SPEECH EMOTION RECOGNITION SURVEY

Husam Ali Abdulmuhsein1, Hala Bahjat Abdul wahab2
Abdul Mohssen Jaber Abdul hossen3

1Assistant professor, Computer Science Department, Faculty of Science, University of Baghdad, Baghdad, Iraq.

2Professor, Faculty of Computer Science, Technology University, Baghdad, Iraq.

3Professor, Computer Science Department, Al-Turath University College, Baghdad, Iraq.

Email: 1abdulmohsein.h.a@scbaghdad.edu.iq, 2110120@uotechnology.edu.iq, 3abdulmoohsen53@yahoo.com

Corresponding Author: Husam Ali Abdulmuhsein

https://doi.org/10.26782/jmcms.2020.09.00016

Abstract

Speech emotion recognition (SER) research field extends back to 1996, but still one main obstacle still exists, which is achieving real-time SER systems. The once-imaginary relationship between humans and robots is rapidly approaching reality. Robots already play major roles, particularly in manufacturing, but until recently, they did only what they were programmed to do. However, with the development of artificial intelligence (AI) approaches, SER researchers are seeking to move robotics to a higher level, giving them the ability to predict human actions and recognize facial expressions and allowing them to interact with humans in more natural and clever ways. Humans are complicated; understanding only what they say is insufficient for all situations. One complication is that humans express identical emotions in multiple ways. For robots to act more like humans, understand them, and follow their orders in more intelligent ways, they need to understand emotions to make appropriate decisions. Thus, to reach the ideal SER state, a more up-to-date survey that considers how SER research has evolved over the past decade is needed. In this survey, our main goal is to explain the different research approaches followed in the SER field particularly Path 6, which represents a new technique in the SER field. To clarify

Husam Ali Abdulmuhsein et al
the techniques for readers, details of the SER systems and their different approaches will be elaborated.

**Keywords**: feature extraction, feature selection and classification, real-time system, robotics, SER.

---

I. Introduction

SER is a 24-year-old research field that goes back to the first article published in 1996 [XI], that aims to achieve an ideal system that can categorized emotions in real time. Thus far, this goal has not been accomplished, and there is much work still to be done. Through our survey, it was noticed that the traditional SER block diagram shown in Fig. 1 has been followed for a long time, but none of the results have fully achieved the expectations. In 2011, [XXXIII], a huge revolution occurred in the world of SER research that resulted in a new approach for recognizing emotions from speech. The main idea behind this invention is that each emotion has its own representation in speech waveforms; this behavior is speaker independent; and discovering that behavior will lead us to a standard identification to each emotion. This research direction will be termed "Path 6" throughout this survey. Yet, few studies have been published that follow Path 6; nevertheless, this path is promising, and the studies that have been conducted have obtained good results.

Based on this survey, a more detailed block diagram will be used for exploring the functionalities of recognition systems and the different approaches used up to the end date of this survey. The standard block diagram, first created in 1994, is constructed from four phases [XVII]: acoustic analysis, statistical pattern matching, pattern selection and pattern recognition (classification), as shown in Fig.1.

Because our research concerns speech, it was considered that the input to the recognition system to be a signal only.

Our SER block diagram is shown in Fig. 2, and this diagram will be named “SERBD-P6”. The name refers to “Speech Emotion Recognition Block Diagram-Path Six”.

---

*Husam Ali Abdulmohsin et al*
Fig 1. Four Principal functions of an automatic speech recognition system.

Through this section, a fast review of all the paths in the SERBD-P6 will be illustrated. Path 1, shown in Fig. 2, reflects the theoretical concept that signals can be classified directly; however, due to variability and complexity issues [XVII], dealing with large amounts of information is a time-intensive process, which is why feature extraction was invented.

Fig 2. The SERBD-P6 Structure.

Path 2 addresses the same theoretical concept but after preprocessing has been conducted. Paths 3 and 4 show that both preprocessing and feature selection (FS) are theoretically optional in a recognition system; however, both play critical roles in system performance.
The remainder of this paper is organized as follows. First, reviewing some of the main concepts in emotion recognition; then, provide a clear definition for the term “emotion,” which has been one of the challenges in this field of research. Next, discuss some of the challenges that researchers in the SER field face. The next section explores few of the used datasets, features, and classification methods used in SER. Paths 4–7 are covered separately in our research through a survey of most of the studies published from 2014 until the present in the SER field. Finally, conclusions will be discussed at the last section.

II. Emotional speech concepts

One concept that must be highlighted involves the two main approaches used in SER, the multimodal and unimodal emotion recognition systems. Researchers in this field state that working with only one model, such as text, speech, or image, to recognize emotions is considerably more challenging than working with many combined models to recognize emotions. The reason is logical: dealing with many facts to reach a decision is much easier than dealing with fewer facts. Although the highest accuracy results have been achieved in SER using multimodal systems, SER must sometimes be implemented using unimodal systems due to the absence of multiple information resources. In 2018, [XX] reported that emotion recognition based on a combination of facial expressions and acoustic information reveals performs better than does using only one information source.

Another basic concept is understanding how emotions can be revealed. There are two ways in which human beings can diagnose emotions: first, by speech, through acoustic, syntactic, or semantic information, and second, by behavior, which may involve facial or body gestures. However, machines enable a new approach for diagnosing emotions—using information obtained by measuring body signals such as heartbeat. This line of research is not covered in this survey; only emotion recognition that depends on speech was considered [L].

Another concept is the difference between local and global features. When a speech signal is divided into smaller units (called frames or segments) and the features are computed from each frame separately from others, those features are called local features. In contrast, features extracted from the entire speech signal are called global features. It has not yet been proven which type (local or global) of features are most suitable for SER.

Another concept involves emotion recognition affected by gender. In 2017, [XXVIII] obtained different accuracy results for male and female samples, and [XXIII] stated that emotion classification conducted on a specific gender leads to higher performance.

In 2017, [XLI] proved an old concept in SER: each emotion is related to specific types of features; thus, every group of features or individual feature is more effective for classifying certain emotions than others. They noticed that after
combining their novel proposed features to the traditional acoustic known features, their recognition accuracy results increased for surprise, joy and anger, less so for fear and sadness. The smallest increase in recognition accuracy was found for the emotion of disgust. In 2018, [XXXIX] showed that the acoustic features generated by human vocal cords are extensively related to emotion recognition, especially for fear. In 2018, [XXVII] proposed an intersegmental emotion recognition system and showed that vowel segments alone can be used to classify emotions when optimal vocal tract and glottal source features exist. In 2019, [XII] classified three emotions using the SVM classifier and achieved 80% accuracy; however, this study also reported that anger and happiness are closely related to the Short Term Energy (STE) features. Their classification accuracy increased by 15%-20% when STE features were added to the feature vector, but the classification accuracy of sadness did not improve.

In the last few years, a new concept called language-agnostic SER was proposed, in which speaking language, gender and speaker are independent. Language-agnostic SER is different from cross-language SER systems. Since SER came into the world of research, most of the papers published have been concerned with cross-language SER systems because of the extensive applicability of such systems in the real world. However, in cross-language systems, gender and speaker are crucial because the absence of such information may reduce the accuracy of the SER system. In 2020, [XLIV] produced a language-agnostic SER system in which speaking language, gender and speaker are independent. They used Gaussian mixture models for the combined temporal modulation spectra and the Mel-scaled representation. Emotion prediction was implemented by the Kullback–Leibler divergence. The obtained accuracy reached 70.1%, classifying six emotions from the Italian language dataset (EMOVO) and the Berlin dataset.

III. Emotion definition

According to Merriam Webster, the definition of emotion is “a conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body” [XXXI]. Philosophers have agreed that having an object is essential for generating emotion. The object can be an attitude toward other persons (e.g., a funeral), any object (e.g., a house), an organism (e.g., a dog), a natural phenomenon (e.g., rain), one’s own behavior (acting positive), or a combination of any of the above. However, objects do not include back pain, pleasant emotions, morning feelings of depression or afternoon feelings from combinations of emotions; none of those have objects; they do not concern nor are they related to specific objects. In conclusion, emotions are intentionally similar to many other psychological states; they provide meaning to objects beyond that of the objects themselves. Psychologists have generally taken the directedness of emotions for granted, partly because most emotion theorists have adopted Darwin’s view that

Husam Ali Abdalmohsin et al
functionally, emotions prepare adaptive responses to challenges—something that obviously requires directedness. Finally, two facts were concluded. First, emotions have objects such as a person’s qualities, dogs, things, properties, processes, causes, situations and events. Second, emotions involve feelings (which tend to reflect bodily reactions). In natural languages, emotions are typically individuated by references to their objects. So it can assumed that a component necessary for an emotion involves an episode of hearing something, seeing something, seeing that something is the case, remembering something, judging or remembering that something is the case, seeming to see that something is the case, and so on. These intellectual episodes, perceptual and cognitive, are sometimes called the emotion basis. Without one of these bases, no emotion can occur—but the bases can occur without the emotion [XXI].

IV. Challenges

SER systems have many challenges:

I. Speech signals can be originate from both static objects, such as a person standing in a certain spot or sitting somewhere, and from moving objects, such as a person walking away from a robot, producing signals from varying distances which causes a change in the vibrations.

II. Emotional states are difficult to identify because different persons have many ways of expressing the same emotional state.

III. Sometimes it is difficult to tell the difference between different emotions even of the same person because some emotions are not easy to differentiate [XXXVIII].

IV. Speech signals can be affected by human health conditions, which subsequently can affect the judgment of the SER system.

V. Choosing the dataset to apply the experiment is another problem in SER. The same dataset can yield different accuracy results using different classification methods. In 2018, [XXV] showed that classifying the features (MFCC and Modulation Spectral (MS)) extracted from two datasets, the Berlin and Spanish datasets, when using two different classifiers (a recurrent neural network (RNN) and MLR) on both datasets, yields different accuracy results. The experiment showed that when applying the RNN to both datasets, RNN/Spanish gives the highest accuracy (90.05%), and when applying the MLR to both datasets, MLR/Berlin yields the highest accuracy (82.41%). From this experiment it can concluded that a relationship exists between the features, dataset and the classification methods. Some classification methods work with some datasets better than other classification methods do.

VI. Through the extensive survey of the SER field, it was noticed that most of
the SER systems that deployed an NN for classification or for feature unification or reduction achieved high accuracy. However, that is not a standard conclusion. One of the challenges of the SER field is the system design—choosing the dataset, choosing the training and testing rates, choosing the feature set and the classification method [XXXIV], utilized an NN for classification but achieved only low accuracy (35.95%) and attributed that to the learning method used and a lack of information in the dataset. However, but we attribute it to poor system design.

V. Emotional datasets

This section will cover few of the frequently used emotional datasets available online and their drawbacks. Most researchers in the speech recognition field know that some percentage of all the available datasets have human performance accuracy problems that occurred during dataset collection. This drawback is caused by humans actors attempting to mimic real emotions in a pretend fashion, regardless of the talent of the professionals employed for such difficult jobs. The instruments and devices used during the recording process can also play a major role that results in serious quality defects in the recorded audio files. The material used to record the datasets is another factor that plays a major role in dataset quality; some sentences are better suited for emotional interpretation than are others, and some sentences reflect emotion more strongly than others. All the above facts will affect the quality of the audio dataset. Through this section, examples of degradation due to human performance, audio recording accuracy and validity for three widely used datasets available online (see Table 1) will be shown. These same problems also apply to other datasets; each dataset has multiple quality measurements that the researcher needs to study before using the dataset.

### Table 1: Some of the Datasets Used in SER and Its Emotional Representation Accuracy

<table>
<thead>
<tr>
<th>Resource</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[VIII]</td>
<td>Berlin</td>
<td>84% for human performance</td>
</tr>
<tr>
<td>[XIX]</td>
<td>RekEmozio</td>
<td>66.5% for audio recording</td>
</tr>
<tr>
<td>[XLVI]</td>
<td>REVDESS</td>
<td>72% Validity task mean proportion correct scores</td>
</tr>
</tbody>
</table>

Some standard criteria are available for evaluating SER research. The main standard criterion is to analyze the dataset used in the research. Studies can achieve high accuracy with single-gender datasets, and the opposite is true for datasets that represent both genders or represent a small number of emotions. Therefore, when choosing a reliable dataset, researchers should consider these key points:

I. The number of actors employed in recording the dataset: larger acting groups result in greater sample variety.

*Husam Ali Abdulmohsin et al*
II. The number of emotions represented in the dataset: more emotions are better for most research purposes. For research that concentrates on a specific emotion, many samples are required to reach the needed level of variety.

III. Gender balance is important in the dataset because male and female frequencies are different; a dataset that includes only one gender will cause the loss of information necessary in the SER research field. SAVEE is an example of a single-gender dataset.

IV. The performance accuracy rate of the dataset, which is measured by many factors mentioned in the beginning of this section, is related to the quality with which the actors mimicked real emotions and the laboratory equipment, textual material, and sentences used when the dataset was compiled.

V. Some emotions are closely related to others; consequently, it is extremely useful to use a dataset that includes similar emotions to test a proposed architecture.

VI. If the accent used in the dataset recordings is limited to a specific accent (for the English language for example), then the results will lose its generality for that language.

VI. Features classes

To cover the bulk of possible features in this survey, it was needed to perform a full classification of all types of features. Many types of features are used for classification. One approach is to classify features as local or global features. Global features are extracted from the complete signal, while local features are extracted from signals segments. Another classification approach is to divide the features into speaker-dependent and speaker-dependent features. It is known that speaker-dependent features are richer in emotional information than are speaker-dependent features. The recognition rate final contribution of speaker-dependent features are higher, but the recognition effect of speaker-dependent features is lower under conditions where the speaker changes [LVI], [XXX].

Another type of feature classification is based on their origin—where they were extracted from speech. These types are termed acoustic and linguistic features.

I. Acoustic Features,
   a. Prosody features include pitch, loudness and intensity, which describe the amplitude and frequency of the speech signal.
   b. Energy features describe loudness perception.
   c. Voiced and unvoiced probabilities.
   d. Spectral features based on the characteristics of the human ear. These features describe the speech formants, which model spoken content and represent speaker characteristics. One new type of spectral feature, invented in 2018 [V], consists of cyclo-stationary spectral features. This spectral analysis reveals the first- and second-order (hidden) periodicities in the

*Husam Ali Abdulmohsin et al*
emotional speech signal using the estimated spectral correlation function (SCF).

e. Cepstral features represent the changes or periodicity in the spectrum features measured in frequencies. These features are modeled using 12 MFCCs calculated based on a Fourier transform of a speech segment [LI].

II. Linguistic features include extracted words and semantics. These types of features are related to grammatical and lexical analyses of the utterance. The latest feature classification was introduced by [XXIX] and shown in Fig. 3.

![Fig 3. The latest feature classification published.](image)

VII. Classification methods

Gaussian mixture models, hidden Markov models, support vector machines and neural networks are the most well-known classification algorithms utilized in speech emotion recognition processes [XXII]. In SER, classification can be divided into two types: uni-classifiers and multi-classifiers.

Uni-Classifier SER systems

Uni-classifier SER recognition systems depend on single classification methods to recognize the emotions represented by the features extracted during the feature extraction phase. In other words, a uni-classifier scheme uses one classifier or
multiple classifiers in parallel, while a multimodal scheme uses many classifiers arranged serially, as shown in Fig. 4.

![Diagram of signal processing]

Fig 4. (a) represents the Uni-modal SER system, and (b) represents the Multimodal SER system, assuming the SER system is constructed from the feature extraction and classification phases only.

In 2015, [XLVII], the gammatone frequency cepstrum coefficient (GFCC) was used for feature extraction, and a backpropagation NN was used to classify two emotions, Sad and Happy, from an English speech dataset. This experiment was performed under various noise conditions, and the accuracy achieved was 97.38%. In 2017, [XLVIII] proposed a uni-classifier SER that measured only the randomness of the MFCC, energy and pitch cues over time to discriminate among the different human emotions. Entropy was used to measure the randomness of cues computed from the cooccurrence matrices and the temporal histogram. These newly proposed features achieved good accuracy compared with the state-of-art methods when applied to the SAVEE dataset. The study used an SVM classifier for classification, and the accuracy reached 82.2%. In 2017, [LV] applied learning automata to error backpropagation to train a modified brain emotional learning (BEL) model based on learning automata (BELBLA) to reduce the complexity of the original method. In BEL, two neural networks form the main structures, namely, the orbitofrontal and amygdala cortex. Formants, energy, amplitude, pitch, MFCC and zero crossing features were extracted from short-term signals from the Berlin dataset. The classification accuracy reached 77.3%, and the study showed that BELBLA outperformed many compared classifiers, including GMM, kNN, SVM, ANN and HMM models. In 2018, [XXXII] proposed an SER framework that used cepstral coefficient (CC) features and the k-means clustering method for classification. The goal was to recognize three emotions, Happy, Anger and Sad. The dataset used was

*Husam Ali Abdulmohsin et al*
recorded by seven males and eight females. The proposed framework outperformed some of the related recent works, achieving an accuracy of 92%. Most importantly, this method reduced the CPU time consumption during clustering. In 2019, [IV], MFCC and STFT features are used for SER. Different approaches were followed using deep-learning techniques such as CNN and context level analysis, and different methods used to explore the performances and accuracies of the employed approaches. Data augmentation and hyperparameter sweeps were applied to improve the performance. To include the majority of emotions in their research, the experiments were applied to the RAVDESS dataset, which represents eight emotions. Three emotions were classified with high accuracy based on their waveforms, Anger, Disgust and Calm, with accuracies of 86.8%, 78% and 72%, respectively. Natural and Calm were classified as similar emotions, and the model was unable to distinguish between them. Poor accuracy rates were obtained when classifying Calm, Neutral and Happy, (55%, 64% and 66%, respectively). The latest example of a uni-classifier SER system was proposed in 2019, by [XXVI], which conducted a comparison study between three classifiers, RNN, SVM and MLR. They extracted MFCC and Modulation Spectral (MS) features from two datasets, Berlin and Spanish. The best results (94%) were obtained by the RNN classifier on the Spanish dataset, but all three classifiers reached the same accuracy on the Berlin dataset (83%).

Multi-Classifier SER systems

Recently, many multi-classifiers have been proposed, some of these works will be discussed in this section.

Combining multiple classifiers in the same system sometimes results in better accuracy but can also sometimes yield weaker results than single-classifier systems [I]. Multi-classifier system development depends on the strategy deployed and must consider the complexity that multiple classifiers will add to the system as well as the increased time consumption. In some cases, time is crucial; thus, increasing accuracy at the expense of time is not acceptable. In this section, a list of studies that used multi-classifier systems will be discussed and report their performances and results. In 2014, [VI] used a multiple parallel classification system (k-nearest neighbor (KNN), Gaussian mixture model (GMM), backpropagation artificial neural network (ANN) and support vector machine (SVM)) to recognize seven emotions from the Berlin dataset using autoregressive (AR) parameters (gain and reflection coefficients) and linear prediction coefficients (LPC). One conclusion reached was that reflection coefficient features represent emotions better than do LPC features. A decision-level fusion technique was used to combine the results. The accuracy obtained was 77.2%. Regardless of the multi-classifier selections, Milton showed that the results could be improved by adding MFCC features to the pool, which improved the accuracy to 83.7%. This study proved that system accuracy is dependent upon both feature quality and the design of the classification.
system. Both must be appropriate to obtain highly accurate results, and both must be synchronized when accounting for [measuring] time consumption. In 2015, [XXIV] proposed an SER system that utilized six emotional models working in parallel. The models use a combination of three sets of acoustic features and two training methods. The framework was designed to recognize emotions from naturally occurring signals in a human robot interaction (HRI) environment. Variance normalization and mean were applied to the features. Two datasets were utilized to train the models, and a third realistic dataset was used to test the framework, which was designed to recognize four emotions. Each model outputs a probability, forming the input to a reliable estimation protocol that outputs the highest probability. The prediction performance reached 73%. In 2019, [II] produced a bagged assembly comparison using SVM and Gaussian kernels for speech emotion recognition. They extracted MFCCs and spectral centroid features from three datasets, EmoDB, RAVDESS and IITKGP-SEHSC, and then deployed the wrapped-based FS method. Their experiment showed that the accuracy of their system increased by 5% when using the bagged SVM multi-classifier system. The performance accuracies attained were 92.45%, 75.69%, 84.11% for the EmoDB, RAVDESS and IITKGP – SEHSC datasets, respectively.

VIII. Path 3

In very rare cases, an input signal does not require preprocessing; however, if it was needed to extract parts of the signal, reorganize a dataset for compatibility with our system, apply low- and high-pass filters, reduce noise, normalize the data, quantize the data and/or remove silence, then a preprocessing phase is essential. Notably, normalization can be applied after feature extraction. In 2017, [XV] proposed a normalization process applied to features after feature extraction, that translates as indicated by the bidirectional arrow in Fig. 2. This image shows that preprocessing can be applied to features as well as to the original signal, after which the system can move directly into the classification phase. In 2018, [VII] obtained results improved by 12.5% after preprocessing, which included aligning and scaling the original signal. In 2019, [XXVI] applied Speaker Normalization (SN) to features extracted from the Berlin and Spanish datasets and reported that SN improved the results on the Berlin dataset but reduced the results on the Spanish dataset. It is known that the SN is used to interpret variation according to speaker diversity rather than variation according to emotional state, which is why the Berlin dataset (recorded with 10 actors) was positively affected by the SN, while the Spanish dataset (recorded using two actors) was not. These results lead to the conclusion that normalization (preprocessing) is case dependent—it depends on the particular case studied by the researcher—and that using preprocessing is not always compatible with the case being studied. Each preprocessing operation must be considered to identify its operation and then a scientific judgment must be made concerning whether to use the preprocessing operation on the data.

Husam Ali Abdulmohsin et al
Feature extraction is an important stage in SER systems specifically and in recognition systems in general and it is shown in the statistical pattern matching in Fig.1. A close attention was paid to the feature extraction process and select the strongest features for classification while avoiding high feature dimensionality that unnecessarily increases system complexity.

We consider the feature extraction phase, to be the most important stage in the recognition lifecycle because strong features can distinguish among multiple classes. Using an accepted classification method, these features alone can be sufficient to achieve good results. Some researchers do not perform any feature reduction; they simply use all the extracted features in the classification phase. However, that approach can result in complex architectures.

In 2018, [XIV] reported that while no correlation exists between SER system accuracy and the number of features used, there is a strong correlation between system complexity and the number of features used. These conclusions were reached from an experiment using three sets of features extracted from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and a convolutional neural network (CNN) for classifying eight emotions. Each set of features yielded different results—regardless of how many features were in each set. In 2019, [XL] proved that there some features are strongly related to certain emotions but poorly related to others. They adopted MFCC, energy, pitch and formant frequency features extracted from the Berlin dataset and used an SVM for classification. Their results clearly showed correlations between the features and emotions. Using the same scheme to classify four emotions, they obtained the following results: 92.59% accuracy for both Happy and Anger emotions, but only 44.44% and 42.58% accuracy for Sad and Fear emotions, respectively. The large difference in accuracy between these two groups of emotions supported the conclusion, and the features mentioned in their research should be considered in future research. In 2020, [XXXV] proved that each set of features has strengths for classifying different emotions. They classified the features into three unequal-length models using duplicated features where required. Then, they applied four classification methods. The best accuracy results they obtained were 94.99% for Anger using the model 2 feature vector and an SVM classifier; 89.11% for Disgust using the model 3 feature vector and a KNN classifier; 90% for Fear using the model 2 feature vector and an NN classifier; 89.33% for Happy using the model 2 feature vector and NN and SVM classifiers; 93.02% for Neutral using the model 3 feature vector and an SVM classifier; 95.24% for Sad using the model 2 feature vector and a Random Forest (RF) classifier; and 90.23% for Surprise using the model 2 feature vector and an RF classifier. Therefore, each feature model and classifier combination is more powerful for classifying a specific emotion, regardless of the number of features that are used in each feature model.

Husam Ali Abdulmohsin et al
Feature extraction affects the accuracy of SER systems. Therefore, spending much time on feature extraction is reasonable—but not too much time. In 2014, [IX] employed a deep neural network (deep belief network, DBN), to extract emotional features. The experiment yielded good results, but the feature extraction process took 136 hours—much longer than traditional feature extraction methods. The experiment was conducted on a dataset of Indian songs downloaded from famous Indian song channels. In 2013, [VI] reported that using autoregressive (AR) features in addition to LPC in multi-classifier SER systems was insufficient. Regardless of which combination of multiple classifiers they used, they also needed to include MFCC features to improve the accuracy of their SER system. In 2017, [XXVIII] showed that deep features extracted by a DBN represent emotional status better than do artificial, statistical emotion features.

X. Path 5

Path 5 represents the SER systems that employ FS in their design to reduce complexity; this path is one of the most important steps in all systems concerned with recognition and is represented as pattern selection in Fig. 1. FS is known as feature reduction, feature optimization or dimensionality reduction of features, and all these terms have the same meaning.

The FS process aims to reduce the correlations between the features in the feature vector and consequently, to reduce the complexity of the whole system, which, in turn, decreases the runtime of the SER system. Many researchers consider that FS is primarily concerned with the system accuracy, and multiple studies have proved that concept. Few FS methods will be surveyed in this section.

In 2015, [XVI], [XVIII] reported that when they employed FS in their systems, the results improved, and the complexity of their systems decreased. The authors of [XVI] deployed a decision tree to select the best features extracted from the eNTERFACE05 audiovisual emotion dataset and achieved an accuracy of 96.59% using an SVM classifier. In 2016, [XXXVII] tested four different FS algorithms (correlation-based, genetic algorithm, forward FS and information gain) with four different classification algorithms (SVM, kNN, NN and decision tree C4.5) and found that the FS algorithms both reduced system complexity and increased system accuracy. The study also proved that the information gain algorithm performs better than do other FS algorithms with two classification algorithms (NN and SVM). It was also proven that the SER system performed better when using any FS algorithm than when FS was not used. In 2017, [LVI] proposed a new framework that utilized correlation analysis and the Fisher criterion for feature selection; applying FS improved their classification accuracy by 2.4% on average. Their framework also required 25% less time for classification after applying FS: prior to applying FS to the feature set, the system required approximately 1.2 s for classification but required only 1.6 s after applying FS. In 2018, [XXVII] applied the maximum relevance minimal redundancy backward wrapping (MRRBW)
algorithm to reduce the feature vector length from 47 to 20 features. They reported that the accuracy of their system increased as the complexity (time consumption) decreased. Some researchers employ two different types of dimensionality reduction methods, each for a different reason. In 2018, [LVII] adopted a new linear discriminant analysis (LDA) combined with principal component analysis (PCA) dimensionality reduction method to project a new subspace to the initial feature set. In this method, PCA is used to remove noise from the initial features and LDA is used to reduce the dimensionality. In 2018, [XXII] proposed an SER system that utilized MATLAB tools to extract MFCC, energy, and statistical speech features. They concluded that system performance was clearly enhanced when the FS algorithm was applied to the system. The system reached an accuracy of 72% using 45 neurons in one hidden layer of an NN. In 2019, [36] extracted spectral centroid and MFCC features from EmoDB (the Berlin dataset) to represent the speech signal, and then utilized the wrapped-based FS method to increase the performance of their system by 7%. In 2019, [XXVI] conducted a comparative study on SER systems and utilized recursive feature elimination through basic linear regression (LR-RFE) to rank the features extracted from two datasets, Berlin and Spanish. This approach achieved accuracies of 83% and 94% on the two datasets, respectively. The FS method was implemented in both experiments and helped to increase the performance of their SER system. In 2019, [XLV] optimized the feature set by utilizing singular value decomposition (SVD) after initialization with semi-nonnegative matrix factorization (Semi-NMF). The MFCC, linear prediction cepstral coefficients (LPCC) and Teager energy operator-autocorrelation (TEO-AutoCorr) were extracted from two datasets (EMO-DB and IEMOCAP). They used KNN and SVM classifiers and a 5-fold cross-validation scheme and achieved remarkable accuracy when using the Semi-NMF. Compared to prior related works, their model outperformed all the previous results, achieving accuracies of 90.12% and 89.3% using the SVM and KNN classifiers, respectively, on Emo-DB and 83.2% and 78% using the SVM and KNN classifiers, respectively, on IEMOCAP.

XI. Path 6

Path 6 discusses a different approach that was invented in 2011 by Jaitly and Hinton [XXXIII], who proposed a new way of representing a signal. Rather than using traditional features and without applying a feature extraction process, they used neural networks to learn an intermediate representation of the raw input signal automatically. The main idea underlying this path is to train a model to learn the intrinsic structures in the raw speech signal by capitalizing on the fact that a strong relationship exists in the speech signal sequence. Some of the approaches applied to this path do not use traditional classification evaluation algorithms; instead, they use cross-validation algorithms to calculate the probability of the results. In 2016, [LIII] proposed a network based on RNNs, CNNs, and time distribution CNNs with no traditional features to classify emotions. Comparing the performance of the proposed time distributed CNN framework with that of CNNs and a long short-term memory
(LSTM) network on the Berlin dataset with seven emotions, they reported that the accuracy of the proposed network was better; it achieved an accuracy of 88.01%. All the results of this experiment were validated via cross-validation. They applied majority voting to reach the final decision. In 2016, [XIII] proposed a novel time-continuous end-to-end SER prediction system that classified two natural and spontaneous emotions (Valence and Arousal) using the RECOLA dataset. They found that the proposed method outperformed all the traditional SER approaches based on signal processing and state-of-the-art features. A combination of convolutional CNNs and LSTM networks was used to learn a perfect representation of the speech signal directly from the raw data. In 2020, [XLII] designed an architecture that involved preprocessing, transformation, and a CNN to learn emotional features and classify them accordingly. This is the main idea underlying Path 6. The feature extraction process and the classification process are merged, forming a single process. An independent feature extraction process is no longer required. Therefore, they proposed a CNN architecture to learn emotional features extracted from a spectrogram representation of the speech signals in the RAVDESS and IEMOCAP datasets. They first applied preprocessing to each signal using a dynamic threshold technique to remove noise and silent segments. The preprocessing output was transformed into spectrograms to reduce the complexity and increase the accuracy of the SER system. They achieved accuracies of 79.5% on the RAVDESS dataset and 81.75% on the IEMOCAP dataset.

XII. Path 7

SER systems that utilize neural networks to convert features from low-level to high-level representations are considered as Path 7 in our survey. Throughout the history of SER research, researchers have used standardized emotion recognition features to recognize emotions without any manipulations of those standard features. In 2014, [X] utilized prosodic and spectral features in a newly proposed architecture. Six emotions extracted from two emotional datasets, eNTERFACE’05 and RML, were used in this experiment. The main idea behind this architecture was to design two main subpaths that work in parallel. Subpath 1 analyzes the prosodic features, and subpath 2 analyzes the spectral features (MFCC). Subpath 2 takes MFCC as the input to two parallel algorithms (bidirectional principal component analysis (BDPCA) and linear discriminant analysis (LDA)), which in turn form the input to the radial basis function (RBF) neural classification. The output of subpaths 1 and 2 form the input to a fusion module used for decision making. The accuracy scores achieved by this architecture according to the datasets are 75.89% and 68.57%, respectively. In 2015, [XLIX] proposed an SER system utilizing DCNN and log Mel-spectrograms (delta, delta-delta and static) as the inputs. An AlexNet DCNN pretrained on ImageNet was used to learn high-level features from each frame segmented from the speech signal, and the learned features are combined by the discriminant temporal pyramid matching (DTPM) strategy. DTPM is intended to generate the utterance level feature representation. The proposed SER used an SVM...
to classifying emotions in four datasets: RML, EMO-DB, BAUM and eNTERFACE05. The resulting accuracy scores were 69.70%, 87.31%, 44.61% and 76.56%, respectively. In 2017, [XXVIII] utilized a deep belief network (DBN) to convert features from low- to high-level representations and used the output of the last hidden layer in the DBN as the input to an SVM. This approach achieved an accuracy of 95.8%, which was higher than when the DBN and SVM were employed separately. In 2017, [XLIII] used a deep learning (DL) neural network for SER. Using a neural network, they learned local and global acoustic emotionally related features extracted from the complete signal and their temporal aggregation according to the utterance, in order to focus on specific regions of the speech signal, looking for the emotionally related features and using a novel feature-pooling strategy. Their experiment was implemented on the interactive emotional dyadic motion capture (IEMOCAP) dataset, which covers four emotions, Happy, Sad, Neutral, and Angry. This strategy results in better performance than does using an SVM, and the learned features work better than does depending on fixed designed features. In 2018, [XXV] proposed a framework that utilized an RNN to classify emotions from the Berlin and Spanish datasets using MFCC and modulation spectral features (MSFs). The RNN results obtained from the proposed framework were compared to multivariate linear regression (MLR) and SVM classifiers. Many experiments were carried out using different combinations of features, datasets and classifiers. The best accuracy on the Spanish dataset (90.05%) was obtained when using both sets of features (MFCC and MSF) and the RNN classifier. The best accuracy on the Berlin dataset (82.41%) was obtained using both sets of features but when using the MLR as a classifier. This result shows that an RNN may work better only under certain conditions. In 2019, [IV] used vocal features (STFT, MFCC) for emotion recognition, an SVM for classification, and then applied a CNN deep-learning approach and context-level analysis of the textual data to improve emotion classification. The results obtained showed that applying the CNN for training improved the performance to 85%, while the accuracy using traditional classification methods such as SVM reached only 48.11%. One reason why Path 7 was added to our new SER diagram was the use of neural networks to convert heterogeneous feature sets to unified feature sets. In 2019, [XXIII] performed multiple comparison experiments for different SER platforms using RAVDESS as the test bed. Their conclusion was that using the log Mel-spectrogram features and a CNN to classify 14 classes (2 gender X 7 emotions) yielded the highest accuracy (68%). In 2019, [LIV] proposed a multitask learning SER method with a self-attention mechanism and extracted features from the spectrogram. The self-attention was used to force the model to concentrate on emotionally rich periods in the speech utterance. Multitask learning was used to make use of the mutual features between gender and emotion classification. The experiment was implemented on the IEMOCAP dataset and their proposed method reached an accuracy of 81.6%. In 2020, [LII] proposed a novel deep neural architecture to convert a heterogeneous acoustic feature set (low- and high-level

Husam Ali Abdulmohsin et al
acoustic features) that might contain unrelated and redundant features into a unified feature set to increase model accuracy. This approach achieved an accuracy of 64% when classifying four emotions extracted from the audio signals in the IEMOCAP dataset. Fuzzy methods are relatively new in the SER field, and very few studies were found until 2019, when [III] proposed a new approach that used fuzzy methods in the SER classification process. An architecture was proposed that facilitated a fuzzy inference system based on fuzzy associative memory. The method combined NN, machine learning (ML), and fuzzy logic based on the observed data. The experiment was applied to the Berlin and SAVEE datasets, and a comparison was performed with other classifiers, such as SVM and Bayes. The best accuracy results on the Berlin and SAVEE datasets were obtained by the fuzzy model: 74.31% and 97.29%, respectively.

XIII. Conclusion

Since the advent of Path 6 in 2011, the field of SER has changed dramatically, and SER methodology has also changed, regardless of how features to recognize emotions are selected. The new trend is to find a specific waveform behavior for each emotion. If that is accomplished, real-time SER will have been achieved. Then, speaker in-dependent SER systems will be invented, and the human-robot emotional relationship will become reality. However, few studies were found when searching for research following Path 6, which means that many methods and techniques remain to be investigated along this path and that it needs more attention from researchers.

Since neural networks were introduced into SER in recent years, high classification accuracies have been reached, and deep learning methods has opened wide research opportunities that may help improve SER.

Through our survey of the SER field from 2011–2020, the following was noticed:

I. Not all emotions have been addresses; many more emotions remain to be investigated in this field. This drawback can be attributed to the available emotion datasets. Although some of the available datasets cover all the primary emotions [XXXVI] and slightly more, many emotions remain to be investigated.

II. Research on real emotions generated from real-life situations, such as telephone system databases and call centers, has not been conducted due to legal issues related to privacy. The datasets that have attempted to reflect real emotions by employing actors have not yet matched the expression levels of real emotions.

III. Acoustic features appear to have the strongest ability to reflect emotions due to their strong relationships with human vocal cords.

IV. The best classification methods used thus far in this field are neural...
networks, which achieve better classification accuracies in less time.

V. Preprocessing is one of the most important phases of speech recognition, especially when working with realistic datasets, which—unlike datasets recorded by actors in standardized laboratories under standardized conditions—require considerable preprocessing.

VI. In 2017, [XXVIII] stated that the five most powerful features used in SER are pitch, short-time zero-crossing, short-time energy, rate formant and MFCC; however, it was noticed that each emotion seems to be related to a specific set of features.

VII. There is a large difference in performance accuracy and quality among the datasets used in SER. Researchers must be careful to select applicable and appropriate datasets for their experiments. Moreover, if possible, it would be better to choose a realistic dataset.

VIII. Features related to the time sequence of the speech signal have been used to obtain highly accurate results because each emotion has a specific effect on the waveform shape.

IX. Local features perform better than do global features in SER.

XIV. Future work

In general, extracting features that are able to classify emotions is the target of most emotion recognition applications, but some emotion recognition applications target a specific emotion, where specific features better than others for that particular emotion; thus, it would be reasonable to conduct a survey on the feature-emotion relationships. For example, many SER applications focus on recognizing fear in human speech more than on other emotions, hoping to capitalize on sudden reactions to save human lives; consequently, a survey of features related to fear would be beneficial to many applications. A survey of the classification algorithms that achieved higher accuracy in the SER research field will be useful for future researches.

Conflict of Interest:

There is no conflict of interest regarding this article.
References


Husam Ali Abdulmohsin et al


Husam Ali Abdulmohsin et al


