the input. Mathematically a convolution means combining two functions to obtain a third one. In CNN, the input is taken and, then a kernel is applied to it. The resulting output is a feature map.

Polling layer is used to reduce the size of convoluted features to decrease computational complexity through dimensionality reduction. It is useful for extracting dominant features of the input data.

After passing input from several convolution and polling layers and extracting the high-level features, the resulting features are used as an input to a fully connected layer by flattening the 2D data to a column array and feeding it to a feed-forward network that operates as an ordinary neural network.

Trigeorgis et al. proposed a system that combines CNN with LSTM networks, where CNN is used to automatically learn the best descriptive characteristics of the speech signal directly from the raw time representation [114]. Speech signals are segmented first, then denoising is applied as preprocessing step. Next, using CNN, acoustic features are extracted. Finally, LSTM layered deep RNN are fed with extracted features. Significantly better performance is obtained by the proposed method compared to traditional designed features. A similar approach is taken by Lim et al. [117]. They compared the proposed time distributed CNN to CNN and LSTM-RNN They obtained 88.01% average precision using the proposed method, whereas single CNN classifier and LSTM classifier obtained average precision 86.32% and 78.31%, respectively.

CNNs can be built by different dimensionalities. Zhao et al. used 1-D and 2-D CNNs with LSTM network for speech emotion recognition [116]. 1-D CNN is built to learn to local and global emotion-related features from speech whereas 2-D CNN is used to learn Mel-spectrogram. Both networks share a similar architecture; both have four local feature learning blocks (LFLBs) and one LSTM network. LFLB network contains a convolution layer and a polling layer. The convolution and pooling kernels in each LFLB are all one-dimensional in 1-D CNN since it learns features from the raw speech signal. The learned features are fed to LSTM layer to get contextual dependencies. In 2-D CNN, Mel-spectgrams in the form of 2-D matrices are fed to CNN to learn high-level emotional features. Features are extracted from and again fed to LSTM. In overall, the 2-D network performed better than the 1-D network. On Berlin EmoDB, 2-D network obtained recognition rates of 95.33% and 95.89% with speaker-dependent and speaker-independent experiments respectively, whereas recognition rates are 89.16% and 52.14% on IEMOCAP database for speaker dependent and speaker independent cases, respectively.

While most of SER systems using CNN, at least one LSTM network is added to deal with temporal dependencies and spectral variations; however, it increases the depth and complexity. Instead, Kim et al. proposed a 3-D CNN to learn spectro-temporal features [115]. First, two seconds of segments are extracted from utterances, padding is applied, then 256 point spectrogram are extracted for every 20 milliseconds. Total of 100 frames is obtained. Finally, a temporal series of 2-D feature maps with a resolution of 10 x 256 are composed. Spectral features are represented as feature maps in short-term windows that are 200 milliseconds long. Each utterance segment has a resolution of 10x10x256 that are denoted as short term (T), Long term (L), and spectral (S). These feature maps are fed into 3-D CNN, and 3-D max polling is applied. Then, two methods are tested for learning. First, 3-D output features are flattened into 1-D vectors and they are forwarded to fully-connected layers with an additional softmax layer. This method is named as 3D-CNN-DNN. In the second method, the 3-D output is transformed into 2-D output features and these are forwarded into a temporal series of a fully connected layer. Extreme Learning Machine (ELM) is used in this method, therefore it is named as 3D-CNN-DNN-ELM. Both methods are compared with off-the-shelf LSTM-ELM and DNN-ELM 1D-CNN-LSTM, 2D-CNN-LSTM, and 2D-CNN-LSTM-DNN. Both of the proposed methods outperformed off-the-shelf methods.

5.4.3. Machine Learning Techniques for Classification Enhancement

5.4.3.1. Autoencoders. Finding labeled data is a challenging task for SER research and applications. Besides, even if the labeled data is obtained, there is no guarantee on the correctness of the emotional labels since there is no standardization on the labeling task. In recent years, auto-encoders gained attention due to their unsupervised, and semi-supervised nature. Autoencoders consists of three layers as other neural networks, an input layer and an output layer of the same size, and hidden layers that contain fewer neurons than the input and output. Autoencoders reconstructs the original input data as output. It has two main parts: an encoder and a decoder. The encoder compresses the input data and transforms into a more dense representation, whereas the decoder part reconstructs the data. In the training phase, the reconstruction error is computed between input and output and tune the network for better representation. Autoencoders generally work as a feature extractor rather than a classifier.

After training the autoencoder, encoder part is connected to a classifier. There are several types of autoencoders such as variational autoencoder (VAE), denoising autoencoder (DAE), sparse autoencoder (SAE), adversarial autoencoder (AAE).

Eskimez et al. used and compared denoising autoencoder, variational autoencoder adversarial autoencoder, and adversarial variational Bayes as feature extractor and fed these learned features into a CNN [133]. These systems compared with baseline SVM and CNN using hand-crafted features. They used USC-IEMOCAP audio-visual dataset to test the systems. Best results are obtained by adversarial variational Bayes which is followed by adversarial autoencoder.

To learn latent representations of speech emotion, Latif et al. proposed a system which is using variational autoencoders [134]. In addition, they tested a type of VAE - Conditional Variational Autoencoder (CVAE). In their study, autoencoders are used to learn the representation while an LSTM network is used as the classifier. They also compared the performance with an autoencoder-LSTM as well as with CNN and BLSTM using hand-crafted features. They used USC-IEMOCAP dataset for testing. The best performance is obtained by CVAE-LSTM with a 64.93 % weighted average.

Sahu et al. inspected Adversarial autoencoder and conducted different experiments on two points [135]. Their first experiment aimed to examine classification performance of autoencoder's compression ability which encodes high dimensional feature vector representation into a lower dimensionality. They also inspected the regeneration of synthetic samples to be used in the training phase. They compared the performance of code vector learned from AAE with openSMILE features classified by SVM and also a lower dimensional representation of these features reduced by PCA and LDA methods. Code vector obtained by AAE showed a close performance compared to openSMILE feature and outperformed other compression techniques. Synthetic samples generated from AAE also showed promising results. Using synthetic samples alongside with original samples increased recognition rate compared to using only original samples.

Deng et al. proposed a Semi-supervised Autoencoder (SS-AE) for speech emotion recognition [136]. It combines discriminative and generative approaches. When supervised classifier learns from the labeled data, it also predicts all unlabelled data in parallel, hence aids explicitly to supervised learning by the incorporating preceding information from unlabelled samples. This is carried out by appending an additional class to the supervised task. A joint objective function is created which minimizes the reconstruction error of unsupervised objective and prediction error of supervise objective. To point out the problem of exploding and vanishing gradient problem, a variation of SA-AR that has skip connections between layers is proposed that is called SA-AE-Skip. With these connections, information can smoothly flow across the several layers during the training. Using the proposed system, need for a large number of the training sample is reduced, as well as the fundamental knowledge from unlabelled data to supervised learning is transferred.

5.4.3.2. Multitask Learning. Most of the SER systems, are focused on single task learning (STL) which aims to learn and predict the emotion in the utterance. However, various studies show that multitask learning (MTL) improves the recognition rate significantly. MTL is a machine learning technique where

several tasks are learned simultaneously by using a shared representation. The learner uses similarities between the tasks leading to improved generalization. It is referred to as inductive transfer that improves generalization by utilizing the domain information extracted from the training signals of tasks as an inductive bias [137]. Generally, emotion recognition is designated as a primary task and several other tasks such as gender, spontaneity, naturalness classification are selected as auxiliary tasks. The succession of the MTL is heavily depended on the selection of the subtasks.

Kim et al. proposed an MTL approach that uses emotion recognition as primary task; and gender and naturalness as auxiliary ones [138]. They tested the proposed method using within-corpora and cross-corpora setups. For experiments, they created two variants of the MTL one using LSTM and the another DNN. They also compared the performance using STL based LSTM and DNN. For cross-corpus setup, they used LDC Emotional Prosody, eNTERFACE, EMODB, FAU-Aibo emotion corpus, and IEMOCAP. For within-corpora experiments, their gain was not significant. However, significant gains have been obtained for large datasets such as AIBO and IEMOCAP. For cross-corpora experiments, MTL outperformed STL and got a significant gain while gender and naturalness subtasks are used together for large corpora.

Mangalam et al. used spontaneity classification as an auxiliary task to MTL [139]. They compared the results with a hierarchical model which performs first a spontaneity detection before the classification process. For classification, they used an SVM classifier. For hierarchical architecture based on spontaneity detection samples classified with different classifiers. The classification is performed using Interspeech 2009 emotion challenge features and USC-IEMOCAP dataset. Proposed methods are compared with SVM, RF, CNN-based and representation learning-based emotion recognition and LSTM baseline classifiers. Best results are obtained by hierarchical classifier followed by MTL classifier.

In most of the speech emotion recognition systems that use dimensional emotional model, each dimensional attribute is learned separately. MTL can be used to classify emotions by jointly learning different dimensional attributes such as arousal, valence, and dominance simultaneously. Parthasarathy et al. using the interrelation between the dimensions proposed a unified framework to jointly predict the arousal, valence and dominance dimensions [140]. Prediction of the emotional attributes is formulated as a regression problem that is solved using DNN. The acoustic features are taken as inputs and mapped into attribute scores. The scores are jointly learned to predict the values of the attributes for dimensions. Two versions of MTL is proposed: one share the hidden layers between all three-dimensional attributes and one that only shares nodes in the first hidden layer which creates a shared feature representation. In the second hidden layer, nodes are separately connected for each representation. Within-corpora and cross-corpora experiments are conducted using a baseline STL classifier and proposed two MTL classifiers. For within-corpora experiments, MSP-PODCAST corpus is used and MTL outperformed STL. For cross-corpora experiments, IEMOCAP and MSP-IMPROV datasets are used. Due to the training and test data mismatch performance is decreased compared to the within-corpora experiment. MTL systems performed better or equal than STL setup.

Lotfian et al. used primary and secondary emotions for MTL within spontaneous emotion recognition context [141]. Spontaneous emotions are annotated with perceptual evaluations. Hence, the annotators

label the utterances based on their perspective and multiple answers can be given by them especially when many related emotional categories are available which creates an ambiguity problem. To overcome this problem, primary and secondary emotions are used. For each sample, primary and secondary labels are generated, then a classifier is trained using MTL. The primary task in the MTL is to find the most relevant category and the auxiliary task is defining all of the labels that are relevant. For experiments, two baseline classifier is created. In the first one, hard labels are obtained from primary emotions using majority voting for draining. In the second one, soft-labels are used that are derived from the primary labels. As a result, higher performance is obtained using MTL than STL with 66.8% accuracy.

Le et al. used MTL BLSTM-RNN to classify emotional speech signals [142]. Their approach has three steps. In the first step the continuous emotional labels(valence, arousal) are discretized and mapped to small set of discrete emotion classes by applying k-means. Then a MTL BLSTM-RNN with cost sensitive Cross Entropy loss is trained to jointly predict label sequences at different granularity. In the end, a decoding framework which incorporates an emotion "language model" to produce more robust time series estimates. The experiments are conducted on RECOLA dataset. They achieved competitive results compared to previously published results.

5.4.3.3. Attention Mechanism. In recent years, the attention mechanism for deep learning gained success within the context of speech emotion recognition [119, 143, 144, 145, 146]. It ensures that the classifier pays attention to the specific locations of the given samples based on the attention weights given each portion of the input. Emotions are not evenly distributed over the whole utterances, rather they are observed on the specific portion of the utterances as mentioned earlier. In speech emotion recognition, this attention mechanism is used to focus on the emotionally salient portion of the given utterance.

Huang et al. proposed a Deep Convolutional Recurrent Neural Network (CLDNN) with an attention mechanism [143]. They investigated the role of the CNN for speech emotion recognition, compared CNNs task-specific spectral decorrelation to discrete cosine transform (DCT) under clean and noisy conditions, and explored context information for attention weight generation. The proposed system consists of a convolutional layer, a temporal layer based on BLSTM, a convolutional attention layer, and a fully connected layer. Convolutional layer extracts the high-level representation from log-Mels that provide complementary information compared to the raw waveform signal and also allow to directly quantify the advantage of CNN's task-specific decorrelation over that by the DCT. The extracted high-level representation is then fed into the BLSTM to learn temporal dependencies. After that, the convolutional attention layer locally gathers context information and to learn the weights. Finally, the output of the attention layer is fed into a fully connected layer for classification. The proposed method is compared with a baseline SVM that uses feature sets from INTERSPEECH Challenges from 2009 to 2013, as well as classifiers such as Sparse kernel reduced rank regression (SKRRR), BLSTM with the fully connected network (LDNN), and CLDNN without attention mechanism. The eNTERFACE'05 dataset is selected for experiments. Proposed method outperformed the other classifiers with unweighted accuracies of 84.00% 91.67% for noisy and clean conditions, respectively. They also showed that CNN's task-specific spectral decorrelation considerably outperforms that of the DCT.

Mirsamadi et al. used RNN to learn features for SER [119]. They introduced a novel weighted-polling method inspired by the attention mechanisms of neural machine translation [147]. Based on the weights which are determined by additional parameters of the attention model, a weighted sum is computed. Simple logistic regression is used as the attention model. The parameters of both the attention model and RNN are trained together. This attention model automatically removes the silent parts of the utterance by assigning small weights to these parts. Additionally, as each part of an utterance carries different emotional power, weights are assigned to each of them accordingly. It has the ability to consider the emotional content of different portions of speech. They compared the proposed system with an SVM based classifier and obtained higher accuracy by the proposed system.

Chen et al proposed a 3-D Convolutional Recurrent Neural Network (3-D ACRNN) with an attention model [144]. For input they used 3-D log Mel Spectrogram which consists of feature channels of static, delta and delta deltas. Delta and delta deltas contain effective emotional information and reduce the influence of emotionally irrelevant factors that increases the recognition rate. The input is passed through a convolutional layer to extract high-level representation. Then, it is passed to LSTM for temporal summarization. The output of LSTM is passed to an attention layer to focus on the emotionally salient part of the utterances. Normalized importance weights are computed by a softmax function. From these weights, utterance level representation is calculated by a weighted sum. Finally, utterance level representation is passed to a fully connected network to classify the emotion. The proposed system is compared with DNN Extreme Learning Machine (ELM) and 2-D ACRNN. The experiments are performed using IEMOCAP dataset and EMO-DB. Proposed systems outperformed DNN-ELM with an improvement of 13.5% and 11.26% for IEMOCAP and EMO-DB, respectively. Also, improvement of 2.34% and 3.44% are obtained on 2-D ACRNN for IEMOCAP and EMO-DB, respectively.

5.4.3.4. Transfer Learning. Finding labeled data to be used in training for speech emotion recognition is relatively hard compared to tasks such as automatic speech recognition, speaker recognition. The low number of data affects the recognition rate in a negative way due to the high variance. One of the methods intended to solve this problem is transfer learning. It's a machine learning technique where the knowledge obtained from a source learning task is transferred to be used as a starting point for a target model on a different but related task. However, in order to use transfer learning the source model need to be general enough. The most common approach for transfer learning is to train a source model with a set of source data or use a pre-trained model, then use the learned knowledge as a starting point on a related task. Optionally, the model may need to be fine-tuned.

Deng et al. used a sparse autoencoder to transfer knowledge from one model to another in order to boost the performance [110]. A single layer autoencoder is used to learn a representation trained on class-specific samples from target data. Then, this new representation is applied to the source data corresponding to a specific class in order to reconstruct and use it for the classification task. They used six databases namely FAU AEC, TUM AVIC, EMO-DB, eNTERFACE, SUSAS, and VAM. The results showed that the proposed system effectively transfer knowledge and increases the recognition rate.

Para-linguistic tasks are appropriate to be used with transfer learning for speech emotion recognition.

Gideon et al. transferred the knowledge between emotion, speaker, and gender recognition tasks [148]. They proposed to used Progressive Neural Networks (ProgresNet) for transfer learning [149] by extending the problem to be used for the cross-corpus task. Progressive neural networks avoid the forgetting effect where the target model lost its ability to solve the source task during the fine-tuning of the model using initial weights learned from a source task. They compared the performance of the proposed system with DNN and transfer learning by pre-training and fine-tuning (PT/FT) on MSP-IMPROV and IEMOCAP datasets. For paralinguistic experiments, knowledge transfer from gender and speaker recognition to emotion recognition are tested. When knowledge from speaker recognition is transferred to emotion recognition, ProgresNet outperformed DNN and PT/FT with both datasets. On the other hand, on the transfer of knowledge from gender recognition to emotion recognition, while ProgresNet surpassed other methods on MSP-IMPROV dataset, on IEMOCAP it failed. In the cross-corpus experiment, ProgresNet outperformed the other methods. Researchers concluded that due to the higher number of weights to transfer knowledge, ProgresNet is most useful when the target dataset is larger.

The knowledge from image processing domain can be also exploited for and transferred to speech emotion recognition domain in order to increase the classification task. In recent years, pre-trained Convolutional Neural Networks such as AlexNet or ImageNet which are trained by millions of images are extensively used for image classification tasks. Stolar et al. by using spectrograms formulated SER task as an image classification problem [150]. They proposed two methods namely AlexNet-SVM and FTAlexNet. For AlexNet-SVM first, the spectrogram images transformed into RGB images. These RGB spectrograms passed into pre-trained Convolutional Neural Network AlexNet as input to provide the feature for SVM which is used for classification. In FTAlexNet method, RGB images are applied to fine-tune AlexNet to provide emotional labels. The experiments are conducted on EMO-DB for males and females. Experiments showed that FTAlexNet obtained an average recognition rate of 76%, whereas the recognition rate for AlexNet-SVM was 68%.

Transfer learning can also be used in cross-corpus and cross-language setting in order to transfer information gained from one corpus can be transferred for another one. Latif et al. used Deep Belief Network (DBN) for their experiments as it has strong generalization power [151]. Their approximation is powerful for any distribution and their building blocks are universal approximator. They compared the proposed system to Sparse Autoencoders with SVM. They used FAU-AIBO, IEMOCAP, EMO-DB, SAVEE, and EMOVO datasets for experiments. In within corpus experiment, proposed system surpass sparse Autoencoder. Later these results are used as a baseline for cross-language experiments. For cross-language experiments, two different settings are tested. In the first one, FAU-AIBO and IEMOCAP datasets are used for training and other datasets for testing. In the second one, leave one out approach is used. Experiments show that leave one out approach obtained the highest accuracy. It yields training the model with a wider range of languages would help to capture intrinsic features from each language which provides a higher recognition rate.

5.4.3.5. Adversarial Training. In recent years, adversarial training gained a lot of attention from Machine learning community. Studies show that machine learning systems are vulnerable to adversarial

examples which are samples that exposed to small but intentionally worst case perturbation. These examples are incorrectly classified with high confidence [152]. Adversarial training is used to increase the recognition rate for speech emotion recognition system where the models are trained with both real and adversarial samples. Large perturbations in model output are penalized by the Adversarial training when small perturbation are added to training samples [153].

In speech emotion recognition systems most of the classifiers are trained by using samples which are recorded in studios. When these systems used in the real-life data, due to the data distribution between training data an testing data, misclassification is observed and recognition rate decreases. The labeling process is costly, and the abundance of unlabeled sample is available, hence we need to take advantage of unlabeled data as much as possible. Abdelwahab et al. used Domain Adversarial Neural Network (DANN) to find a common representation between training data and test data [154]. The proposed system has an adversarial multitask training phase to extract this common representation. The primary task is to predict emotional attributes such as arousal, valence, and dominance whereas the auxiliary task is to learn a common representation between source and target domains. The network consists of two classifiers: the main emotion classifier and a domain classifier that determines whether an input sample is from source domain or target domain. Both classifiers have common layers for starting, then each classifier branch out. The primary task is trained with source data whereas the domain classifier uses data from both labeled and unlabeled data. The classifiers are trained in parallel. The representation is learned using a gradient reversal layer (GRL). The experiments are conducted using IEMOCAP and MSP-IMPROV datasets as source data and MSP-Podcast dataset as the target. Two different baselines are established. The first network is trained and tested using only source samples. In the second one, the network is trained and validated using target samples. Both of the baselines are lack of GRL which is responsible for domain classification. With the proposed system on average 27.3% relative improvement in concordance correlation coefficient is obtained.

Han et al. proposed a conditional adversarial training framework to predict dimensional emotional representation namely arousal and valence [155]. The proposed system contains two networks. The first network tries to generate predictions for emotions from the given features, whereas the second one tries to differentiate the prediction obtained from the first network and the original samples while acoustic features used as conditionals. The proposed method is tested using RECOLA dataset and compared to LSTM-RNN as baseline classifier, CNN-LSTM, prediction based learning [156], and Reconstruction-error-based learning [157]. The proposed method gives the best results for both arousal and valence dimension.

Sahu et al. proposed a system to smoothing model predictions using adversarial training [153]. The smoothness is enforced by manifold regularization. They investigated two different training procedures; adversarial training and virtual adversarial training. In the first one, adversarial direction is determined based on the given labels for the training sample, whereas for the second one it is determined based on the output distribution of the training samples. The performance was evaluated using IEMOCAP dataset whereas cross-corpus experiments are conducted on SAVEE, EMA, and LDC datasets. Deep Neural Network is selected used as a classifier for the experiments. DNN with adversarial training surpassed the

baseline DNN. On the cross-corpus setting, DNN with adversarial training is also performed best with accuracies of 46.25%, 61.65%, and 43.18% for SAVEE, EMA, and, LDC, respectively.

Table 4: List of studies

Paper	**Dataset**	**Features**	**Classifier**	**Results**
Albornoz et al.[121]	Berlin Emo DB	Mean of log spectrum, MFCC, and prosodic features	Hierarchical classifier using HMM, GMM, and MLP	71.5% average recognition rate
Bitouk et al.[67]	LDC, Berlin Emo DB	Spectral features	SVM	46.1% recognition rate for LDC by using group-wise feature selection with class level spectral features, 81.3% recognition rate for EMODB by rank search subset evaluation feature selection with combined class level spectral features and utterance level prosodic features.
Borchert et al.[74]	Berlin Emo DB	Formants, spectral energy distribution in different frequency bands, HNR, jitter, and shimmer.	SVM, J48	90% recognition rate for single emotion recognition, 70% for all emotions
Busso et al.[61]	EPSAT, EMA, GES, SES, WSJ	Features derived from the F0 contour.	GMM	77% average recognition rate

Journal Pre-proof

Table 4: List of studies

Paper	**Dataset**	**Features**	**Classifier**	**Results**
Deng et al.[110]	AVIC, EMODB, eNTERFACE, SUSAS, VAM	LLDs such as ZCR, RMS, energy, pitch frequency, HNR, MFCC	Denoising autoencoder	For AVIC 62.7% recognition rate, for EMODB 57.9%, for eNTERFACE 59.1% for SUSAS 59.5%, for VAM 60.2%
Deng et al.[111]	AIBO DB, ABC DB, SUSAS DB	Low-Level Descriptors	Denoising autoencoders and SVM	64.18% average recognition rate for ABC DB, 62.74% average recognition rate for SUSAS DB
Grim et al.[23]	EMA DB, VAM I-II DBs	Pitch related features, speaking rate related features, spectral features	Rule based fuzzy estimator and SVM	0.27, and 0.23 mean errors for VAMI, and VAMII, respectively for both gender. 0.19 mean error for EMA DB.
Han et al.[112]	IEMOCAP DB	MFCC features, pitch-based features, and their delta feature across time frames	DNN and Extreme Learning Machine	54,3% average recognition rate

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Hu et al. [105]	8 native Chinese speakers (4 females and 4 males) uttered each sentence in five simulated emotional states, resulting in 1,600 utterances in total.	Spectral features	GMM supervector based SVM	82.5% recognition rate for mixed gender, 91.4% for male, 93.6% for female
Kwon et al. [100]	SUSAS DB and AIBO DB	Prosodic and spectral features	GSVM and HMM	For SUSAS DB using GSVM 90% 92.2% recognition rates are obtain for neutral and stress speech, resp vely. Using HMM 96.3% recognition rat is obtained. Recognition rate is 70 for 4-class style classification with M. For multiclass classification on AI DB, GSVM achieved an average recogn n rate of 42.3%.The average recogni rate is 40.8% using HMM.

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Lee et al. [109]	AIBO DB USC IEMOCAP DB .	ZCR, root mean square energy, pitch, harmonics-to-noise ratio, and 12 MFCCs and their deltas.	Decision tree	For AIBO, 48.37% using leave-one speaker out cross validation on the training dataset. For IEMOCAP, average unweighte recall of 58.46% using leave-one speaker out cross-validation
Luengo et al.[51]	Emotional speech database for Basque, recorded by the University of the Basque Country, with single actress	Prosodic and spectral features	SVM, GMM	98% accuracy for GMM-MFCC, 92.32% with SVM & prosodic features, 86.71% with GMM & prosodic feature
Mirsamadi et al.[119]	IEMOCAP corpus	Automatically learned by Deep RNN, as well as hand-crafted LLDs consisting of F0, voicing probability, frame energy, ZCR, and MFCC	Deep RNN	Proposed system with raw spectral features have 61.8% recognition rate Proposed system with LLD features have 63.5% recognition ra
Mao et al.[113]	SAVEE DB, Berlin EMO DB, DES DB, MES DB	Automatically learned by CNN	CNN	73.6% accuracy for SAVEE DB, 85.2% for EMODB, 79.9% for DES DB 78.3% for MES DB

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Nakatsu et al.[106]	100 utterance 50 male, 50 female	Speech power, pitch, LPC	Neural networks	50% recognition rate
Nogueiras et al.[60]	Spanish corpus of INTERFACE, Emotional Speech Synthesis Database	Prosodic and spectral features	HMM	Recognition rate higher than 70% for all emotions
Nwe wt al.[102]	3 female 3 male for Burmese language 3 female 3 male for Mandarin language	LFPC	HMM	Average recognition rates of classification for the Burmese and the Mandarin utterances are 7 % and 75.7%, respectively.
Rao et al.[55]	Telugu emotion speech corpus	Prosodic features	SVM	66% average recognition rate with sentence level prosodic featur 65.38% average recognition rate wi word level prosodic features, 63% average recognition rate with syllable level prosodic feature
Rong et al.[53]	One natural and one acted speech corpora in Mandarin	Pitch, energy, ZCR, and spectral features	kNN	66.24% average recognition rate with all 84 features, 61.18% with PCA/MDS, 60.40% with ISOMAP, and 69.21% with proposed ERFTrees method

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Sato et al.[66]	Database from Linguistic Data Consortium	MFCC	HMM	66.4% recognition rate
Schuller et al.[62]	German and English 5 speaker 5250 sample. Acted and natural data	Energy and Pitch based features	continuous HMM	86.8% average recognition rate with global prosodic features and 77.8% average recognition rate for instantaneous features
Schuller et al.[103]	3,947 movie and automotive interaction dialog-turns database consisting of 35 speakers.	Pitch, energy, and duration related features	StackingC SVM NB C4.5 kNN	63.51% recognition rate for 276 dimensional features and 71.62% for 100 dimensional features
Schuller et al.[52]	Berlin EmoDb	The raw contours of ZCR, pitch, first seven formants, energy, spectral development, and HNR and linguistic features	StackingC MLR NB 1NN SVM C4.5	76.23% recognition rate with all 276 features, 80.53% with top 75 features selected by SVM SFFS
Schuller et al.[54]	VAM DB	Low level descriptors such as signal contour, spectral pitch, formants, HNR, MFCCs, or energy of the signal and linguistic features	Support Vector Regression	Best result for Valence dimension 6.7% using linguistic features. For Activation dimension 85.1% recognition rate with acoustic and bag of n-grams features. For Dominance acoustic and bag of character n-grams features recognition rate is 82.5%

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Shen et al.[104]	Berlin Emo DB	Energy, pitch, LPCC, MFCC, LPCMCC	SVM	Best results with energy and pitch features is 66.02%, 70.7% for only LPCMCC features,and 82.5% for using both of them.
Trigeorgis et al.[114]	RECOLA DB.	Automatically learned by deep CNN	Deep CNN with LSTM	MSE in arousal dimension is .684, and MSE in valence domain is .261
Ververidis et al.[101]	1300 utterances from DES	Statistical properties of formants, pitch, and energy contours of the speech signal	GMM	48.5% recognition rate for GMM with one Gaussian density, 56% fo males and 50.9% for females
Wang et al.[158]	Berlin EMO DB, CASIA DB, Chinese elderly emotion database (EESDB)	Fourier Parameters, MFCC	SVM	For EMO DB 88.88% recognition r , while for CASIA DB 79% recognit rate, and for EESDB 76% recognit rate
Wollmer et al.[159]	Sensitive Artificial Listener (SAL) database	Low Level audio features such as pitch, MFCC, energy, HNR and also linguistic features	BLSTM, LSTM, SVM, and conventional RNN	Quadrant prediction F1-measure o up to 51.3 %,

Journal Pre-proof

Table 4: List of studies

Paper	Dataset	Features	Classifier	Results
Wu et al.[97]	Two corpora; corpora A and B consist of the utterances from six and two volunteers, total 2,033 sentences	Pitch, intensity, formants 1-4 and formant bandwidths 1-4, four types of jitter-related features, six types of shimmer-related features, three types of harmonicity-related features, MFCCs	Meta Decision Tree (MDT) containing SVM, GMM, MLP classifiers	80% recognition rate with mixed utterances from corpora A and B.
Wu et al. [160]	Berlin Emo DB, VAM DB	Prosodic features, speaking rate features, ZCR and TEO based features	SVM	91.3% by proposed modulation spectral features and prosodic features for EMODB, 86% by prosodic and spectral modulation features for VAM DB
Yang et al.[161]	Berlin Emo DB	Prosodic, spectral and voice quality features	Bayesian classifier	73.5% average recognition rate
Zhang et al.[162]	ABC, AVIC, DES, eNTERFACE, SAL, and VAM.	LLDs such as energy, pitch, voice quality, spectral, MFCC features	Unsupervised Learning	Mean unweighted recognition rate is 66.8% using Z-normalization of arousal classification, 58.2% for valence classification with centering normalization on cross-corpus emotion recognition

Journal Pre-proof

6. Challenges

Although there are many advancements on speech emotion recognition systems, there are still several obstacles that need to be removed for successful recognition.

One of the most important problems is the generation of the dataset that is used for the learning process. Most of the data sets used for SER are acted or elicited that are recorded in special silent rooms. However, the real-life data is noisy and has far more different characteristics than the others. Although natural data sets are also available, they are fewer in numbers. There are legal and ethical problems to record and use natural emotions. Most of the utterances in natural data sets are taken from talk-shows, call-center recordings, and similar cases where the involved parties are informed of the recording. These data sets do not contain all emotions and may not reflect the emotions that are felt. In addition, there are problems during the labeling of the utterances. There are human annotators labeling the speech data after the utterances are recorded. The actual emotion felt by the speaker and emotions perceived by human annotators may show differences. Even the recognition rates of human annotator are not over 90%. In favor of humans, however, we believe that we also depend on the content and the context of the speech as we are evaluating.

There are also cultural and language effects on SER. There are several studies available working on cross-language SER. However, the results show that current systems and features used are not sufficient for it. The intonation of emotions on speech among various languages may show differences for example.

An overlooked challenge is the case of multiple speech signals, where the SER system has to decide which signal to focus on. Although it can be handled via a speech separation algorithm in the preprocessing stage, current systems fail to notice this problem.

7. Conclusion

We have identified and detailed the parts that make up a speech emotion recognition system. These systems require training data provided by speech databases that are created using either acted, elicited, or natural sources. The signals are then preprocessed to make them fit for feature extraction. SER systems most commonly use prosodic and spectral features since they support a wider range of emotion and yield better results. The results can further be improved by adding features from other modalities, such as the ones that depend on visual or linguistic features.

Once all the features are extracted, SER systems have a wide range of classification algorithms to choose from. While most use classical approaches, there are an increasing number of studies that incorporate recent advances, such as Convolutional or Recurrent Neural Networks.

All of these preprocessing and feature extraction are done to detect the emotion in the speech signal, yet emotions are still an open problem in psychology. There are several models that define them. SER systems use manual labeling for their training data, which, as mentioned earlier, is not always exactly correct.

Although there are systems and realizations of real-time emotion recognition, SER systems are not yet part of our every day life, unlike speech recognition systems that are now easily accessible even with

mobile devices. To reach this goal, SER systems need more powerful hardware so that processing can be done faster; more correctly labeled data so that the training is more accurate; and more powerful algorithms so that the recognition rates increase. We believe that the research will continue towards solutions that apply deep learning algorithms, and since they require more data and more powerful processors, and these advances are likely to follow.

We believe that, as SER systems become more part of our daily lives, there will be more data available to learn from, which will improve their performance, even when at times humans can fail. The subtle differences which may not be registered by humans can be picked up by these networks that will improve the areas where emotion recognition is applicable, such as human computer interaction, healthcare, and alike.

References

[1] Björn W. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM*, 61(5):90–99, April 2018.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, Jan 2001.

[3] X. Huahu, G. Jue, and Y. Jian. Application of speech emotion recognition in intelligent household robot. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 537–541, Oct 2010.

[4] Won-Joong Yoon, Youn-Ho Cho, and Kyu-Sik Park. A study of speech emotion recognition and its application to mobile services. In Jadwiga Indulska, Jianhua Ma, Laurence T. Yang, Theo Ungerer, and Jiannong Cao, editors, *Ubiquitous Intelligence and Computing*, pages 758–766, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[5] Princess Florianne O. Boco, Diana Karen B. Tercias, Khristina G. Judan Cruz, Carlo R. Raquel, Rowena Cristina L. Guevara, and Prospero C. Naval. Emsys : An emotion monitoring system for call center agents. 2010.

[6] Mariusz Szwoch and Wioleta Szwoch. Emotion recognition for affect aware video games. In Ryszard S. Choraś, editor, *Image Processing & Communications Challenges 6*, pages 227–236, Cham, 2015. Springer International Publishing.

[7] Diana Van Lancker, Cathleen Cornelius, and Jody Kreiman. Recognition of emotionalprosodic meanings in speech by autistic, schizophrenic, and normal children. *Developmental Neuropsychology*, 5(2-3):207–226, 1989.

[8] L. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, March 2011.

[9] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162 – 1181, 2006.

[10] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.

[11] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117, 2012.

[12] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, Feb 2015.

[13] S. Ramakrishnan. Recognition of emotion from speech: A review. In S. Ramakrishnan, editor, *Speech Enhancement, Modeling and Recognition- Algorithms and Applications*, chapter 7. IntechOpen, Rijeka, 2012.

[14] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin. A review on emotion recognition using speech. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 109–114, March 2017.

[15] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28, Apr 2018.

[16] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[17] Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

[18] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.

[19] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[20] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

[21] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[22] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[23] Michael Grimm, Kristian Kroschel, Emily Mower Provost, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49:787–800, 10 2007.

[24] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[25] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *INTERSPEECH*, pages 1517–1520. ISCA, 2005.

[26] Jianhua Tao Fangzhou Liu Meng Zhang and Huibin Jia. Design of speech corpus for mandarin text to speech. In *The Blizzard Challenge 2008 workshop*, 2008.

[27] Iemocap database. `https://sail.usc.edu/iemocap/`. Accessed: 2019-05-15.

[28] Surrey audio-visual expressed emotion database. `https://sail.usc.edu/iemocap/`. Accessed: 2019-05-15.

[29] Toronto emotional speech database. `https://tspace.library.utoronto.ca/handle/1807/24487`. Accessed: 2019-05-15.

[30] Xia Mao, Lijiang Chen, and Liqin Fu. Multi-level speech emotion recognition based on hmm and ann. In *2009 WRI World congress on computer science and information engineering*, volume 7, pages 225–229. IEEE, 2009.

[31] Aijun Li, Fang Zheng, William Byrne, Pascale Fung, Terri Kamm, Yi Liu, Zhanjiang Song, Umar Ruhi, Veera Venkataramani, and Xiaoxia Chen. Cass: A phonetically transcribed corpus of mandarin spontaneous speech. In *Sixth International Conference on Spoken Language Processing*, 2000.

[32] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924, 2017.

[33] Inger S Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[34] KX Wang, QL Zhang, and SY Liao. A database of elderly emotional speech. In *Proc. Int. Symp. Signal Process. Biomed. Eng Informat.*, pages 549–553, 2014.

[35] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan. An articulatory study of emotional speech production. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[36] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3501–3504. European Language Resources Association (ELRA), 2014.

[37] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[38] Keio university japanese emotional speech database (keio-esd). `http://research.nii.ac.jp/src/en/Keio-ESD.html`. Accessed: 2019-05-15.

[39] Emotional prosody speech and transcripts. `http://olac.ldc.upenn.edu/item/oai:www.ldc.upenn.edu:LDC2002S28`. Accessed: 2019-05-15.

[40] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[41] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011.

[42] John HL Hansen and Sahar E Bou-Ghazale. Getting started with susas: A speech under simulated and actual stress database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[43] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pages 865–868. IEEE, 2008.

[44] A Batliner, S Steidl, and E Nöth. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In *Proc. of a Satellite Workshop of LREC*, volume 2008, page 28, 2008.

[45] Björn Schuller, Ronald Müller, Florian Eyben, Jürgen Gast, Benedikt Hörnler, Martin Wöllmer, Gerhard Rigoll, Anja Höthker, and Hitoshi Konosu. Being bored? recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, 27(12):1760–1774, 2009.

[46] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[47] Caglar Oflazoglu and Serdar Yildirim. Recognizing emotion from turkish speech using acoustic features. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):26, 2013.

[48] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2017.

[49] R.G. Bachu, S. Kopparthi, B. Adapa, and B.D. Barkana. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In Khaled Elleithy, editor, *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282, Dordrecht, 2010. Springer Netherlands.

[50] Jouni Pohjalainen, Fabien Fabien Ringeval, Zixing Zhang, and Björn Schuller. Spectral and cepstral audio noise reduction techniques in speech emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 670–674. ACM, 2016.

[51] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez. Automatic emotion recognition using prosodic parameters. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[52] Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[53] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. Acoustic feature selection for automatic emotion recognition from speech. *Information processing & management*, 45(3):315–328, 2009.

[54] Bjorn Schuller. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on affective computing*, 2(4):192–205, 2011.

[55] K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2):143–160, 2013.

[56] Jen-Chun Lin, Chung-Hsien Wu, and Wen-Li Wei. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 14(1):142–156, 2012.

[57] Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412, 1985.

[58] Jo-Anne Bachorowski. Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57, 1999.

[59] Roddy Cowie and Ellen Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Fourth International Conference on Spoken Language Processing*, 1996.

[60] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B Mariño. Speech emotion recognition using hidden markov models. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[61] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE transactions on audio, speech, and language processing*, 17(4):582–596, 2009.

[62] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 2, pages II–1. IEEE, 2003.

[63] Swarna Kuchibhotla, HD Vankayalapati, RS Vaddi, and Koteswara Rao Anne. A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, 17(4):401–408, 2014.

[64] Eddie Wong and Sridha Sridharan. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pages 95–98. IEEE, 2001.

[65] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Detection of stress and emotion in speech using traditional and fft based log energy features. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1619–1623. IEEE, 2003.

[66] Nobuo Sato and Yasunari Obuchi. Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848, 2007.

[67] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625, 2010.

[68] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[69] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

[70] Marko Lugger and Bin Yang. The relevance of voice quality features in speaker independent emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–17. IEEE, 2007.

[71] Marko Lugger and Bin Yang. Psychological motivated multi-stage emotion classification exploiting voice quality features. In *Speech Recognition*. InTech, 2008.

[72] Xi Li, Jidong Tao, Michael T Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. Stress and emotion classification using jitter and shimmer features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1081. IEEE, 2007.

[73] Shiqing Zhang. Emotion recognition in chinese natural speech by combining prosody and voice quality features. In *International Symposium on Neural Networks*, pages 457–464. Springer, 2008.

[74] Martin Borchert and Antje Dusterhoft. Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 147–151. IEEE, 2005.

[75] Christer Gobl and Ailbhe Nı Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1-2):189–212, 2003.

[76] John. Laver. *The phonetic description of voice quality / John Laver*. Cambridge University Press Cambridge [Eng.] ; New York, 1980.

[77] Klaus R Scherer. Vocal affect expression: A review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[78] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[79] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2011.

[80] HM Teager and SM Teager. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech production and speech modelling*, pages 241–261. Springer, 1990.

[81] James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 381–384. IEEE, 1990.

[82] James F Kaiser. Some useful properties of teager's energy operators. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 3, pages 149–152. IEEE, 1993.

[83] Guojun Zhou, John HL Hansen, and James F Kaiser. Nonlinear feature based classification of speech under stress. *IEEE Transactions on speech and audio processing*, 9(3):201–216, 2001.

[84] Chung-Hsien Wu, Jen-Chun Lin, Wen-Li Wei, and Kuan-Chun Cheng. Emotion recognition from multi-modal information. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, Kaohsiung, Taiwan, October 29 - November 1, 2013*, pages 1–8, 2013.

[85] Zhihong Zeng, Jilin Tu, Brian Pianfetti, and Thomas S. Huang. Audio-visual affective expression recognition through multistream fused HMM. *IEEE Trans. Multimedia*, 10(4):570–577, 2008.

[86] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1136–1139. IEEE, 2006.

[87] Rosalind Wright Picard et al. *Affective computing*. MIT Press, 1995.

[88] Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 366–371. IEEE, 1998.

[89] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[90] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[91] Zhihong Zeng, Jilin Tu, Brian M Pianfetti, and Thomas S Huang. Audio–visual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10(4):570–577, 2008.

[92] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Two-level hierarchical alignment for semi-coupled hmm-based audiovisual emotion recognition with temporal course. *IEEE Transactions on Multimedia*, 15(8):1880–1895, 2013.

[93] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[94] Jonghwa Kim and Elisabeth André. Emotion recognition using physiological and speech signal in short-term observation. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 53–64. Springer, 2006.

[95] Jonghwa Kim. Bimodal emotion recognition using speech and physiological changes. In *Robust speech recognition and understanding*. InTech, 2007.

[96] Florian Eyben, Martin Wöllmer, Alex Graves, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3(1-2):7–19, 2010.

[97] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011.

[98] George A Tsihrintzis, Maria Virvou, Efthymios Alepis, and Ioanna-Ourania Stathopoulou. Towards improving visual-facial emotion recognition through use of complementary keyboard-stroke pattern information. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, pages 32–37. IEEE, 2008.

[99] Preeti Khanna and M Sasikumar. Recognising emotions from keyboard stroke pattern. *International journal of computer applications*, 11(9):1–5, 2010.

[100] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[101] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *2005 IEEE International Conference on Multimedia and Expo(ICME)*, volume 00, pages 1500–1503, 07 2005.

[102] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[103] Björn Schuller, Stephan Reiter, Ronald Muller, Marc Al-Hames, Manfred Lang, and Gerhard Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 864–867. IEEE, 2005.

[104] Peipei Shen, Zhou Changjun, and Xiong Chen. Automatic speech emotion recognition using support vector machine. In *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*, volume 2, pages 621–625. IEEE, 2011.

[105] Hao Hu, Ming-Xing Xu, and Wei Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–413. IEEE, 2007.

[106] Ryohei Nakatsu, Joy Nicholson, and Naoko Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 343–351. ACM, 1999.

[107] Valery Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering*, volume 710, page 22, 1999.

[108] Joy Nicholson, Kazuhiko Takahashi, and Ryohei Nakatsu. Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4):290–296, 2000.

[109] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

[110] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 511–516. IEEE, 2013.

[111] Jun Deng, Zixing Zhang, Florian Eyben, and Björn Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

[112] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[113] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, 16(8):2203–2213, 2014.

[114] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.

[115] Jaebok Kim, Khiet P Truong, Gwenn Englebienne, and Vanessa Evers. Learning spectro-temporal features with 3d cnns for speech emotion recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 383–388. IEEE, 2017.

[116] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[117] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

[118] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, pages 597–600, 2008.

[119] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[120] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[121] Enrique M Albornoz, Diego H Milone, and Hugo L Rufiner. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, 25(3):556–570, 2011.

[122] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *2005 international conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE, 2005.

[123] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. In *Ninth International Conference on Spoken Language Processing*, 2006.

[124] Yixiong Pan, Peipei Shen, and Liping Shen. Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108, 2012.

[125] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[126] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech communication*, 54(9):1049–1063, 2012.

[127] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools and Applications*, 46(1):119, 2010.

[128] Leimin Tian, Johanna Moore, and Catherine Lai. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 565–572. IEEE, 2016.

[129] Seliz Gülsen Karadoğan and Jan Larsen. Combining semantic and acoustic features for valence and arousal recognition in speech. In *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, pages 1–6. IEEE, 2012.

[130] Meysam Asgari, Géza Kiss, Jan Van Santen, Izhak Shafran, and Xubo Song. Automatic measurement of affective valence and arousal in speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 965–969. IEEE, 2014.

[131] Heysem Kaya, Dmitrii Fedotov, Ali Yesilkanat, Oxana Verkholyak, Yang Zhang, and Alexey Karpov. Lstm based cross-corpus and cross-task acoustic emotion recognition. In *Interspeech*, pages 521–525, 2018.

[132] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. The usc creativeit database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, page 55, 2010.

[133] Sefik Emre Eskimez, Zhiyao Duan, and Wendi Heinzelman. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103. IEEE, 2018.

[134] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. *arXiv preprint arXiv:1712.08708*, 2017.

[135] Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Wael AbdAlmageed, and Carol Espy-Wilson. Adversarial auto-encoders for speech based emotion recognition. *arXiv preprint arXiv:1806.02146*, 2018.

[136] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):31–43, 2017.

[137] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[138] Jaebok Kim, Gwenn Englebienne, Khiet P Truong, and Vanessa Evers. Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning. *arXiv preprint arXiv:1708.03920*, 2017.

[139] Karttikeya Mangalam and Tanaya Guha. Learning spontaneity to improve emotion recognition in speech. *arXiv preprint arXiv:1712.04753*, 2017.

[140] Srinivas Parthasarathy and Carlos Busso. Jointly predicting arousal, valence and dominance with multi-task learning. In *INTERSPEECH*, pages 1103–1107, 2017.

[141] Reza Lotfian and Carlos Busso. Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning. *Proc. Interspeech 2018*, pages 951–955, 2018.

[142] Duc Le, Zakaria Aldeneh, and Emily Mower Provost. Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. In *INTERSPEECH*, pages 1108–1112, 2017.

[143] Che-Wei Huang and Shrikanth Shri Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 583–588. IEEE, 2017.

[144] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018.

[145] Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai. An attention pooling based representation learning method for speech emotion recognition. *Proc. Interspeech 2018*, pages 3087–3091, 2018.

[146] Michael Neumann et al. Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773. IEEE, 2018.

[147] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[148] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. Progressive neural networks for transfer learning in emotion recognition. *arXiv preprint arXiv:1706.03256*, 2017.

[149] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

[150] Melissa N Stolar, Margaret Lech, Robert S Bolia, and Michael Skinner. Real time speech emotion recognition using rgb image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8. IEEE, 2017.

[151] Siddique Latif, Rajib Rana, Shahzad Younis, Junaid Qadir, and Julien Epps. Transfer learning for improving speech emotion classification accuracy. *arXiv preprint arXiv:1801.06353*, 2018.

[152] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[153] Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, and Carol Espy-Wilson. Smoothing model predictions using adversarial training procedures for speech based emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4934–4938. IEEE, 2018.

[154] Mohammed Abdelwahab and Carlos Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, 2018.

[155] Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, and Björn Schuller. Towards conditional adversarial training for predicting emotions from speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6822–6826. IEEE, 2018.

[156] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5005–5009. IEEE, 2017.

[157] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Reconstruction-error-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2367–2371. IEEE, 2017.

[158] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li. Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, 6(1):69–75, 2015.

[159] Martin Wollmer, Björn Schuller, Florian Eyben, and Gerhard Rigoll. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):867–881, 2010.

[160] Siqing Wu, Tiago H Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[161] Bin Yang and Marko Lugger. Emotion recognition from speech signals using new harmony features. *Signal processing*, 90(5):1415–1423, 2010.

[162] Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller. Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 523–528. IEEE, 2011.