electronics

MDPI

Review

A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism

Eva Lieskovská *, Maroš Jakubec, Roman Jarina and Michal Chmulík

Faculty of Electrical Engineering and Information Technology, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia; maros.jakubec@feit.uniza.sk (M.J.); roman.jarina@uniza.sk (R.J.); michal.chmulik@uniza.sk (M.C.)

* Correspondence: eva.lieskovska@feit.uniza.sk

Abstract: Emotions are an integral part of human interactions and are significant factors in determining user satisfaction or customer opinion. speech emotion recognition (SER) modules also play an important role in the development of human–computer interaction (HCI) applications. A tremendous number of SER systems have been developed over the last decades. Attention-based deep neural networks (DNNs) have been shown as suitable tools for mining information that is unevenly time distributed in multimedia content. The attention mechanism has been recently incorporated in DNN architectures to emphasise also emotional salient information. This paper provides a review of the recent development in SER and also examines the impact of various attention mechanisms on SER performance. Overall comparison of the system accuracies is performed on a widely used IEMOCAP benchmark database.

Keywords: speech emotion recognition; deep learning; attention mechanism; recurrent neural network; long short-term memory

check for updates

Citation: Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* **2021**, *10*, 1163. https://doi.org/10.3390/electronics10101163

Academic Editors: Matus Pleva, Yuan-Fu Liao, Patrick Bours and Chiman Kwan

Received: 11 March 2021
Accepted: 9 May 2021
Published: 13 May 2021

1. Introduction

The aim of human–computer interaction (HCI) is not only to create a more effective and natural communication interface between people and computers, but its focus also lies on creating the aesthetic design, pleasant user experience, help in human development, online learning improvement, etc. Since emotions form an integral part of human interactions, they have naturally become an important aspect of the development of HCI-based applications. Emotions can be technologically captured and assessed in a variety of ways, such as facial expressions, physiological signals, or speech. With the intention of creating more natural and intuitive communication between humans and computers, emotions conveyed through signals should be correctly detected and appropriately processed. Throughout the last two decades of research focused on automatic emotion recognition, many machine learning techniques have been developed and constantly improved.

Emotion recognition is used in a wide variety of applications. Anger detection can serve as a quality measurement for voice portals [1] or call centres. It allows adapting provided services to the emotional state of customers accordingly. In civil aviation, monitoring the stress of aircraft pilots can help reduce the rate of a possible aircraft accident. Many researchers, who seek to enhance players' experiences with video games and to keep them motivated, have been incorporating the emotion recognition module into their products. Hossain et al. [2] used multimodal emotion recognition for quality improvement of a cloud-based gaming experience through emotion-aware screen effects. The aim is to increase players' engagement by adjusting the game in accordance with their emotions. In the area of mental health care, a psychiatric counselling service with a chatbot is suggested in [3]. The basic concept consists of the analysis of input text, voice, and visual clues in order to assess the subject's psychiatric disorder and inform about diagnosis and treatment. Another suggestion for emotion recognition application is a conversational chatbot, where speech emotion identification can play a role in better conversation [4]. A real-time SER application should find an optimal trade-off between less computing power, fast processing times, and a high degree of accuracy.

In this review, we focus on works dealing with the processing of acoustic clues from speech to recognise the speaker's emotions. The task of speech emotion recognition (SER) is traditionally divided into two main parts: feature extraction and classification, as depicted in Figure 1. During the feature extraction stage, a speech signal is converted to numerical values using various front-end signal processing techniques. Extracted feature vectors have a compact form and ideally should capture essential information from the signal. In the back-end, an appropriate classifier is selected according to the task to be performed.

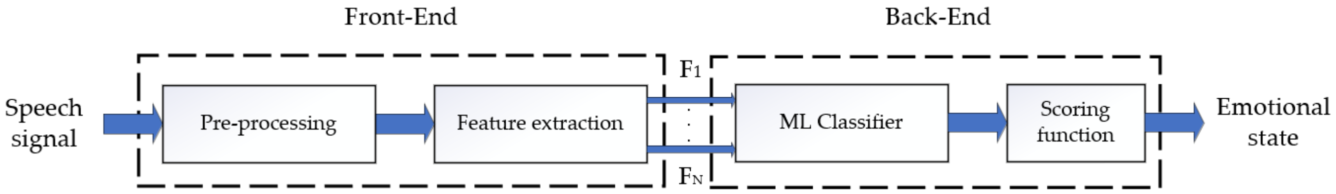

Figure 1. Block scheme of general speech emotion recognition system.

Examples of widely used acoustic features are mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCC), short-time energy, fundamental frequency (F0), formants [5,6], etc. Traditional classification techniques include probabilistic models such as the Gaussian mixture model (GMM) [6–8], hidden Markov model (HMM) [9], and support vector machine (SVM [10–12]. Over the years of research, also various artificial neural network architectures have been utilised, from the simplest multilayer perceptron (MLP) [8] through extreme learning machine (ELM) [13], convolutional neural networks (CNNs) [14,15], to deep architectures of residual neural networks (ResNets) [16] and recurrent neural networks (RNNs) [17,18]. In particular, long short-term memory (LSTM) and gated recurrent units (GRU)-based neural networks (NNs), as state-of-the-art solutions in time-sequence modelling, have been widely utilised in speech signal modelling. In addition, various end-to-end architectures have been proposed to learn jointly both extraction of features and classification [15,19,20].

Besides LSTM and GRU networks, the introduction of an attention mechanism (AM) in deep learning may be considered as another milestone in sequential data processing. The purpose of AM is, as with human visual attention, to select relevant information and filter out irrelevant ones. The attention mechanism, first introduced for a machine translation task [21], has become an essential component of neural architectures. Incorporating AM into encoder–decoder-based neural architectures significantly boosted the performance of machine translation even for long sequences [21,22]. Motivated by the success of attention on machine translation, many researchers have considered it as an essential component of neural architectures for a remarkably large number of applications including natural language processing (NLP) and speech processing. Since emotional salient information is unevenly distributed across speech utterances, an integration of AM into NN architecture is also of interest among the SER research community.

Although several review articles have been devoted to automatic speech emotion recognition [23–29], to the best of the authors' knowledge, a comprehensive overview of SER solutions containing attention mechanisms is lacking. Motivated by this finding, in this article, we provide a review of the recent development in the speech emotion recognition field with a focus on the impact of AM in deep learning-based solutions on SER performance.

The paper is organised as follows: Firstly, the scope and methodology of the survey are discussed in Section 2. In Section 3, we address some of the key issues in deep learning-based SER development. Section 4 provides a theoretical background of the most commonly used neural architectures incorporating AM. Then, we review recently proposed

SER systems incorporating different types of AM. Finally, we compare the accuracy of the selected systems on the IEMOCAP benchmark database in Section 5. The section is concluded by a short discussion on the impact of AM on SER system performance.

2. Scope and Methodology

The paper is divided into two main parts: the first part discusses a general concept of SER and related works, including a description of the novel and deep features, and transfer learning and generalisation techniques, and the focus of the second part is on DNN models incorporating attention mechanism. We used Scopus and Web of Science (WoS) citation databases to search for relevant publications. A number of published papers by year of publication is given in Table 1. This is a general amount of works when searching by the keywords: speech, emotion, recognition, attention. Due to the excessive amount of research work dealing with this topic, only selected papers from the last 4 to 5 years of intensive research are reported in our study. In this review, the speech-related works were mainly taken into consideration; papers dealing with other physiological signals such as EEG, heart rate variability, as well as a fusion of multiple modalities were excluded.

Table 1. Number of publications during the initial search for speech emotion recognition and attention speech emotion recognition.

Year	Scopus		WoS	
	General SER	Attention SER	General SER	Attention SER
2016	519	34	344	30
2017	631	42	348	24
2018	829	82	446	54
2019	979	125	415	63
2020	886	133	325	59

For an additional overview of research work dealing with SER from previous and latest years, we refer a reader to reviews and surveys listed in Table 2. Note, our review does not cover all the topics related to SER such as detailed descriptions of speech features, classifiers, and emotional models, which are addressed more closely in other survey papers. We assume a reader's knowledge in probabilistic and machine learning-based approaches in data classifiers as well as in the basic DNN architectures. To the best of the authors' knowledge, none of the other reviews or surveys (listed in Table 2) deal with attention mechanism in more detail; hence, we consider it to be our main contribution.

Table 2. A brief summary of reviews and surveys related to SER.

References	Description of the Content
[23]; 2011	A comprehensive survey discusses acoustic features, classification methods (both traditional and artificial neural networks (ANNs)), and multimodal approaches. The authors pointed out that some of the existing databases were not sufficient for automatic SER and the development of benchmark emotional speech databases is necessary.
[24]; 2015	Survey from 2000 to 2011 describing various features (considering non-linguistic and linguistic information) and feature selection methods, and providing a comparison of classification performance of traditional classifiers, ANNs, and their combinations. The major shortcoming for direct comparison of SER systems is considered to be a lack of uniformity in the way the methods are evaluated and assessed.

Table 2. *Cont.*

References	Description of the Content
[25]; 2018	The review provides a thorough description of emotion datasets and speech features (excitation source, prosodic and vocal tract features) from 2000 to 2017. It also discusses the classification of emotions in general.
[26]; 2018	A review article, which traces 20 years of progress in SER. The author discusses the techniques of representation of emotional speech (considering audio and textual features) and the ongoing trends. Benchmark results of the SER challenges are also provided.
[27]; 2018	The survey covers existing emotion detection research efforts, emotion models, datasets, detection techniques, their features, limitations, and some possible future directions. Emotion analysis from text is also thoroughly described.
[28]; 2019	Review of the deep learning techniques for SER: RNN, recursive neural network, deep belief network (DBN), CNN, and auto encoder (AE).
[29]; 2020	A review discusses current methodologies in SER. It covers a wide area of SER topics such as emotional models, databases, features, pre-processing methods, supporting modalities, and classifiers. The authors address challenges in SER: the need for natural datasets with a sufficient amount of data; they also pointed out that unsatisfactory results are still being achieved with cross-language scenarios.

Evaluation Metrics

In this section, common metrics of accuracy evaluation are listed. For a multiclass classification task, accuracy is assessed per class firstly and then the average accuracy is determined. This is denoted as unweighted accuracy hereafter. If the class accuracies are weighted according to the number of per-class instances, then the evaluation metric may not reflect the unbalanced nature of data (which is very common with databases of emotional speech). Therefore, the unweighted accuracy is often a better indicator of the system's accuracy. The common evaluation metrics for the SER tasks are as follows:

- Precision is the ratio of all correctly positively classified samples (true positive—TP) to all positive classified samples (TP and false positive—FP). For K-class evaluation, the precision is computed as follows:

$$\mathrm{precision} = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}}{K}. \tag{1}$$

- Recall is the ratio of all correctly positively classified samples (TP) to the number of all samples in a tested subgroup (TP and false negative FN). Recall indicates a class-specific recognition accuracy. Similarly, as in the case of precision, the recall for a multiclass classification problem is computed as the average of recalls for individual classes.

$$\mathrm{recall} = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{K}. \tag{2}$$

- In the literature, the multiclass recall is referred to as unweighted average recall (UAR), which is recommended metric for SER. UAR corresponds to unweighted accuracy (UA), computed similarly as the average over individual class accuracies.
- Weighted accuracy is often given as weighted average recall (WAR), which is computed as the class-specific recalls weighted by the number of per-class instances s_k according to (3). This metric is also interchangeable with weighted accuracy (WA; or accuracy), which is defined as correct predictions over a total number of predictions. Note that evaluation metrics were not clearly defined in previous works. Thus, we unified them as described above.

$$\mathrm{WAR} = \frac{\sum_{k=1}^{K} |s_k| \cdot \mathrm{recall}_k}{\sum_{k=1}^{K} |s_k|} \quad (3)$$

- F1 score is defined as the harmonic mean of the precision and recall.

$$\mathrm{F1} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (4)$$

Note, all of the above-mentioned classification metrics are in the range of [0, 1] (×100 %).

A regression problem is often encountered when dealing with a continuous emotional scale. The appropriate metric for the regression is the correlation coefficient determined in two ways:

- Pearson's correlation coefficient (PCC; ρ) measures the correlation between the true and predicted values (x and y, respectively). Given the pairs of values $\{(x_n, y_n)\}$, n = 1, 2, . . . , N, Pearson's correlation coefficient is computed as follows:

$$\rho = \frac{\sum_{n=1}^{N} (x_n - \mu_x)\left(y_n - \mu_y\right)}{\sqrt{\sum_{n=1}^{N} (x_n - \mu_x)^2 \sum_{n=1}^{N} \left(y_n - \mu_y\right)^2}}, \quad (5)$$

where n denotes the index of the current pair, and μ_x and μ_y are mean values of x_n and y_n, respectively.

- Concordance Correlation Coefficient (CCC; ρ_c) examines the relationship between the true and predicted values from a machine learning model. CCC lies in the range of [−1, 1], where 0 indicates no correlation and 1 is perfect agreement or concordance.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + \left(\mu_x - \mu_y\right)^2}, \quad (6)$$

where μ is the mean value and σ is standard deviation, and ρ is Pearson's correlation coefficient.

A comparison of published SER solutions is difficult due to the different experimental conditions used. Thus, we tried to do at least an intuitive comparative analysis of the published DNN-based SER systems performance. We grouped the systems according to the emotional datasets used for the conduction of experiments. Since the settings of the datasets differ significantly, we also group the compared works according to emotional labelling (discrete/continuous SER) and/or the number of classes being recognised and common cross-validation scenario. For the evaluation, we use the most widely used IEMOCAP database, on which most of the state-of-the-art systems have been tested. For comparison, we also listed the performance of the systems tested on EmoDB and RECOLA datasets.

3. Speech Emotion Recognition and Deep Learning

In this section, we review the most relevant issues in today's SER system development in general: (1) emotional speech database development, (2) speech features extraction and DL based emotion modelling, and (3) selected techniques for SER performance improvement, such as data augmentation, transfer learning, and cross-domain recognition (the attention mechanism is addressed in Sections 4 and 5). A comparison of the state-of-the-art works (excluding AM) based on common criteria is provided at the end of this Section.

3.1. Databases of Emotional Speech

Since the state-of-the-art SER solutions are exclusively based on data-driven machine learning techniques, the selection of a suitable speech database is naturally a key task in building such SER systems. Several criteria have to be taken into account when selecting a proper dataset, such as the degree of naturalness of emotions, the size of the database, and the number of available emotions. The databases can be divided into three basic categories:

- Simulated (acted): Professional actors express emotions through scripted scenarios.
- Elicited (induced): Emotions are created via artificially induced situations. With this approach, it is possible to achieve more natural recordings and simultaneously to have control over the emotional and lexical content of recordings.
- Spontaneous (natural): Spontaneous audio recordings are being extracted from various reality shows. The disadvantage of real-world audio samples is that they may be distorted by background noise and reverberation [30]. Another drawback is that the natural or spontaneous databases often contain unbalanced emotional categories.

Naturally, speech databases are created in various languages, and they may consist of a variety of emotional states. However, emotion labelling is not unified. Recognised emotion can be labelled into several discrete emotional classes, as shown in Table 3. The common way is labelling to six basic (known as the big six) emotional categories—anger, disgust, fear, happiness, sadness, surprise, and neutral. If SER is considered a regression problem, the emotions are mapped to continuous values representing the degree of emotional arousal, valence, and dominance. Valence is a continuum ranging from unhappiness to happiness, arousal ranges from sleepiness to excitement, dominance is in a range from submissiveness to dominance (e.g., control, influence) [31]. In Table 3, the most widely used databases of emotional speech are listed.

We would like also to draw attention to the following issue related to speech emotion rating and annotation. It has to be distinguished between emotion perceived (or observed) and emotion elicited (induced). Unlike in music emotion recognition, or affective analysis of movies where attention is paid to the listener's or spectator's experience, in the case of speech emotion recognition, the focus is on the speaker and his emotional state. The way the data is annotated is of much importance, especially in the case of annotation of spontaneous and induced emotions of the speaker. The emotion in speech is usually annotated by a listener. Another option is to use the rating provided by the speaker himself (felt or induced emotions) or obtained by analysis of the speaker's physiological signals. Since the experimental studies have shown a considerable discrepancy between emotion ratings by speaker and observer, correct and unambiguous emotion rating is still an open issue [32].

Table 3. Comparison of databases of emotional speech.

Database	Language	Num. of Subjects	Num. of Utterances	Discrete Labels	Dim. Labels	Modality
AESDD [33]	Greek	3 F/2 M	500	A, D, F, H, S	–	A
EmoDB [34]	German	5 F/5 M	500	A, B, D, F, H, N, S	–	A
eNTERFACE'05 [35]	English	42	5 utt./emotion	A, D, F, H, N, S, S_r	–	A, V
FAU-AIBO [36]	German	30 F/21 M (children)	18 216	A, B, E_m, H_e, I, J, M, N, O, R, S_r	–	A
IEMOCAP [37]	English	5 F/5 M	10,039	A, D, E, F, F_r, H, N, s, S_r	√	A, V, T, MCF
MSP-PODCAST [38]	English	–	62,140	A, D, F, H, S, S_r, N, C, O	√	A
Polish DB [39]	Polish	4 F/4 M	240	A, B, F, J, N, S	–	A
RAVDESS [40]	English	12 F/12 M	104	A, D, F, H, N, S, S_r	√	A, V
RECOLA [41]	French	46 (27) [1]	–	–	√	A, V, ECG, EDA
SAVEE [42]	English	4 M	480	A, D, F, H, S, S_r, N	–	A, V

Meaning of acronyms are as follows: Num. of subjects: F—female, M—male; Discrete labels: A—anger, B—boredom, C—contempt, D—disgust, E—excitement, E_m—emphatic, F—fear, H—happiness, H_e—helplessness, I—irritation, J—joy, M—motherese, N—neutral, O—other, R—reprimanding, S—sadness, S_r—surprise; Dim. Labels: dimensional labels (arousal, valence, dominance); Modality: A—audio, V—video, T—text, MCF—motion capture of face, ECG—electrocardiogram, EDA—electrodermal activity. [1] Overall, 46 subjects participated in samples recording; however, only 27 subjects were available for audio–visual emotion recognition challenge (AVEC) [43].

3.2. *Acoustic Features*

The purpose of SER is to automatically determine the emotional state of the speaker via a speech signal. Changes in the waveform's frequency and intensity may be observed when comparing different emotionally coloured speech signals [9]. The aim of SER is to capture these variations using different discriminative acoustic features. Acoustic features (referred to as low-level descriptors (LLDs) are often aggregated by temporal feature integration methods (e.g., statistical and spectral moments) in order to obtain features at a global level [44]. High-dimensional feature vectors can be transformed into a compact representation using feature selection (FS) techniques. The aim is to find substantial information from the feature set and discard redundant values simultaneously. In this way, it is possible to optimise the time complexity of the system while maintaining similar accuracy.

Over the many years of research, the focus has been placed on the selection of the ideal set of descriptors for emotional speech. MFCCs originally proposed for speech/speaker recognition are well established also for the derivation of emotional clues. Prosodic descriptors (such as pitch, intensity, rhythm, and duration), as well as voice quality features (jitter and shimmer), are common indicators of human emotions as well [8]. In addition, numerous novel features and feature selection techniques have been developed and successfully applied to SER [7,44–50]. For instance, Gammatone frequency cepstral coefficients proposed by Liu [45] yielded a 3.6% average increase in accuracy compared to MFCCs. Epoch-based features extracted by the zero time windowing also provided emotion-specific and complementary information to MFCCs [46]. Ntalampiras et al. [44] proposed a multiresolution feature called perceptual wavelet packet based on critical-band analysis. It takes into account that not all parts of the spectrum affect human perception in the same way. In [7], the nonlinear Teager–Kaiser energy operator (TEO) was used in combination with MFCC for the detection of stressed emotions. Kerkeni et al. [47] proposed modulation spectral features and modulation frequency features—based on empirical mode decomposition of the input signal and TEO extraction of the instantaneous amplitude and instantaneous frequency of the AM–FM components. Yogesh et al. [48] extracted nonlinear bispectral features and bicoherence features from speech and glottal waveforms.

However, despite great research efforts, there is still no single solution for the most appropriate features. For better comparability of SER systems and their obtained results, attempts to unify feature extraction have been made. When selecting appropriate audio features for SER, it is a common practice to use the openSMILE open-source audio feature extraction toolkit. It contains several feature sets intended for automatic emotion recognition, some of which were proposed in several emotion-related challenges and benchmark initiatives.

- The INTERSPEECH 2009 (IS09) [51] feature set consists of fundamental frequency, voicing probability, frame energy, zero-crossing rate, and 12 MFCCs and their first-order derivatives. With statistical functionals applied to LLDs, 384-dimensional feature vectors can be obtained.
- The feature set of the INTERSPEECH 2010 (IS10) paralinguistic challenge [52] contains 1582 features, which are obtained in three steps: (1) a total of 38 LLDs are smoothed by low-pass filtration, (2) their first order regression coefficients are added, and (3) 21 functionals are applied.
- The extended Geneva minimalistic acoustic parameter set (eGeMAPS) [53] contains LLD features, which paralinguistic studies have suggested as most related to emotions. The eGeMAPS consists of 88 features: the arithmetic mean and variation of 18 LLDs, 8 functionals applied to pitch and loudness, 4 statistics over the unvoiced segments, 6 temporal features, and 26 additional cepstral parameters and dynamic parameters.
- The INTERSPEECH 2013 computational paralinguistic challenge (ComParE) [54] is another feature set from the openSMILE extractor, which is mostly used to recognise emotions. ComParE consists of 6373 features based on extraction of 64 LLDs (prosodic, cepstral, spectral, sound quality), adding their time derivates (delta features), and applying statistical functions.

3.3. Data-Driven Features

Apart from speech parameterisation from handcrafted features, another popular approach is to let a neural network (NN) to perform feature extraction. A typical example is the utilisation of CNN to learn from 2D speech spectrograms, log-mel spectrograms, or even from the raw speech signals [19,55]. CNN is usually supplemented by fully connected (FC) layers and softmax for classification [56]. Architecture, which consists of multiple convolutional layers, is often referred to in literature as deep CNN (DCNN). Huang and Narayanan [55] examined the ability of CNN to perform task-specific spectral decorrelation using log-mel filter-bank (MFB, or log-mel spectrogram) as input features. Since MFCCs are log-mels decorrelated by discrete cosine transform (DCT), the authors demonstrated that the CNN module was a more effective task-specific decorrelation technique under both clean and noisy conditions (experiments were conducted on eNTERFACE'05 [35] database). Aldeneh and Provost [14] experimentally proved that a system based on the minimum set of 40 MFB features and CNN architecture can achieve similar results as SVM trained on a large feature set (1560). Compared to a complex system based on deep feature extraction derived from 1582-dimensional features and an SVM classifier [10], the proposed 40 MFB-CNN provides a more effective and end-to-end solution. Fayek et al. [15] proposed various end-to-end NN architectures to model intra-utterance dynamics. CNN had better discriminative performance than DNN and LSTM architectures, all trained with MFB input features. Vrysis et al. [57] conducted a thorough comparison between standard features, temporal feature integration tactics, and 1D and 2D DCNN architectures. The designed convolutional algorithms delivered excellent performance, surpassing the traditional feature-based approaches. The best 2D DCNN architecture achieved higher accuracy than 1D DCNN with the comparable number of parameters. Moreover, 1D DCNN was four times slower on execution. Hajarolasvadi and Demirel [58] proposed 3D CNN model for speech emotion recognition. The utterances in form of overlapping frames were processed in two ways—88 dimensional features and spectrogram were extracted for each frame. The representation of 3D spectrogram was based on the selection of k most discriminant frames with k-means clustering algorithm applied to the extracted features. Using this approach, it is possible to capture both spectral and temporal information. The proposed architecture was able to outperformed pretrained 2D CNN model transferred to SER task. Mustaqeem and Kwon [59] proposed plain CNN architecture called deep stride CNN, which used strides for downsampling of input feature maps instead of the pooling layer. The authors dealt with proper pre-processing in form of noise reduction through novel adaptive thresholding and decreasing of computational complexity by utilising simplified CNN structure. This stride CNN improved accuracy by 7.85% and 4.5% on IEMOCAP and RAVDESS datasets, respectively and significantly outperformed state-of-the-art systems.

3.4. *Temporal Variations Modelling*

Emotional content in speech varies through time; therefore, it is appropriate to leverage the techniques which are effective for temporal modelling, such as stochastic HMM or neural networks with recurrent units (e.g., LSTM or GRU).

Tzinis and Potamianos [17] studied the effects of variable sequence lengths for LSTM-based recognition (see Section 4 for RNN–LSTM description). Recognition on sequences concatenated at frame-level yielded better results on phoneme length (90 ms). The best results were achieved over statistically aggregated segments at the word level (3 s)—64.16% WA and 60.02% UA (IEMOCAP). In this case, extraction of higher-level statistical functions from multiple LLDs over speech segments led to a more salient representation of underlying emotional dynamics. The proposed solution yielded comparable results to a more complex system based on deep feature extraction and SVM classifiers [10,60].

Recurrent layers are often used in combination with CNN (referred to as CRNN) for the exploitation of temporal information from emotional speech [61]. In this way, both local and global characteristics are modelled. Zhao et al. [62] compared the performance of 1D and 2D-CNN LSTM architectures with raw speech and log-mel spectrograms as input, respectively. Moreover, 2D-CNN LSTM performed better in the modelling of local and global representations than its 1D counterpart. The 2D-CNN LSTM outperformed traditional approaches such as DBN and CNN. Luo et al. [63] proposed a two-channel system with joint learning of handcrafted HSFs/DNN and log-mel spectrogram/CRNN learned features. In this way, it is possible to obtain different kinds of information from emotional speech. The authors also designed another jointly learned architecture—multi-CRNN with one CRNN channel learning from a longer time scale of spectrogram segment and a second CRNN channel for deeper layer-based feature extraction. Their CRNN baseline consisted of CNN–LSTM with a concatenation of three pooling layers (average, minimum, and maximum). Jointly learned SER systems extracted more robust features than the plain CRNN system and HSF–CRNN outperformed multi-CRNN. Satt et al. [64] proposed CNN–BiLSTM architecture with spectrogram as input and worked with two different frequency resolutions. The results indicated that lower resolution yields lower accuracy by 1–3%. The combination of CNN and BiLSTM achieved better results in comparison with the stand-alone CNN model. Moreover, unweighted accuracy was improved by the proposed two-step classification, where special emphasis was put on a neutral class. Ma et al. [65] dealt with the accuracy loss introduced by the speech segmentation process, i.e., division of longer utterances into segments of the same length. They proposed a similar approach to Satt et al. [64] (a combination of CNN and BiGRU), except that spectrogram of the whole sentence, was used as input. They introduced padding values and dealt with the appropriate processing of valid and padded sequences. Moreover, different weights were assigned to the loss so that the length of the sentence does not affect the bias of the model. There was a significant performance improvement over segmentation methods with fixed-length inputs. Compared to [64], the proposed model using variable-length input spectrograms achieved absolute improvements of 2.65% and 4.82%, in WA and UA.

A significant part of the works on SER prefers to model emotions on continuous scale (usually in the activation–valence emotional plane). Several works on continuous SER have also proven that CNN-based data-driven features outperform traditional hand-engineered features [19,66,67]. For example, authors of [19,67] proposed end-to-end continuous SER systems, in which 1D CNN was applied on the raw waveform and temporal dependencies were then modelled by the Bi-LSTM layers. Khorram et al. [66] proposed two architectures for continuous emotions recognition—dilated CNN with a varying dilation factor for different layers and downsampling/upsampling CNN—with different ways of modelling long-term dependencies. AlBadawy and Kim [68] further improved the accuracy of valence with joint modelling of the discrete and continuous emotion labels. Table 4 summarises the top performances of the continuous SER systems tested on the RECOLA dataset.

Table 4. Comparison of continuous SER on RECOLA datasets: A–V = activation–valence, ρ_c—concordance correlation coefficient.

References	Audio Parametrization	Classification Method	Reported Accuracy (ρ_c)	
Trigeorgis et al. [19]; 2016	Raw signal (6 s long sequences)	end-to-end CNN–BiLSTM	0.686 A	0.261 V
Khorram et al. [66]; 2018	MFB	Down/Up CNN	0.681 A	0.502 V
Tzirakis et al. [67]; 2018	Raw signal (20 s long sequences)	end-to-end CNN–LSTM	0.787 A	0.440 V
AlBadawy and Kim [68]; 2018	MFB	Deep BLSTM	0.697 A	0.555 V

3.5. Transfer Learning

The methods based on leveraging pretrained neural networks can often obtain better results than traditional techniques [11,12]. As a result of some studies, pretrained neural networks also outperform randomly initialised networks [69]. The use of transfer learning is especially appropriate for SER, due to the lack of large speech emotion corpora. The deep spectrum features proposed in [12], which were derived from feeding spectrograms through the pretrained network designed for the image classification task, AlexNet [70], is reported to match and even outperform some of the conventional feature extraction techniques. Zhang et al. [11] proposed the use of the AlexNet DCNN pretrained model to learn from three-channel log-mel spectrograms extracted from emotional speech (the additional two channels contained first and second-time derivates of the spectra, known as delta features). The authors also proposed discriminant temporal pyramid matching (DTPM) pooling strategy to aggregate segment-level features (obtained from the DCNN block) to the discriminative utterance-level representations. According to the results obtained with four different databases, AlexNet fine-tuned on emotional speech performed better in comparison with the simplified DCNN model and at the same time, DTPM based pooling outperformed the conventional average pooling method. Xi et al. [16] conducted several experiments with the utilisation of a pretrained model for speaker verification tasks. The authors proposed a residual adapter which is the residual CNN ResNet20 trained on the VoxCeleb2 speaker dataset with adapter modules trained on IEMOCAP emotion data. The residual adapter outperformed ResNet20 trained on emotional data only. This proved the inadequacy of using a small dataset for training with the ResNet architecture.

3.6. Generalisation Techniques

The lack of sufficient size of datasets and their imbalanced nature are problems often encountered in SER. With the increase in complexity and size of DNNs, the need for a large dataset is essential for their good performance. One of the solutions is to extend the dataset by various deformation techniques. This approach is limited by the possibility of losing the emotional content by inappropriate deformation of speech samples. The insufficient amount of data can also be addressed by utilising data from other emotional databases. However, there arises a problem of mismatched conditions between training and testing data or in other words problem of mismatched domains.

3.6.1. Data Augmentation

Audio datasets can be effectively expanded (or augmented) using various deformation techniques such as pitch and/or time shifting, the addition of background noise, and volume control [71]. The addition of various noise levels can expand the dataset up to several times [72]. In this subsection, data augmentation techniques applied specifically for the SER task are briefly listed.

In [14], the augmentation based on speed perturbation resulted in an improvement of 2.3% and 2.8% on IEMOCAP and MSP–IMPROV datasets, respectively. Etienne et al. [73] applied several augmentation techniques on highly unbalanced samples from the IEMOCAP database: vocal tract length perturbation based on rescaling of the spectrograms along the frequency axis, oversampling of classes (happiness and anger), and the use of a higher frequency range. Compared to baseline, the application of all three techniques increased the UA by about 4% (absolute improvement). Vryzas et al. [74] pointed out the fact that changes in the timing and tempo characteristics could result in an undesired loss of emotional clues. They used pitch alterations with constant tempo based on sub-band sinusoidal modelling synthesis for augmentation of data. Although augmentation has not increased the accuracy of the proposed CNN system (for the AESDD dataset [33]), it can improve its robustness and generalisation.

The popular approach of data augmentation is the use of generative adversarial networks (GANs) for generating new in-distribution samples. GAN consists of two networks, which are trained together: generator for generating new samples and discriminator for deciding the authenticity of samples (generated vs. true sample) [75]. Sahu et al. [76] employed vanilla and conditional GAN networks (trained on the IEMOCAP dataset) for generating synthetic feature vectors. The proposed augmentation made slight improvements in SVM's performance when real data were appended with synthetic data. The authors pointed out that a larger amount of data is needed to have a successful GAN framework. Chatziagapi et al. [77] leveraged GAN for spectrogram generation to address the data imbalance. Compared to standard augmentation techniques, authors achieved 10% and 5% relative performance improvement on IEMOCAP and FEEL-25k, respectively.

Fu et al. [78] designed an adversarial autoencoder (AAEC) emotional classifier, through which the dataset was expanded in order to improve the robustness and generalisation of the classifier. The proposed model generated most of the new samples almost within the real distribution.

3.6.2. Cross-Domain Recognition

In the domain adaptation approach, there is an effort to generalise the model for effective emotion recognition across different domains. The performance of a speech emotion recognition system tuned for one emotional speech database can deteriorate significantly for different databases, even if the same language is considered. One may encounter mismatched domain conditions such as different environments, speakers, languages, or various phonation modes. All these conditions worsen the accuracy of the SER system in a cross-domain scenario. Therefore, a tremendous effort has been made to improve the generalisation of the classifier.

Deng et al. [79] proposed unsupervised domain adaptation based on autoencoder. The idea was to train the model on a whispered speech from the GeWEC emotion corpus, while normal speech data were used for testing. Inspired by Universum learning, the authors enhanced the model by integration of the margin-based loss, which adds information from unlabelled data (from another database) to the training process. The results showed that the proposed method outperformed other domain adaptation methods. Abdelwahab and Busso [80] discussed the negative impact of mismatched data distributions between training and testing dataset (target and source domain) on the emotion recognition task. To compensate for the differences between the two domains, the authors used domain adversarial neural network (DANN) [81], which is an adversarial multitask training technique for performing emotion classification tasks and the domain classification. DANN effectively reduced the gap in the feature space between the source and target domains. Zheng et al. [82] presented a novel multiscale discrepancy adversarial (MSDA) network for conducting multiple timescales domain adaptation for cross-corpus SER. The MSDA is characterised by three levels of discriminators, which are fed with global, local, and hybrid levels of features from the labelled source domain and unlabelled target domain. MSDA integrates multiple timescales of deep speech features to train a set of hierarchical domain discriminators and an emotion classifier simultaneously in an adversarial training network. The proposed method achieved the best performance over all other baseline methods. Noh et al. [83] proposed a multipath and group-loss-based network (MPGLN), which supports supervised domain adaptation from multiple environments. It is an ensemble learning model based on a temporal feature generator using BiLSTM, a transferred feature extractor from the pretrained VGG-like audio classification model, and simultaneous minimisation of multiple losses. The proposed MPGLN was evaluated over five multidomain SER datasets and efficiently supported multidomain adaptation and reinforced model generalisation.

Language dependency and emotion recognition with consideration of different languages are common issues that may be encountered in SER. One of the solutions would be to identify language firstly and then to perform language-dependent emotion recognition [5]. Another solution would be to share different language databases and to process them jointly. This is denoted as a multilingual scenario. In the case of a cross-lingual scenario, one dataset is used for training and the other one for testing. Tamulevičius et al. [72] put together a cross-linguistic speech emotion dataset with the size of more than 10.000 emotional utterances. It consists of six emotion datasets of different languages. Moreover, augmentation of data was performed with the addition of white noise and application of Wiener filtering (expansion of dataset up to nine times). For the representations of speech emotion, authors chose several two-dimensional acoustic feature spaces (cochleagrams, spectrograms, mel-cepstrograms, and fractal dimension-based features), and they used CNN for classification. The results showed the superiority of cochleagrams over the other utilised feature spaces and confirmed that emotions are language dependent. With the increase of different language datasets in the training partition, the results obtained by testing with remaining datasets slightly increased.

3.7. DNN Systems Comparison

In this subsection, we tried to do at least a coarse comparison of the performance of related works discussed above (remark, it is not possible to make an exact comparison due to different test conditions, even if the same dataset was used). Note this summary does not contain works incorporating attention mechanisms. The attention mechanism is discussed in Section 4.

We focused on finding common criteria and the selection of datasets for comparative analysis. From literature review, we selected the two most widely used databases—EmoDB and IEMOCAP—and sorted out the related works in terms of the number of emotions used for classification and cross-validation scheme. The resulted comparison of the SER systems on EmoDB and IEMOCAP is in Tables 5 and 6 respectively.

For the EmoDB dataset, we considered research works that used all emotion classes and the leave-one speaker-out (LOSO) method of cross-validation—speaker-independent scenario. The human evaluation of emotions from EmoDB showing the average recognition rate of 84.3% was surpassed by most of the works under comparison.

As seen in Table 5, the system incorporating handcrafted features with proper temporal feature integration method yielded state-of-the-art results (>90% WA) in [44]. Thus, the aggregation of different descriptors carries significant emotional information. However, the disadvantage is that the high dimensional feature sets often cause an increase in computational complexity. The low accuracy of pretrained AlexNet in [84] was caused by the reduction of bandwidth and μ-law companding for the purpose of the development of a real-time SER system (7% reduction in accuracy). Table 5 shows that end-to-end CRNN architecture [62], outperformed other works under comparison.

Table 5. Comparison of SER systems based on classification using a complete EmoDB dataset.

References	Audio Parametrisation	Applied Techniques	Reported Accuracy
Ntalampiras et al. [44]; 2012	Log-likelihood fusion level with optimally integrated feature sets	Simple logistic recognition	93.4% WA
Huang et al. [85]; 2014	Spectrogram	semi-CNN with SVM	85.2% WA
Yogesh et al. [48]; 2017	BSFs, BCFs, IS10 (1632 features) FS: PSOBBO	ELM	90.31% WA
Zhang et al. [11]; 2018	3D Log-mels (static, Δ, ΔΔ) DCNN–DTPM	linear SVM	87.31% WA
Zhao et al. [62]; 2019	Log-mel spectrograms	2D CNN LSTM	95.89% WA
Lech et al. [84]; 2020	Spectrograms converted into RGB	AlexNet (real-time SER)	82% WA

In the case of IEMOCAP, the expansion of highly underrepresented class Happiness, by merging it together with Excitement, naturally yields better results, especially in UA measure. This effect can be seen in the first part of Table 6. (Emotions: A, E + H, N, S). The common procedure for dataset partition is to employ a leave-one session-out cross-validation (fivefold). A common approach is to use data from one speaker for validation and data from the remaining speakers for testing. IEMOCAP contains both scripted and improvised scenarios. Scripted recordings are often not incorporated into SER systems, due to possible correlation with lingual content (systems working with improvised data are marked with an asterisk in Table 6). Note that the SER system trained on the improvised dataset outperformed the system applied on the scripted dataset [86,87]. The degree of naturalness of emotional speech has a significant impact on recognition accuracy. Learning on improvised data only can result in better performance than the combination of improvised and scripted data. This means that better accuracies can often be achieved with smaller data set.

Table 6. Comparison of SER systems for IEMOCAP dataset. Meaning of acronyms: A—anger, E—excitement, H—happiness, N—neutral, S—sadness.

References	Audio Parametrisation	Applied Techniques	Weighted Accuracy	Unweighted Accuracy
Emotions: A, E + H, N, S				
Fayek et al. [15]; 2017	MFB	LSTM–RNN DNN CNN	61.71% WA 62.55% WA 64.78% WA	58.05% UA 58.78% UA 60.89% UA
Aldeneh and Provost [14]; 2017	40 MFB Speed data augment.	CNN	–	61.8% UA
Xia and Liu [10]; 2017	1582 features from IS10 DBN with MTL	SVM	60.9% WA	62.4% UA
Kurpukdee et al. [60]; 2017	ConvLSTM–RNN phoneme-based feature extractor	SVM	65.13% WA	–
Sahu et al. [76]; 2018	1582-dimensional openSMILE feature space Augment. with GAN	SVM	–	60.29% UA
Luo et al. [63]; 2018	6373 HSFs features Log-mel spec.	DNN/CRNN	60.35% WA	63.98% UA
Chatziagapi et al. [77]; 2019	Mel-scaled Spectrograms Augment. with GAN	CNN(VGG19)	–	53.6% UA
Emotions: A, H, N, S				
Lee and Tashev [13]; 2015	Segment-level features + DNN	ELM	52.13% WA *	57.91% UA *
Tzinis and Potamianos [17]; 2017	Statistical features over 3 s segments	LSTM	64.16% WA	60.02% UA
Satt et al. [64]; 2017	STFT spectrograms	CNN–BiLSTM	68.8% WA *	59.4% UA *
Ma et al. [65]; 2018	Variable length spectrograms	CNN–BiGRU	71.45% WA *	64.22% UA *
Yenigalla et al. [4]; 2018	Phoneme embedding and spectrogram	2 CNN channels	73.9% WA *	68.5% UA *
Wu et al. [88]; 2019	Spectrograms	CNN–GRU–SeqCap	72.73% WA	59.71% UA
Xi et al. [16]; 2019	Magnitude spectrograms	Residual Adapter on VoxCeleb2	72.73% WA *	67.58% UA *
Mustaqeem and Kwon [59]; 2019	Noise reduction Spectrograms	DSCNN	84% WA	82% UA

* Improvised data only.

For the IEMOCAP database, with the fivefold cross-validation technique and four emotions for classification (anger, sadness, happiness, and neutral), DNN–ELM [13], based on deep feature extraction and ELM classifier, yielded an accuracy of about 52.13% in WA and 57.91% in UA. These results were considered as a baseline for further evaluation. These results were surpassed by the RNN architecture with the proper extraction of higher-level statistical functionals from multiple LLDs over speech segments. The results of 64.16% WA and 60.02% UA were obtained even on a full dataset (improvised and scripted).

Deep features extracted by CNN often surpass the traditional feature-based approaches [57,89]. A combination of CNN and BiLSTM (CRNN) is effective in the derivation of both local and global characteristics. CRNN often achieves better results in comparison with the stand-alone CNN models [62,64]. Ma et al. [65] emphasised the importance of using the whole sentences for classification because the segmentation of utterances caused the degradation of accuracy. The proposed CRNN architecture with variable-length spectrograms as input features increased the baseline results by 19% and 6% in WA and UA, respectively. Compared to hybrid models, the CRNN end-to-end approach is more effective for implementation.

There is also discussion about the performance of 1D and 2D convolutions. In our study, 2D DCNN outperformed 1D DCNN with a similar number of parameters [57]. Moreover, 1D DCNN was four times slower on execution. In the case of CRNN, 2D-CNN–LSTM outperformed its 1D counterpart in [62]. Yenigalla et al. [4] used phoneme embeddings in addition to spectrograms as input to a model consisting of two separate CNN channels. This two-channel solution further improved results obtained by CRNN proposed by Ma et al. [65] (from 71.45% WA* to 73.9% WA* and from 64.22% UA* to 68.5% UA*). The approach based on transfer learning utilising a pretrained model from a speaker verification task yielded similarly high-performance [16]. The authors further proved the benefits of applying domain-agnostic parameters for SER and the inadequacy of using a small dataset for training with the ResNet architecture. According to Table 6, the deep stride CNN architecture [59] achieved the highest accuracy for both WA and UA. The proposed stride CNN increases the accuracy by using salient features extraction from raw spectrograms and reducing computational complexity. However, the experiments were conducted with an 80/20% split of the dataset, which differs from the LOSO model with an additional validation data partition.

4. Speech Emotion Recognition with Attention Mechanism

Before discussing the attention mechanism, we provide the theoretical background of the LSTM recurrent networks, which were first used as the base architecture for AM.

4.1. LSTM–RNN

Let the input sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$, $\mathbf{X} \in R^{T\times d}$, be transformed by RNN into hidden state vectors representation $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$, $\mathbf{H} \in R^{T\times n}$. Here, d and n denote the dimension of input vectors and the number of hidden units, respectively. A basic principle of RNN lies in the fact that the previous information from sequence $\mathbf{h}_{t-1}$ contributes to shaping the current outcome $\mathbf{h}_t$. Output vector $\mathbf{y}_t$ of the simple RNN is obtained as follows:

$$\mathbf{h}_t = f\,(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1}), \tag{7}$$

$$\mathbf{y}_t = g(\mathbf{V}\mathbf{h}_t), \tag{8}$$

where $\mathbf{W} \in R^{n\times d}$, $\mathbf{U} \in R^{n\times n}$, $\mathbf{V} \in R^{n\times n}$ are learnable weights, and f, g are activation functions.

Note that long-term dependencies in a sequence cannot be captured by a simple RNN unit due to the gradient vanishing problem [90]. Various recurrent units (such as Long short-term memory (LSTM), gated recurrent unit (GRU)) with different internal infrastructure were developed to enable capture dependencies over a longer period.

LSTM [91] uses internal gates to overcome the above-mentioned constraints of the simple recurrent units. The input sequence flows through three types of gates—forget gate $\mathbf{f}_t$ (9), input gate $\mathbf{i}_t$ (10), and output gate $\mathbf{o}_t$ (13). Another component of LSTM is a memory cell $\mathbf{c}_t$ (12), whose state is updated at each time step. The process of cell state update depends on the previous hidden state vector $\mathbf{h}_{t-1}$, current input vector $\mathbf{x}_t$, and the previous cell state $\mathbf{c}_{t-1}$ (previous cell state can be also included into gates, and this is called peephole connection). The inner structure of LSTM is shown in Figure 2. Here, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ denotes input sequence, where T is the length of the sequence. The individual operations in LSTM are formalised as follows:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{V}_f\mathbf{c}_{t-1} + \mathbf{b}_f), \tag{9}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i x_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{V}_i\mathbf{c}_{t-1} + \mathbf{b}_i), \tag{10}$$

$$\mathbf{z}_t = \tan h(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{b}_z), \tag{11}$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \mathbf{z}_t\,, \tag{12}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{V}_o\mathbf{c}_t + \mathbf{b}_o), \tag{13}$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tan h(\mathbf{c}_t). \tag{14}$$

Here, $\mathbf{W}_l \in R^{n\times d}$, $\mathbf{U}_l \in R^{n\times n}$, $\mathbf{V}_l \in R^{n\times n}$, and $\mathbf{b}_l \in R^n$, $l \in \{f, i, z, o\}$ are weight matrixes and bias terms. Tanh and σ are the hyperbolic tangent function and sigmoid function. Sign ∘ denotes the Hadamard product.

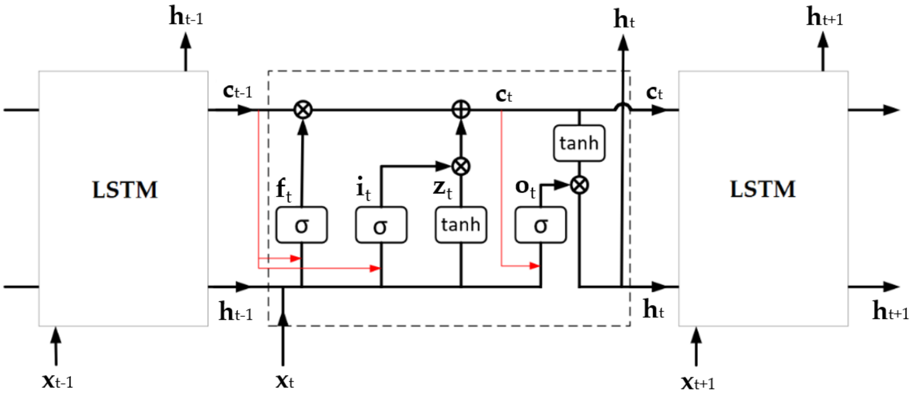

Figure 2. Detail of inner structure of LSTM. The peephole connections are depicted with red lines.

In contrast to LSTM, which incorporates past information into DNN, the ability to look into the future is added in bidirectional LSTM architecture (BiLSTM). As the name implies, BiLSTM is composed of forward and backward LSTM layers. The calculation process of layers depends on the way from which direction the input sequence is read.

4.2. Attention Mechanism

Incorporation of the attention mechanism (AM) into DNN-based SER systems was often motivated by research in the NLP field [18,91,92] and computer vision [92]. We give a brief explanation of the attention mechanism from the NLP's point of view due to the similarity of the tasks. "Language" attention can be traced back to work related to neural machine translation [21]. Here, the typical encoder–decoder approach was supplemented by the network's ability to soft-search for salient information from a sentence to be translated. The authors used BiRNN/RNN as encoder/decoder, both with the GRU inner structure [93]. The machine translation decoding process can be described as the prediction of the new target word $\mathbf{y}_t$, which is dependent on context vector $\mathbf{c}$ obtained from a current sentence and previously predicted words [93].

$$\mathrm{P}\left(\mathbf{y}_t \mid \mathbf{y}_{<\,t}\,,\,\mathbf{c}\right) = \mathrm{g}\left(\mathbf{h}_t\,,\,\mathbf{y}_{t-1},\,\mathbf{c}\right) \tag{15}$$

Fixed encoding of sentences, which was considered to be a drawback in performance, was substituted by a novel attention mechanism. The main idea behind the attention is to obtain a context vector created as a weighted sum of encoded annotations (18), while attention weights $\mathbf{a}$ are learned by the so-called alignment model (16)—i.e., jointly trained feedforward neural network.

$$\mathrm{e}_{kj} = \mathbf{v}_a^T \tan h\left(\mathbf{W}_a \mathbf{h}_{k-1} + \mathbf{U}_a \mathbf{h}_j\right) \tag{16}$$

$$\mathbf{a}_{kj} = \frac{\exp(\mathrm{e}_{kj})}{\sum_{\tau=1}^{T} \exp(\mathrm{e}_{k\tau})} \tag{17}$$

$$\mathbf{c}_k = \sum_{j=1}^{T} \mathbf{a}_{kj} \mathbf{h}_j \tag{18}$$

where $\mathbf{v}_a \in \mathrm{R}^n$, $\mathbf{W}_a \in \mathrm{R}^{n\times n,}$ and $\mathbf{U}_a \in \mathrm{R}^{n\times 2n}$ are weight matrices. Assuming two RNNs as the encoder and decoder, the attention weights are obtained by considering hidden states of the encoder $\mathbf{h}_j$ and hidden states of the decoder $\mathbf{h}_{k-1}$ of the last predicted word. A context vector is computed at each time step and the proposed network architecture is trained jointly. Figure 3 shows a general scheme of the described process incorporating AM.

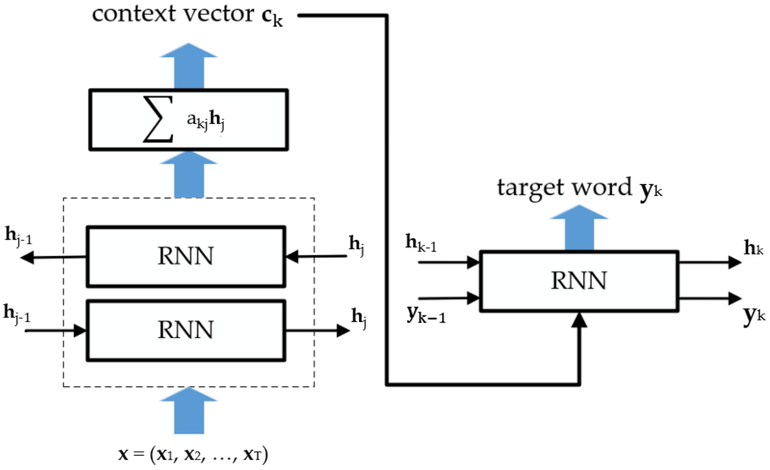
context vector $\mathbf{c}_k$
$\sum a_{kj}\mathbf{h}_j$
$\mathbf{h}_{j-1}$
RNN
$\mathbf{h}_j$
$\mathbf{h}_{j-1}$
RNN
$\mathbf{h}_j$
target word $\mathbf{y}_k$
$\mathbf{h}_{k-1}$
$\mathbf{y}_{k-1}$
RNN
$\mathbf{h}_k$
$\mathbf{y}_k$
$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$

Figure 3. Encoder–decoder framework with an attention mechanism.

AM Modifications

As during the last years, numerous AM concepts and variations have been proposed and implemented, several different taxonomies of AM already exist. Different strategies of classification of AM can a reader find e.g., in [94,95]. Here, we point out some of the key works addressing different implementations of AM.

Luong et al. [22] proposed implementing AM globally and locally. Global attention uses whole information from a source sentence. In this case, the context vector was computed as the weighted average of all source hidden states, while attention weights were obtained from the current target hidden state $\mathbf{h}_k$ and each source hidden state $\mathbf{h}_j$. This approach works on a principle similar to Bahdanau et al. [21], but it differs in simplified computation. Moreover, various alignment functions were examined (see Table 7). As the name implies, *local attention* focuses only on the subset from the source sentence. It is a computationally more efficient method. Context vector takes into account a preselected range of source hidden states with an aligned position corresponding to each target word. Thus, this type of context vector has a fixed length. The aligned position is either at the current target word at time t or can be learned to be predicted. According to results, dot alignment worked well for the global attention and general was better for the local attention. The best performance achieved local attention model with predictive alignments. The machine translation model with the attention mechanism outperformed conventional non-attentional models.

Table 7. Computation of different alignment scores.

Dot	$\mathbf{h}_k^T\mathbf{h}_j$
General	$\mathbf{h}_k^T\mathbf{W}_a\mathbf{h}_j$
Concatenation	$\mathbf{v}_a^T \tanh\left(\mathbf{W}_a\left[\mathbf{h}_k, \mathbf{h}_j\right]\right)$

Lin et al. [96] applied AM on sentiment analysis tasks. This approach allowed the system to perform a standalone search for significant parts of a sentence and thus reducing redundant information. Firstly, BiLSTM encoded words from source sentences into individual hidden states **H** and then the attention weights are computed as an alignment model from **H**. Sentence embedding vector was created as a weighted sum of hidden states. It was not enough to focus only on a certain component of the sentence. Therefore, a concept of multiple hops of attention was proposed, where more such embeddings for different parts of the sentence were created. The sentence embeddings in a form of 2D matrices were then used for sentiment recognition. Moreover, the authors proposed a penalisation technique to ensure that the summation weights cannot be similar.

AM is also a powerful tool for fine-grained aspect-level sentiment classification. Based on the aspect information, the sentiment of the sentence can take on different meanings. Wang et al. [97] firstly proposed an embedding representation of each aspect. Then attention-based LSTM learns the sentiment of a given sentence and is able to focus on important parts by considering a given aspect. Aspect embeddings were incorporated as concatenation to hidden states vectors and attention weights were obtained subsequently. Embeddings could be additionally appended to word vectors as well. In this way, the information from the aspect is preserved in a hidden vector. This novel approach for aspect-level sentiment classification outperformed baseline systems. In [98], the aspect expression from sentences was formed as a weighted summation of aspect embeddings. The number of aspects was preselected and the weights were computed so that context information, as well as aspect expression, were included. An unsupervised objective was applied to improve the training procedure. Another way how to improve the attention model was the inclusion of words, which are in vicinity to the target aspect expression. This method takes advantage of the fact that that context words closer to the target offer complementary clues in sentiment classification. The application of both methods improved results in comparison with various LSTM attention systems.

Chorowski et al. [99] divided encoder–decoder-based attention mechanism into three different categories according to parameters used during the alignment process. Here, the computation of attention weights vector $\mathbf{a}_k$ can be based on location in form of previous attention vector $\mathbf{a}_{k-1}$, current content **H,** or a combination of both in hybrid AM. Table 8 shows different implementations of AM. Even though hybrid AM seeming to be the best solution for encoder–decoder based speech recognition [99], the decoder part is omitted in SER, and therefore, the AM for SER task is simplified.

Table 8. The implementations of the attention mechanisms.

Location-based AM	$\mathbf{a}_k = Attend(\mathbf{h}_{k-1}, \mathbf{a}_{k-1})$
Content-based AM	$\mathbf{a}_k = Attend(\mathbf{h}_{k-1}, \mathbf{H})$
Hybrid AM	$\mathbf{a}_k = Attend(\mathbf{h}_{k-1}, \mathbf{a}_{k-1}, \mathbf{H})$

4.3. Attention Mechanism in Speech Emotion Recognition

This section provides a description of various implementations of AM for speech emotion recognition. As for emotional speech, one label is often used to characterise the whole utterance, although it is clear that the sentence may contain unemotional and silent intervals as well. Therefore, the searching techniques for important parts of emotional speech have been developed.

The first attempts to make the model focus on emotionally salient clues were proposed before the invention of the attention weights. Han et al. considered the speech segments with the highest energy to contain the most prominent emotional information [100]. Lee and Tashev [13] proposed the BiLSTM–ELM system for SER and the importance of each frame is decided using the expectation maximisation algorithm. Moreover, to represent the uncertainty of emotional labels, a speech sample is able to acquire one of two possible states—given emotion and "zero" emotion. The benefit of this system was leveraging RNN's ability to model long contextual information from emotional speech and addressing the uncertainty of emotional labels. The BiLSTM–ELM outperformed the DNN–ELM system, implemented according to [100], with 12% and 5% absolute improvement in UA and WA, respectively.

Most of the attention mechanisms in the SER field are based on the previously described method of attention weights computation using Equations (16) and (17). However, various modifications of AM were proposed, e.g., different input features can be used (feature maps) and simplified computations were developed (the decoder part is omitted for SER systems).

4.3.1. Attentive Deep Recurrent Neural Networks

Huang and Narayanan [101] implemented two types of attention weights: content-based AM (19) inspired by [21,99] and its simplified version (20).

$$\mathbf{a}_\mathrm{j}= \mathrm{softmax}\left(\mathbf{v}_\mathrm{a}^\mathrm{T}\sigma_\mathrm{a}\left(\mathbf{W}_\mathrm{a}\mathbf{h}_\mathrm{j}\right)\right) \tag{19}$$

$$\mathbf{a}_\mathrm{j}= \mathrm{softmax}\left(\mathbf{v}_\mathrm{a}^\mathrm{T}\mathbf{h}_\mathrm{j}\right) \tag{20}$$

In order to avoid overfitting, the authors proposed separate training of BiLSTM and AM components as well as application of dropout before the summation of hidden vectors. According to the results, the simplified implementation of the attention weights defined by (20) yielded better results. The AM-based system outperforms the non-AM system—an improvement from 57.87% to 59.33% in WA and from 48.54% to 49.96% in UA was observed. Moreover, the authors experimentally proved that the attention selection distribution was not just correlated to the frame energy curve.

In [18], Mirsamadi et al. pointed out the fact that only a few words in the labelled utterance were emotional. They highlighted the importance of considering silence intervals and emotionless parts of the utterance as well. Here, the attention weights were computed using the softmax function on the inner product between trainable attention vector $\mathbf{u}$ and RNN output $\mathbf{y_t}$ at each time step, similarly as (20). In the subsequent step, the weighted average in time was performed, and the softmax layer was applied for final emotion classification. This deep RNN architecture with AM is able to focus on emotionally significant cues and on their temporal variations at the utterance level. The proposed combination of BiLSTM and the novel mean-pooling approach with local attention revealed improved performance over many-to-one training and slightly increased results over the mean-pooling method. With only 32 LLDs, the absolute improvement of 5.7% and 3.1% (in WA and UA) was achieved over the traditional SVM model, which needed additional statistical functions for satisfactory results. Tao and Liu [102] discussed the limitation of the time-dependent RNN model and the proposed advanced LSTM (A–LSTM) for better temporal context modelling. Unlike LSTM, which uses the previous state to compute a new one, A–LSTM makes use of multiple states by combining information from preselected time steps. The weights were learned and applied to the inner states of LSTM. The authors proposed the DNN–BiLSTM model with the learning of multiple tasks—emotion, speaker, and gender classification. Moreover, BiLSTM was followed by an attention-based weighted pooling layer. A relative improvement of 5.5% was achieved with A–LSTM, compared to conventional LSTM. Thus, the time dependency modelling capability of LSTM was improved. The proposed solution did not outperform Mirsamadis attentive RNN [18].

AM was also introduced into the forgetting gate $\mathrm{f_t}$ of LSTM cell in [103]. Here, the updating of the cell state (21) is viewed as a weighted sum of the previous cell state $\mathbf{c_{t-1}}$ and the current value for update $\mathbf{z_t}$.

$$\mathbf{c}_\mathrm{t}= \mathbf{f}_\mathrm{t} \circ \mathbf{c}_{\mathrm{t}-1} + (1 - \mathbf{f}_\mathrm{t}) \circ \mathbf{z}_\mathrm{t} \tag{21}$$

$$\mathbf{f}_\mathrm{t}= \sigma(\mathbf{W}_\mathrm{f}\tan\mathrm{h}(\mathbf{V}_\mathrm{f}\,\mathbf{c}_{\mathrm{t}-1})) \tag{22}$$

The weights for the cell state updating were obtained by training of the self-attention model (20), with $\mathbf{W}_\mathrm{f} \in \mathrm{R}^{\mathrm{n}\times\mathrm{n}}$ and $\mathbf{V}_\mathrm{f} \in \mathrm{R}^{\mathrm{n}\times\mathrm{n}}$ as trainable parameters. Calculation complexity of the proposed attention gate was reduced by taking into account only the cell state at the previous moment $\mathbf{c}_{\mathrm{t}-1}$. The ComParE frame-level features were used for classification, while the proposed network had the ability to learn high-level dependencies. The second AM was utilised in the output gate. It was in form of weights applied in both time and feature dimensions. Compared to the traditional LSTM, the obtained results showed an absolute improvement of 2.8%, 13.8%, and 8.5% in UAR for CASIA, eNTERFACE, and GEMEP, respectively. Xie et al. [104] proposed a dense LSTM with attention-based skip connections between the layers. In order to address the variable distribution of significant emotional information in speech, attention weights were incorporated into the LSTMs output in the time dimension. This approach was inspired by the global attention described in [22]. Assuming that different speech features have different abilities to distinguish emotion categories, weighting on feature dimension was also implemented. Results showed that attention applied to the output of each layer improved unweighted average recall and accelerated convergence speed in comparison with the general LSTM approach.

4.3.2. Attentive Deep Convolutional Neural Network

Neumann and Vu [86] performed a comparison of different speech features with an attentive CNN architecture. It contains an attention layer based on a linear scoring function. Additionally, the authors applied MTL for both categorical and continuous labels (activation and valence). The results indicated a small difference in performance between MFB, MFCC, and eGeMAP features and a slight improvement of accuracy with the MTL approach. The best results were reported with a combination of MFB features, attentive CNN with MTL learning. Li et al. [92] used two types of convolution filters for extraction of time-specific and frequency-specific features from the spectrograms. Feature extraction was followed by CNN architecture for modelling high-level representation. Inspired by attention-based low-rank second-order pooling proposed for the task of action classification from single RGB images [105], the authors applied this novel pooling method after the last convolutional layer. It was based on a combination of two attention maps—the class-specific top-down and class-agnostic bottom-up attention. The authors reported on the strong emotional representation ability of the proposed architecture. In order to preserve the information from variable length utterance as a whole without the need for segmentation, Zhang et al. [69] designed fully convolutional network (FCN) architecture—adapted AlexNet with removed fully connected layers. The proposed pretrained FCN architecture takes spectrograms of variable length as input without the need for division of utterances or padding to the required length [64,65]. Furthermore, the attention mechanism identifies important parts of spectrograms and ignores nonspeech parts. FCN architecture outperformed the nonattentive CNN–LSTM method proposed in [64] and achieved comparable results with attention-based convolutional RNN [106]. Thus, the proposed FCN architecture is able to capture the temporary context without the need for additional recurrent layers.

4.3.3. Attentive Convolutional–Recurrent Deep Neural Network

In many cases, the extraction of large feature sets is replaced by direct learning of emotional speech characteristics by deep CNN architectures. Satt et al. [64] segmented utterances into 3 s intervals firstly. Then, the spectrograms were extracted and were directly fed to the CNN–LSTM architecture. Harmonic modelling was applied on spectrogram to eliminate nonspeech parts of the emotional utterance. This step was particularly useful for the classification of emotion in noisy conditions. Lastly, the attention mechanism was added to the LSTM layer, which did not improve the achieved results. Zhao et al. [107] used two streams for feature extraction—fully convolutional network (FCN) with temporal convolutions and Attention–BiLSTM layers—and concatenated the outputs for further DNN based classification. The results indicated improvements over attention–BiLSTM and Att–CNN [86] architectures. Sarma et al. [20] proposed a raw speech waveform-based end-to-end time delay neural network (TDNN) with LSTM–attention architecture. Accuracy improvement on the IEMOCAP database, as well as reduction of confusion among individual categories, was observed with the use of AM. Huang and Narayanan [55] proposed CLDNN architecture with the convolutional AM. System leveraged task-specific spectral decorrelation of CNN applied on log-mel features and temporal modelling by BiLSTM layers. Main modules were frozen during the training of attention weights. Improved results were achieved with the use of AM under the clean test-set conditions. Chen et al. [106] discussed the negative impact that the personalised features (containing speaker’s characteristics, content, etc.) have on the ability of the SER system to generalise well. Assuming that the time derivates of the coefficients (delta features) reduce these undesirable effects, a 3D log-mel spectrogram (consisted of log-mels including delta and delta–delta features) was proposed for the compensation of the personalised features. The authors proposed an attention-based convolutional RNN system (ACRNN) for emotion recognition. When compared with DNN–ELM-based system [100], 3D-ACRNN achieved significant improvement in recognition accuracy on IEMOCAP and EmoDB databases. 3D-ACRNN also outperformed 2D-ACRN based on standalone log-mels. Li et al. [108] proposed an end-to-end self-attentional CNN–BiLSTM model. The attention mechanism based on the same procedure as in [96] concentrates on salient parts of speech. Additionally, the gender recognition task was added to improve emotion recognition in a multitask learning manner. As the gender of the speaker affects the emotional speech, these variations can be taken advantage of. The state-of-the-art results were reported with increased overall accuracy on the IEMOCAP database. Dangol et al. [109] proposed an emotion recognition system based on 3D CNN–LSTM with a relation-aware AM that integrates pairwise relationships between input elements. The 3D spectrogram representations provided both spectral and temporal information from the speech samples. In order to increase the accuracy of emotion recognition, the computation process of attention weights was modified and the synthetic individual evaluation oversampling technique was used to update the feature maps.

In [110], the authors used prosodic characteristics with a fusion of three classifiers working at the syllable, utterance, and frame levels. They used a combination of methods such as the mechanism of attention and the feature selection based on RFE. System performance was improved by identification of relevant features, incorporating attention and score-level fusion. Zheng et al. [111] performed ensemble learning by the integration of three models/experts, each focusing on different feature extraction and classification tactics. Expert 1 is a two-channel CNN model that effectively learns time- and frequency-domain features. Expert 2 is GRU with AM that learns short-term speech characteristics from the principal component analysis (PCA) processed spectrograms with a further fusion of mean value features of the spectrograms. Expert 3 performs end-to-end multilevel emotion recognition using BiLSMT with attention mechanism with a combination of local (CRNN model learning from speech spectrum) and global features (HSFs). Each expert accessed emotional speech in a different way and their combination reduced the negative effects of data imbalance and results in better generalization ability.

For better clarity, the AM-based SER systems are also summarised in Table 9.

Table 9. Comparison of SER systems with an attention mechanism. Meaning of acronyms: A—anger, E—excitement, F_r—frustration, H—happiness, N—neutral, S—sadness; A/V—activation/valence.

References	Techniques of Audio Parametrisation	Proposed Machine Learning Method	Database (Emotions)
	IEMOCAP		
Huang and Narayanan [101]; 2016	28 LLDs: 13 MFCC, F0, Δ	BiLSTM	A, H, N, S
Mirsamadi et al. [18]; 2017	F0, voice probab., frame energy, ZCR, 12 MFCC, Δ	BiLSTM—weighted-pooling with local attention	A, H, N, S
Neumann and Vu [86]; 2017	Max. length of the utterance: 7.5 s MFB (26)	Attentive CNN with MTL	A, E + H, N, S A–V
Tao and Liu [102]; 2018	13 MFCC, ZCR, energy, entropy of energy, spectral characteristics, 12 D chroma, chroma dev., HR, pitch	DNN–BiLSTM–MTL with Advanced LSTM	A, H, N, S
Zhao et al. [107]; 2018	743 features + PCA	Att–BiLSTM–FCN	A, E + H, N, S

Table 9. *Cont.*

References	Techniques of Audio Parametrisation	Proposed Machine Learning Method	Database (Emotions)
Sarma et al. [20]; 2018	Raw waveform front end	TDNN–LSTM–attention	A, H, N, S
Chen et al. [106]; 2018	3D Log-mel spectrograms	Attention-based convolutional RNN (ACRNN)	A, H, N, S
Li et al. [92]; 2018	Spectrogram (2 s segments with 1 s overlap)	CNN–TF–Att.pooling	A, H, N, S
Xie et al. [104]; 2019	The ComParE frame-level features	LSTM with skipped connections	A, E, F_r, N, S
Zhang et al. [69]; 2019	Spectrogram (variable utterance length)	Fully convolutional network + attention layer	A, H, N, S
Li et al. [108]; 2019	Mel spectrogram + Δ, ΔΔ (max. length of the utterances: 7.5 s)	CNN–BiLSTM–MTL: + Attention mechanism	A, E + H, N, S
Alex et al. [110]; 2020	Prosodic and spectral features extracted at various levels + RFE	Fusion of three separate DNNs + Attention at the syllable-level	A, E + H, N, S
Zheng et al. [111]; 2020	(1) Spectrogram (2) Spectrogram + PCA (3) LLDs and their HSFs; spectrogram and CRNN with attention m.	Ensemble model: (1) two-channel CNN (2) GRU with attention m. (3) BiLSTM with attention m.	A, E + H, N, S
Dangol et al. [109]; 2020	Silence/noise removal 3D Log-mel spectrograms	Relation-aware attention-based 3D CNN–LSTM	A, H, N, S
	Other databases		
Huang and Narayanan [55]; 2017	MFB	CLDNN with convolutional attention mechanism	eNTERFACE'05
Chen et al. [106]; 2018	3D Log-mel spectrograms	Attention-based convolutional RNN (ACRNN)	EmoDB full data set
Xie et al. [103]; 2019	The ComParE frame-level features (openSMILE)	LSTM with attention gate and time/frequency attention	CASIA, (6) eNTERFACE (6) GEMEP (12)
Xie et al. [104]; 2019	The ComParE frame-level features (openSMILE)	LSTM with skipped connections	eNTERFACE (6)
Dangol et al. [109]; 2020	Silence/noise removal 3D Log-mel spectrograms	Relation-aware attention-based 3D CNN and LSTM	EmoDB SAVEE

5. Impact of Attention Mechanism on SER

We performed a comparison of related works based on the most common settings to study the impact of AM on speech emotion recognition. We applied the same methodology as in Section 3.7. Since IEMOCAP is the most commonly used database in the published works, we chose it for further analysis.

Tables 10 and 11 show the comparison of SER systems on IEMOCAP for two kinds of classes of emotions: (1) anger, happiness, neutral and sad and (2) an extension of the 'excitement' class. As previously explained, it is not possible to make an exact comparison of the systems due to different test conditions, even if the same dataset was used. Thus, the reported accuracies listed in Tables 10 and 11 provide only coarse information in terms of their performance comparison.

Table 10. Comparison of system accuracies on IEMOCAP database for four emotions. Meaning of acronyms: AM—attention mechanism, A—anger, H—happiness, N—neutral, S—sadness.

References	AM	Description of System	Emotions	WA	UA
		Recurrent architectures			
[101]; 2016	√	28 LLDs BiLSTM	A, H, N, S	59.33%	49.96%
[18]; 2017	√	32 LLDs BiLSTM—with local AM	A, H, N, S	63.5%	58.8%
[17]; 2017	×	Statistical features over 3 s segments and LSTM	A, H, N, S	64.16%	60.02%
[102]; 2018	√	LLDs Advanced LSTM	A, H, N, S	55.3%	–
		Convolutional architectures			
[92]; 2018	√	Spectrograms CNN–TF–Att.pooling	A, H, N, S (improvised)	71.75%	68.06%
[4]; 2018	×	Phoneme embedding and spectrogram Two CNN channels	A, H, N, S (improvised)	73.9%	68.5%
[69]; 2019	√	Spectrogram and FCN + attention layer	A, H, N, S (improvised)	70.4%	63.9%
[16]; 2019	×	Magnitude spectrograms Residual Adapter on VoxCeleb2	A, H, N, S (improvised)	72.73%	67.58%
		Combination of CNN and RNN			
[64]; 2017	×	Spectrograms CNN–BiLSTM	A, H, N, S (improvised)	68.8%	59.4%
[20]; 2018	√	Raw waveform front end TDNN–LSTM–attention	A, H, N, S	70.1%	60.7%
[65]; 2018	×	Spectrograms CNN–BiGRU	A, H, N, S (improvised)	71.45%	64.22%
[106]; 2018	√	3Dlog-mel spec.; Att.–CRNN	A, H, S, N (improvised)	–	64.74%
[88]; 2019	×	Spectrograms CNN–GRU–SeqCap	A, H, N, S	72.73%	59.71%
		Hybrid systems			
[13]; 2015	×	Segment-level features DNN–ELM	A, H, N, S (improvised)	52.13%	57.91%
[13]; 2015	×	32 LLDs BiLSTM–ELM	A, H, N, S (improvised)	62.85%	63.89%
[10]; 2017	×	DBN–MTL feat. Extract. SVM classifier	A, H, N, S	60.9%	62.4%

Table 11. Comparison of system accuracies on IEMOCAP database for additional combination of excitement and happiness. Meaning of acronyms: AM—attention mechanism, A—anger, E—excitement, H—happiness, N—neutral, S—sadness.

References	AM	Description of System	Emotions	WA	UA
		Convolutional architectures			
[86]; 2017	√	MFB; Attentive CNN with MTL	A, E + H, N, S A–V	56.10%	–
[14]; 2017	×	MFB and CNN	A, E + H, N, S	–	61.8%

Table 11. *Cont.*

References	AM	Description of System	Emotions	WA	UA
[15]; 2017	×	Log-mel spectrogram ConvNet	A, E + H, N, S	64.78%	60.89%
[77]; 2019	×	Mel-scaled spectrograms Augment. With GAN CNN(VGG19)	A, E + H, N, S	–	53.6%
		Combination of CNN and RNN			
[107]; 2018	√	743 features + PCA Att–BiLSTM–FCN	A, E + H, N, S	59.7%	60.1%
[108]; 2019	√	Log-mel spectrograms, Δ, ΔΔ; CNN–BiLSTM with MTL	A, E + H, N, S	81.6%	82.8%
		Hybrid systems and ensemble models			
[60]; 2017	×	ConvLSTM feature extractor	SVM	65.13%	–
[63]; 2018	×	HSFs–DNN Log-mel spec.-CRNN	A, E + H, N, S	60.35%	63.98%
[76]; 2018	×	1582-dimensional openSMILE feature space Augment. With GAN	SVM	–	60.29%
[111]; 2020	√	Ensemble model	A, E + H, N, S	75%	75%

The following conclusions, in particular, can be drawn from the works under study:

- AM has improved over the last years and a growing trend of AM use can be observed. Certainly, the performance improvement when using AM is evidenced by many research studies on SER [18,20,69,92,102–104,107,108,111]. On the other hand, two works [63,68] did not report improvements when using AM. Learning the attention weights for emotional representations of speech seems to be a reasonable way to address the variability of emotional clues across utterance; however, we have to note that the resulting benefit in terms of accuracy increment is not always so obvious. As seen from Tables 10 and 11, the properly configured systems without AM may outperform the systems with AM (although one may argue about the correctness of such judgment due to different testing conditions among published works). The reason for ambiguity might be that AM-based SER system performance is subject to implementation issues as follows:
    - The implementation of appropriate AM can be linked to various factors such as the derivation of accurate context information from speech utterances. As in NLP, the better the contextual information obtained from the sequence, the better the performance of the system. The duration of divided segments significantly influences the accuracy of emotion recognition [20,63,86]. Therefore, appropriate input sequence lengths must be determined in order to effectively capture the emotional context.
    - Proper representation of emotional speech is also an important part of deriving contextual information. RNN is suitable for modelling long sequences. Extraction of higher-level statistical functions from multiple LLDs over speech segments with a combination of LSTM [18] can be compared to 32 LLDs with BiLSTM and local AM [18]. Transfer learning is a suitable solution particularly for small emotional datasets [16]. However, more works should be considered to make conclusions. End-to-end systems that combined CNN as feature extractor and RNN for modelling of the long-term contextual dependencies achieved high performance on IEMOCAP data and on EmoDB [62,106]. Various combinations of RNN and CNN are able to outperform separate systems [62,107]. The two-channel CNN taking phoneme embeddings and spectrograms on input seem to further improve the accuracy [4]. Thus, it can be beneficial to allow the model

to learn different kinds of features. Moreover, leveraging multitask Learning for both the discrete and continuous recognition tasks improves the accuracy of SER systems [10,112]. CRNN architecture together with multitask learning was a part of the state-of-the-art solution on IEMOCAP proposed in [108]. Here, AM clearly improved system performance.

- Recurrent networks provide temporal representation for the whole utterance and better results are obtained with its aggregation by pooling for further recognition [18,20]. Several works compare different pooling strategies. The attention pooling is able to outperform global max pooling and global average pooling (GAP) [18,102,107]. The same was true for the attention pooling strategy for convolutional feature maps in [92] (attention-based pooling outperformed GAP). It can be concluded that learning of the attention weights indeed allows the model to adapt itself to changes in emotional speech.

6. Conclusions

This study provides a survey on speech emotion recognition systems from very recent years. The aim of the SER research can be summarised as the search for innovative ways how to appropriately extract emotional context from speech. We can observe a trend in the use of deep convolutional architectures that can learn from spectrogram representations of utterances. Together with recurrent networks, they are considered as a strong base for SER systems. Throughout the years, more complex SER architectures were developed with an emphasis on deriving emotionally salient local and global contexts. As can be inferred from our study, the attention mechanism can improve the performance of the SER systems; however, its benefit is not always evident. Although AM modules have become a natural part of today's SER systems, AM is not an indispensable element for the achievement of high accuracies or even state-of-the-art results.

Author Contributions: Conceptualisation, E.L., R.J. and M.J.; methodology, E.L. and M.J.; writing—original draft preparation, E.L. and M.J.; writing—review and editing, R.J. and M.C.; supervision, R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations and Acronyms

Abbreviation	Meaning
AM	Attention Mechanism
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
CCC	Concordance Correlation Coefficient
CLDNN	Convolutional Long Short-Term Memory Deep Neural Network
CNN	Convolutional Neural Network
DANN	Domain Adversarial Neural Network
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Networks
DSCNN	Deep Stride Convolutional Neural Network
DTPM	Discriminant Temporal Pyramid Matching
ECG	Electro-Cardiogram
EDA	Electro-Dermal Activity
ELM	Extreme Learning Machine
FC	Fully Connected layer
FCN	Fully Convolutional Network
FS	Feature Selection

GAN	Generative Adversarial Network
GeWEC	Geneva Whispered Emotion Corpus
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
HSF	High-Level Statistical Functions
LSTM	Long Short-Term Memory
MFB	Log-Mel Filter-Bank
MFCCs	Mel-Frequency Cepstral Coefficients
MTL	Multitask Learning
NLP	Natural Language Processing
NN	Neural Network
PCA	Principal Component Analysis
ResNet	Residual Neural Network
RFE	Recursive Feature Elimination
RNN	Recurrent Neural Network
SER	Speech Emotion Recognition
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
WoS	Web of Science

References

1. Burkhardt, F.; Ajmera, J.; Englert, R.; Stegmann, J.; Burleson, W. Detecting anger in automated voice portal dialogs. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
2. Hossain, M.S.; Muhammad, G.; Song, B.; Hassan, M.M.; Alelaiwi, A.; Alamri, A. Audio–Visual Emotion-Aware Cloud Gaming Framework. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 2105–2118. [CrossRef]
3. Oh, K.; Lee, D.; Ko, B.; Choi, H. A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation. In Proceedings of the 2017 18th IEEE International Conference on Mobile Data Management (MDM), Daejeon, Korea, 29 May–1 June 2017; pp. 371–375.
4. Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018.
5. Deriche, M.; Abo absa, A.H. A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks. *Arab. J. Sci. Eng.* **2017**, *42*, 5231–5249. [CrossRef]
6. Pravena, D.; Govind, D. Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. *Int. J. Speech Technol.* **2017**, *20*, 787–797. [CrossRef]
7. Bandela, S.R.; Kumar, T.K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2017; pp. 1–5.
8. Koolagudi, S.G.; Murthy, Y.V.S.; Bhaskar, S.P. Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. *Int. J. Speech Technol.* **2018**, *21*, 167–183. [CrossRef]
9. New, T.L.; Foo, S.W.; Silva, L.C.D. Classification of stress in speech using linear and nonlinear features. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 Proceedings (ICASSP '03), Hong Kong, China, 6–10 April 2003; Volume 2, p. II-9.
10. Xia, R.; Liu, Y. A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space. *IEEE Trans. Affect. Comput.* **2017**, *8*, 3–14. [CrossRef]
11. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
12. Cummins, N.; Amiriparian, S.; Hagerer, G.; Batliner, A.; Steidl, S.; Schuller, B.W. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 478–484.
13. Lee, J.; Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015.
14. Aldeneh, Z.; Provost, E.M. Using regional saliency for speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2741–2745.
15. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
16. Xi, Y.; Li, P.; Song, Y.; Jiang, Y.; Dai, L. Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 513–518.

17. Tzinis, E.; Potamianos, A. Segment-based speech emotion recognition using recurrent neural networks. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 190–195.
18. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
19. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
20. Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N. Emotion Identification from Raw Speech Signals Using DNNs. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3097–3101.
21. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473.
22. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1412–1421.
23. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
24. Anagnostopoulos, C.-N.; Iliou, T.; Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.* **2015**, *43*, 155–177. [CrossRef]
25. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [CrossRef]
26. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
27. Sailunaz, K.; Dhaliwal, M.; Rokne, J.; Alhajj, R. Emotion detection from text and speech: A survey. *Soc. Netw. Anal. Min.* **2018**, *8*, 28. [CrossRef]
28. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345. [CrossRef]
29. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
30. Kamińska, D.; Sapiński, T.; Anbarjafari, G. Efficiency of chosen speech descriptors in relation to emotion recognition. *EURASIP J. Audio Speech Music Process.* **2017**, *2017*, 3. [CrossRef]
31. Bakker, I.; van der Voordt, T.; Vink, P.; de Boon, J. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Curr. Psychol.* **2014**, *33*, 405–421. [CrossRef]
32. Truong, K.P.; Van Leeuwen, D.A.; De Jong, F.M. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Commun.* **2012**, *54*, 1049–1063. [CrossRef]
33. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C.A.; Kalliris, G. Speech Emotion Recognition for Performance Interaction. *J. Audio Eng. Soc.* **2018**, *66*, 457–467. [CrossRef]
34. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; Volumn 5, pp. 1517–1520.
35. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE' 05 Audio-Visual Emotion Database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 8.
36. Steidl, S. *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*; Logos-Verlag: Berlin, Germany, 2009.
37. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
38. Lotfian, R.; Busso, C. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Trans. Affect. Comput.* **2019**, *10*, 471–483. [CrossRef]
39. Kamińska, D.; Sapiński, T. Polish Emotional Speech Recognition Based on the Committee of Classifiers. *Przeglad Elektrotechniczny* **2017**, *2016*, 101–106. [CrossRef]
40. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
41. Ringeval, F.; Sonderegger, A.; Sauer, J.S.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013. [CrossRef]
42. Haq, S.; Jackson, P. Speaker-dependent audio-visual emotion recognition. In Proceedings of the AVSP, Norwich, UK, 10–13 September 2009.

43. Ringeval, F.; Schuller, B.; Valstar, M.; Jaiswal, S.; Marchi, E.; Lalanne, D.; Cowie, R.; Pantic, M. AV + EC 2015—the first affect recognition challenge bridging across audio, video, and physiological data. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Brisbane, Australia, 26–30 October 2015.
44. Ntalampiras, S.; Fakotakis, N. Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2012**, *3*, 116–125. [CrossRef]
45. Liu, G.K. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. *arXiv* **2018**, arXiv:1806.09010.
46. Fahad, M.S.; Deepak, A.; Pradhan, G.; Yadav, J. DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features. *Circuits Syst. Signal Process.* **2021**, *40*, 466–489. [CrossRef]
47. Kerkeni, L.; Serrestou, Y.; Raoof, K.; Mbarki, M.; Mahjoub, M.A.; Cleder, C. Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Commun.* **2019**, *114*, 22–35. [CrossRef]
48. YogeshC, K.; Hariharan, M.; Ngadiran, R.; Adom, A.H.; Yaacob, S.; Berkai, C.; Polat, K. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Syst. Appl.* **2017**, *69*, 149–158. [CrossRef]
49. Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; Tan, G.-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* **2018**, *273*, 271–280. [CrossRef]
50. Chen, L.; Mao, X.; Xue, Y.; Cheng, L.L. Speech emotion recognition: Features and classification models. *Digit. Signal Process.* **2012**, *22*, 1154–1160. [CrossRef]
51. Schuller, B.; Steidl, S.; Batliner, A. The Interspeech 2009 Emotion Challenge. In Proceedings of the 10th Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009; pp. 312–315.
52. Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C.; Narayanan, S. The INTERSPEECH 2010 paralinguistic challenge. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 2794–2797.
53. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
54. Weninger, F.; Eyben, F.; Schuller, B.W.; Mortillaro, M.; Scherer, K.R. On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common. *Front. Psychol.* **2013**, *4*, 292. [CrossRef] [PubMed]
55. Huang, C.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 583–588.
56. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
57. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [CrossRef]
58. Hajarolasvadi, N.; Demirel, H. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. *Entropy* **2019**, *21*, 479. [CrossRef]
59. Mustaqeem; Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2020**, *20*, 183. [CrossRef]
60. Kurpukdee, N.; Koriyama, T.; Kobayashi, T.; Kasuriya, S.; Wutiwiwatchai, C.; Lamsrichan, P. Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1744–1749.
61. Lim, W.; Jang, D.; Lee, T. Speech emotion recognition using convolutional and Recurrent Neural Networks. In Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea, 13–16 December 2016; pp. 1–4.
62. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [CrossRef]
63. Luo, D.; Zou, Y.; Huang, D. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 152–156.
64. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017.
65. Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 3683–3687.
66. Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; McInnis, M.; Provost, E.M. Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *arXiv* **2017**, arXiv:1708.07050. Cs.
67. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-End Speech Emotion Recognition Using Deep Neural Networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.

68. AlBadawy, E.A.; Kim, Y. Joint Discrete and Continuous Emotion Prediction Using Ensemble and End-to-End Approaches. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 366–375.
69. Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention Based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1771–1775.
70. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
71. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
72. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. *Electronics* **2020**, *9*, 1725. [CrossRef]
73. Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L.; Schmauch, B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. *arXiv* **2018**, arXiv:1802.05630, 21–25.
74. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *J. Audio Eng. Soc.* **2020**, *68*, 14–24. [CrossRef]
75. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montréal, QC, Canada, 18–22 November 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 2672–2680.
76. Sahu, S.; Gupta, R.; Espy-Wilson, C. On Enhancing Speech Emotion Recognition using Generative Adversarial Networks. *arXiv* **2018**, arXiv:1806.06626. Cs.
77. Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019; pp. 171–175.
78. Fu, C.; Shi, J.; Liu, C.; Ishi, C.T.; Ishiguro, H. AAEC: An Adversarial Autoencoder-based Classifier for Audio Emotion Recognition. In Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20), Seattle, WA, USA, 16 October 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 45–51.
79. Deng, J.; Xu, X.; Zhang, Z.; Frühholz, S.; Schuller, B. Universum Autoencoder-Based Domain Adaptation for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 500–504. [CrossRef]
80. Abdelwahab, M.; Busso, C. Domain Adversarial for Acoustic Emotion Recognition. *arXiv* **2018**, arXiv:1804.07690. Cs Eess. [CrossRef]
81. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *arXiv* **2016**, arXiv:1505.07818. Cs Stat.
82. Zheng, W.; Zheng, W.; Zong, Y. Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition. *Virtual Real. Intell. Hardw.* **2021**, *3*, 65–75. [CrossRef]
83. Noh, K.J.; Jeong, C.Y.; Lim, J.; Chung, S.; Kim, G.; Lim, J.M.; Jeong, H. Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets. *Sensors* **2021**, *21*, 1579. [CrossRef]
84. Lech, M.; Stolar, M.; Best, C.; Bolia, R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci.* **2020**, *2*, 14. [CrossRef]
85. Huang, Z.; Dong, M.; Mao, Q.; Zhan, Y. Speech Emotion Recognition Using CNN. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 801–804.
86. Neumann, M.; Vu, N.T. Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *arXiv* **2017**, arXiv:1706.00612. [CrossRef]
87. Latif, S.; Rana, R.; Qadir, J.; Epps, J. Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study. *arXiv* **2020**, arXiv:1712.08708.
88. Wu, X.; Liu, S.; Cao, Y.; Li, X.; Yu, J.; Dai, D.; Ma, X.; Hu, S.; Wu, Z.; Liu, X.; et al. Speech Emotion Recognition Using Capsule Networks. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6695–6699.
89. Papakostas, M.; Spyrou, E.; Giannakopoulos, T.; Siantikos, G.; Sgouropoulos, D.; Mylonas, P.; Makedon, F. Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition. *Computation* **2017**, *5*, 26. [CrossRef]
90. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; ISBN 978-0-262-03561-3.
91. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
92. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.-R. An Attention Pooling based Representation Learning Method for Speech Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; International Speech Communication Association: Hyderabad, India, 2018.

93. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qata, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1724–1734.
94. Karmakar, P.; Teng, S.W.; Lu, G. Thank you for Attention: A survey on Attention-based Artificial Neural Networks for Automatic Speech Recognition. *arXiv* **2021**, arXiv:2102.07259.
95. Chaudhari, S.; Mithal, V.; Polatkan, G.; Ramanath, R. An attentive survey of attention models. *arXiv* **2019**, arXiv:1904.02874.
96. Lin, Z.; Feng, M.; dos Santos, C.N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-attentive Sentence Embedding. *arXiv* **2017**, arXiv:1703.03130.
97. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 606–615.
98. He, R.; Lee, W.S.; Ng, H.T.; Dahlmeier, D. Effective Attention Modeling for Aspect-Level Sentiment Classification. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Santa Fe, NM, USA, 2018; pp. 1121–1131.
99. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015), Montréal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 577–585.
100. Han, K.; Yu, D.; Tashev, I. Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
101. Huang, C.-W.; Narayanan, S.S. Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 1387–1391.
102. Tao, F.; Liu, G. Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2906–2910.
103. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech Emotion Classification Using Attention-Based LSTM. *IEEEACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
104. Xie, Y.; Liang, R.; Liang, Z.; Zhao, L. Attention-Based Dense LSTM for Speech Emotion Recognition. *IEICE Trans. Inf. Syst.* **2019**, *E102.D*, 1426–1429. [CrossRef]
105. Girdhar, R.; Ramanan, D. Attentional Pooling for Action Recognition. *arXiv* **2017**, arXiv:1711.01467. CsCV.
106. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444. [CrossRef]
107. Zhao, Z.; Zheng, Y.; Zhang, Z.; Wang, H.; Zhao, Y.; Li, C. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 272–276.
108. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019.
109. Dangol, R.; Alsadoon, A.; Prasad, P.W.C.; Seher, I.; Alsadoon, O.H. Speech Emotion Recognition UsingConvolutional Neural Network and Long-Short TermMemory. *Multimed. Tools Appl.* **2020**, *79*, 32917–32934. [CrossRef]
110. Alex, S.B.; Mary, L.; Babu, B.P. Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features. *Circuits Syst. Signal Process.* **2020**, *39*, 5681–5709. [CrossRef]
111. Zheng, C.; Wang, C.; Jia, N. An Ensemble Model for Multi-Level Speech Emotion Recognition. *Appl. Sci.* **2020**, *10*, 205. [CrossRef]
112. Parthasarathy, S.; Busso, C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1103–1107.