technologies

MDPI

Article

Optimizing Speech Emotion Recognition with Machine Learning Based Advanced Audio Cue Analysis

Nuwan Pallewela [1], Damminda Alahakoon [1,*], Achini Adikari [1], John E. Pierce [2] and Miranda L. Rose [2]

[1] Centre for Data Analytics and Cognition, La Trobe Business School, La Trobe University, Melbourne, VIC 3083, Australia; n.pallewela@latrobe.edu.au (N.P.); a.adikari@latrobe.edu.au (A.A.)
[2] Centre of Research Excellence in Aphasia Recovery and Rehabilitation, La Trobe University, Melbourne, VIC 3083, Australia; j.pierce@latrobe.edu.au (J.E.P.); m.rose@latrobe.edu.au (M.L.R.)
* Correspondence: d.alahakoon@latrobe.edu.au

Abstract: In today's fast-paced and interconnected world, where human–computer interaction is an integral component of daily life, the ability to recognize and understand human emotions has emerged as a crucial facet of technological advancement. However, human emotion, a complex interplay of physiological, psychological, and social factors, poses a formidable challenge even for other humans to comprehend accurately. With the emergence of voice assistants and other speech-based applications, it has become essential to improve audio-based emotion expression. However, there is a lack of specificity and agreement in current emotion annotation practice, as evidenced by conflicting labels in many human-annotated emotional datasets for the same speech segments. Previous studies have had to filter out these conflicts and, therefore, a large portion of the collected data has been considered unusable. In this study, we aimed to improve the accuracy of computational prediction of uncertain emotion labels by utilizing high-confidence emotion labelled speech segments from the IEMOCAP emotion dataset. We implemented an audio-based emotion recognition model using bag of audio word encoding (BoAW) to obtain a representation of audio aspects of emotion in speech with state-of-the-art recurrent neural network models. Our approach improved the state-of-the-art audio-based emotion recognition with a 61.09% accuracy rate, an improvement of 1.02% over the BiDialogueRNN model and 1.72% over the EmoCaps multi-modal emotion recognition models. In comparison to human annotation, our approach achieved similar results in identifying positive and negative emotions. Furthermore, it has proven effective in accurately recognizing the sentiment of uncertain emotion segments that were previously considered unusable in other studies. Improvements in audio emotion recognition could have implications in voice-based assistants, healthcare, and other industrial applications that benefit from automated communication.

Keywords: speech emotion recognition; audio; sentiment; machine learning; artificial intelligence

Citation: Pallewela, N.; Alahakoon, D.; Adikari, A.; Pierce, J.E.; Rose, M.L. Optimizing Speech Emotion Recognition with Machine Learning Based Advanced Audio Cue Analysis. *Technologies* **2024**, *12*, 111. https://doi.org/10.3390/technologies12070111

Academic Editor: Juan Gabriel Avina-Cervantes

Received: 30 May 2024
Revised: 30 June 2024
Accepted: 9 July 2024
Published: 11 July 2024

1. Introduction

In the fields of human–computer interaction (HCI) and artificial intelligence (AI), accurate identification of human emotions in conversation remains a crucial challenge. However, it is a key element in facilitating more human-like interactions with computers. Extensive research has been carried out in the realm of detecting emotionally relevant information from diverse sources, presenting both positive and compelling evidence of its impact across various domains, including hospitality [1,2], healthcare [3], human resource management [4,5], and applications of AI. The complexity of human emotions, represented through the expression of modalities such as text, acoustic, and visual cues, challenges AI technologies to accommodate multiple modalities to be effective in real-world contexts. Recent advancements in multi-modal fusion techniques have captured the attention of researchers in the field of emotion recognition, achieving notable success [6,7]. Yet, uni-modal emotion recognition remains essential for specific applications and scenarios, such

as analysing online textual conversations, over-the-phone communication without video, and some health care applications such as health hotlines, support services, and telephonic consultations. As emotions are inherently subjective, discerning the emotions of others poses a significant challenge, even for humans [8]. This challenge is observable even in benchmark datasets [9] for emotion recognition, where different human annotators may assign different emotion labels to the same segments. The incorporation of intermediate feature representation in speech emotion recognition (SER) models helps to understand the possible causes of such confusion and develop solutions to remove or reduce the uncertainty via computational means. In this study, we primarily focused on improving the accuracy of SER using audio modality in conversations and representing audio cues that narrate emotions.

Current state-of-the-art approaches [6,7] leverage deep learning (DL) models to identify emotions from audio features extracted from the speech signals [10,11]. This process generates an internal hidden representation of audio patterns, enabling the prediction of emotions, but results in limiting the interpretability of the prediction. Research conducted on models utilizing audio representations [12,13] has demonstrated the feasibility of achieving competitive accuracy while maintaining interpretability. However, a major problem for identifying decisive audio patterns is the presence of low confidence in the annotated labels in the training datasets. These low-confidence labels can potentially lead to the identification of patterns that lack clear indicators for a specific emotion. To address this limitation, we utilised a segmentation-based approach where data with high-confidence labels were used to identify audio patterns for recognising emotions, while data with low confidence were used only for testing and validation. Subsequently, we utilised the audio patterns from these segments to generate audio representations, following the approach outlined in [12], which utilises the bag-of-audio-words (BoAW) technique. Effectiveness of this representation was tested with well-known conversational emotion recognition models [13,14].

The remainder of the paper is structured as follows: Section 2 delves into related work in the field. Section 3 presents a comprehensive description of the methodology employed. Sections 4 and 5 showcase the experimental results. Finally, Section 6 concludes the paper.

2. Related Work

The process of SER is conventionally segmented into two primary phases: feature extraction and classification. In the feature extraction phase, speech signals are transformed into numerical values through a range of front-end signal processing methods. The resulting feature vectors are concise and aim to encapsulate the critical information within the signal. Following this, in the classification phase, a suitable classifier is chosen based on the specific requirements of the task at hand.

There has been a substantial amount of research dedicated to identifying emotionally relevant information from audio sources. The conventional approach for predicting emotions involves frame-based feature extraction, low-level descriptors (LLDs), followed by the aggregation of utterance-level information, which is then used with a classification or regression technique [15–20]. However, these low-level features as they are may lack the necessary ability to effectively capture and convey subjective emotions in machine learning models. There is a definite requirement for the creation of autonomous algorithms that can learn features on their own and extract advanced emotional characteristics from speech in order to improve the accuracy of recognising emotions. To address this, deep learning algorithm approaches like deep neural networks (DNN) [21], convolutional neural networks (CNN) [22], and deep convolutional neural networks (DCNN) [23] have been used in the recent past. In [14], BoAW, a more interpretable audio feature vector, was employed to extract emotions, with promising results.

In terms of classification, conventional approaches have utilised probabilistic models such as Gaussian mixture models (GMM) [20], hidden Markov models (HMM) [19], and support vector machines (SVM) [24,25]. Over time, research has explored different

types of artificial neural network structures, including CNNs [26], residual neural networks (ResNets) [27], recurrent neural networks (RNNs) [14,28], long short-term memory (LSTM) [15], and gated recurrent units (GRUs) [12,14]. LSTMs and GRUs have emerged as the most advanced approaches for modelling temporal sequences and have been extensively employed in the processing of speech signals. In addition, a multitude of comprehensive designs have been proposed with the objective of concurrently acquiring feature extraction and classification skills [29–31].

Although progress has been made, there is still significant potential for improving the accuracy and interpretability of current methods. This potential for improvement stems from the current limitations in accurately predicting a wider range of basic emotions [32,33] and the comparatively lower performance of acoustic-based emotion detection methods relative to text-based approaches [34].

3. Method

Figure 1 presents the high-level architecture of the proposed approach, with different components for feature extraction and classification. Features are extracted from audio files to represent dialogue, BoAW feature embeddings are generated for audio segments using a codebook trained with utterances of high-confidence emotion labels, and emotions are classified with a Bi-Dialogue RNN that considers speaker, context, and emotion of preceding utterances into the prediction.

Our proposed method of only using the audio features of segments with high-confidence emotion labels in the training phase builds upon the fact that audio emotional prosody clues can be utilised to perceive the emotion of utterances in the absence of other modalities [35], and when the emotion is clear in an utterance, it can be considered more reliable to be used in training a machine learning model.

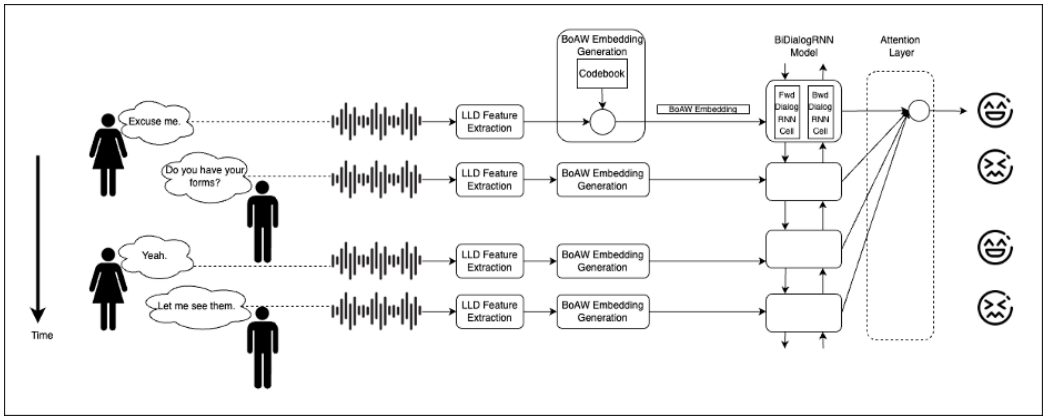

Figure 1. Overall architecture of the proposed model.

3.1. Audio Feature Extraction

Deriving effective feature sets for detecting emotions is crucial in developing an accurate emotion detection model. In the pre-processing stage, audio conversations need to be divided into individual utterances, and the corresponding speakers must be identified. This segmentation process can be achieved through either manual or automatic speaker diarization techniques. Subsequently, the audio files containing utterances are processed, and acoustic LLDs are extracted with various features, including aspects such as root mean square energy, zero-crossing rate, mel-frequency cepstral coefficients (MFCC) 1–14, and spectral flux. Additionally, voicing-related features include fundamental frequency (F0), logHNR, jitter, and shimmer. These features are recognized as significant for conveying emotional content in the human voice [15–20], and we assume that emotion-related features will remain constant during the utterance timeframe. Feature vectors for the audio signal were generated using the openSMILE [36] toolkit. The audio stream is divided into fixed length time slices, and LLDs are generated for each time slice in the audio segment.

Following the feature extraction phase, a comprehensive feature corpus is generated, containing time-varying feature sequences for subsequent feature engineering tasks. In the subsequent step, variable-length time series feature vectors, composed of low-level features from the audio signal, undergo encoding through a BoAW-based mechanism before being input into prediction models.

3.2. Bag of Audio Words (BoAW) Embeddings

The purpose of using BoAW feature extraction methodology is to represent the acoustic speech signal with a discrete set of patterns of the audio waveform [37]. Having this representation would also facilitate the interpretability and the study of causality on machine learning model outcomes. This methodology originates from natural language processing (NLP), wherein documents are represented using bag-of-words (BoW) representations. In a parallel fashion, this technique introduces BoAW as an encoding mechanism for audio data. As shown in Figure 2, the BoAW method keeps a codebook that represents the frequently occurring distinct feature vectors of the audio segments. For each feature vector of a time slice in an utterance, the nearest pattern from the generated codebook is mapped to build the histogram vector for the utterance.

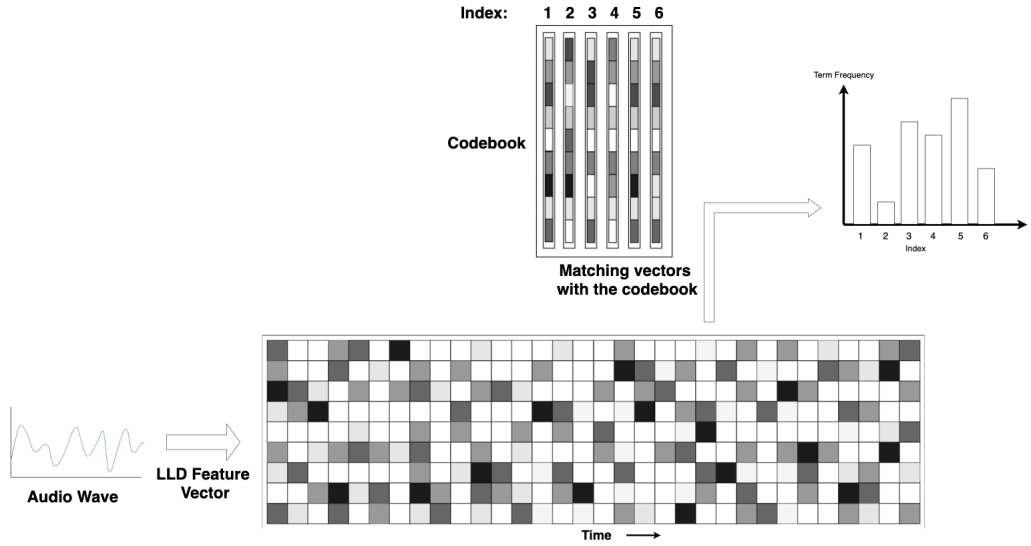

Figure 2. BoAW codebook and feature vector generation.

3.3. Codebook Generation

The initial phase of BoAW involves establishing an indexed codebook, which comprises distinct feature patterns derived from training audio data. These patterns result from a set of low-level features extracted from the audio, and the algorithm achieves this by either randomly selecting patterns or utilising the kmeans++ clustering algorithm. The codebook vectors are iteratively chosen randomly, emphasizing higher dissimilarity among them based on the Euclidean distance measure, and they act as the vocabulary for train and test feature vectors. The term frequency (TF) of the highest matching pattern's index is increased, and this process is repeated for the five highest similar term frequencies. At the end, we obtain feature embeddings of the desired dimension for each utterance, with the TF matrix log-transformed before inputting it into the emotion detection model. As the downstream emotion classification task depends on the effectiveness of this feature encoding, the codebook needs to represent frequent feature vector patterns that would be easy to differentiate for the respective emotions. Therefore, only the audio segments with highly confident labels (majority tagging) are used to generate a BoAW codebook. Each annotator's labelling data are considered rather than just depending on the summarized output. Equation (1) is used to filter the samples with high-confidence emotion tags. Here, E_j represents the jth emotion where $j = 1, 2, 3, ..., n$. (n is number of emotion classes).

$$Confidence = Max[(Count(E_1)/(\Sigma_{j=1}^{n} Count(E_j))), ..., (Count(E_n)/(\Sigma_{j=1}^{n} Count(E_j)))] \quad (1)$$

An audio segment is selected for codebook generation if the confidence is higher than the empirically determined selected threshold value. These segments are filtered only to generate the codebooks. We used all segments of the dataset in the emotion classification task.

3.4. Embedding Generation

The generated codebook is used to determine the embedding vector for the audio segment. Each audio segment contains an array of LLD feature vectors and are compared with vectors in the codebook. When faced with an unseen low-level feature pattern, the codebook pattern with the closest proximity to the given pattern is selected, and the TF of the corresponding index is incremented. After determining all term frequencies at the utterance level (forming a bag/histogram of audio words), a TF matrix is generated. The entries in this matrix serve as fixed-size encodings, which can be fed into a sequence-to-sequence modeling approach for predicting latent emotion categories. Similar to the standard NLP BoW approach in document classification, TF and inverse document frequency (IDF) can also be used. We experimented with TF-IDF transformations, bi-gram, and tri-gram with different combinations to explore the impact on prediction accuracy. As the accuracy of these experiments did not improve with the use of above feature variations, uni-gram BoAW features were used for the further experiments with the classification models.

3.5. Bi-Dialogue RNN Model

The BiDialogueRNN [14] is a recently developed model designed for retrieving emotion predictions from conversations by incorporating contextual information from both preceding and succeeding utterances. In this model, there are three key factors influencing the understanding of emotion variations in conversations, each at a different scale. These factors include the global context of the conversation, the speaker's state, and the emotion of the utterance, each of which is modelled using distinct RNNs, as shown in Figure 3: Global State GRU (GRU_g), Party(Speaker) State GRU (GRU_p), and Emotion State GRU (GRU_e), respectively. As in [38], GRU cells were utilised to capture the long-term dependencies within sequences, enabling the identification of conversation dynamics while preserving inter-party relationships by maintaining the conversation context. In this context, each GRU cell computes a hidden state (h_t), defined as $h_t = GRU(h_{t-1}, x_t)$, which represents the current input (x_t) and the previous GRU state (h_{t-1}). The state h_t serves as the current GRU output.

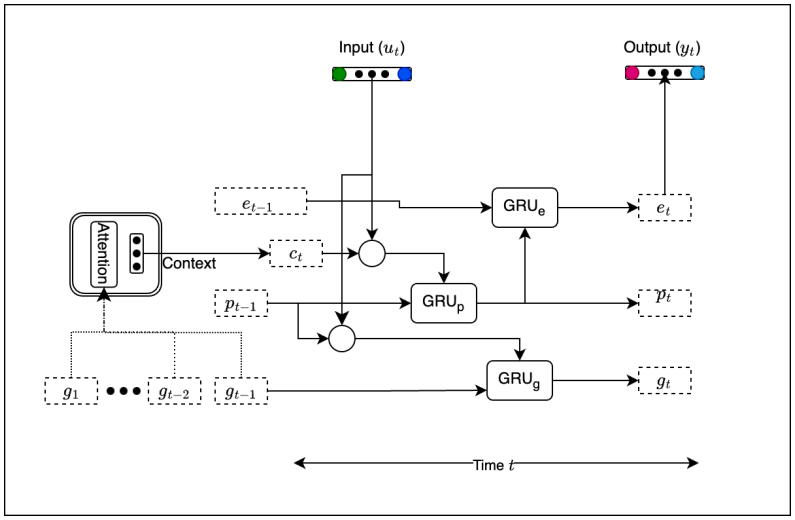

Figure 3. BiDialogueRNN architecture with attention.

3.6. Global State GRU

The context of the utterance is represented as a global state that is shared among and co-constructed by the parties. When a speaker is engaged in the conversation, at each turn, the context of the conversation must be updated. As the most recent context of the conversation has a relatively high impact on the emotional state of the speaker, previous states information is preserved by encoding the utterance (u_t) and speaker's previous state (p_{t-1}), which are concatenated and fed to the GRU_g at each time step, as shown in Equation (2).

$$g_t = GRU(g_{t-1}, u_t \oplus p_{t-1}) \tag{2}$$

3.7. Party State GRU

Each participant's state evolves throughout the conversation, and these states contain valuable information that can be utilised to gauge the speaker's emotions. The expression GRU_p introduces a computational framework aimed at capturing the speaker's evolving state during the conversation. This is achieved by continually updating the respective current states (p_t) of the speakers whenever an utterance (u_t) is made, as shown in Equation (3); c_t, as shown in Equation (4), represents the context of the conversation evaluated with attention mechanism, which is described in Section 3.9.

$$p_t = GRU(p_{t-1}, u_t \oplus c_t) \tag{3}$$

$$c_t = \alpha[g_1, g_2, \ldots, g_t] \tag{4}$$

3.8. Emotion State GRU

To predict the emotional state, emotionally relevant features extracted from the participants' states (p_t) are fed into the emotion state GRU, along with the speaker's previous emotional state (e_{t-1}) (Equation (5)). The bidirectional emotion state RNN conducts both forward and backward passes to generate emotional representations of the speakers across the conversation. These forward and backward emotion states are concatenated, and an attention-based mechanism is employed separately to highlight emotionally significant segments, thereby facilitating a more intuitive emotion classification process.

$$e_t = GRU(e_{t-1}, p_t) \tag{5}$$

3.9. Attention Mechanism

Given the influence of past conversation states on the speaker, it is crucial to focus on emotionally relevant segments to predict the speaker's next state. This model employs an attention mechanism proposed in [12], which involves considering the transposed vector of the current global state, as depicted in Equation (6). The findings suggest that this attention mechanism is more effective, as it contains dense information compared to the high-dimensional (2000-length) sparse utterance encodings obtained directly from BoAW embedding. In Equation (6), g_t is the global state, and W_α is the weight vector for the softmax layer.

$$\alpha = \text{softmax}(g_t^T W_\alpha [g_1, g_2, \ldots, g_t]) \tag{6}$$

4. Experiments

This section presents the dataset and experimental setup used in the study. In each experiment, we evaluated our models on the IEMOCAP dataset [9].

4.1. Interactive Emotional Dyadic Motion Capture (IEMOCAP) Dataset

The IEMOCAP dataset is published by Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). It contains conversations from ten actors in five dyadic sessions during scripted and spontaneous communication scenarios. This dataset contains approximately 12 h of data with categorical (happiness, sadness,

frustrated, angry, exited, neutral, fear, disgust, surprise, and other) and attribute (valance, activation, and dominance) emotion annotations for segments of speech recordings. Each segment is labelled by three annotators, and the majority emotion is taken. Each segment may be annotated with one or multiple emotion tags per annotator, depending on personal judgment. Figure 4 shows the distribution of emotion labels for segments. Frustrated, angry, sadness, neutral, excited, and happiness are the most prominent emotions labelled in the dataset. Label 'xxx' is given if the tags given by the three annotators are not agreed to assume a clear emotion label. Most of the past studies have left out this confusing label tag and other emotion tags due to not having enough data points to train and evaluate machine learning models.

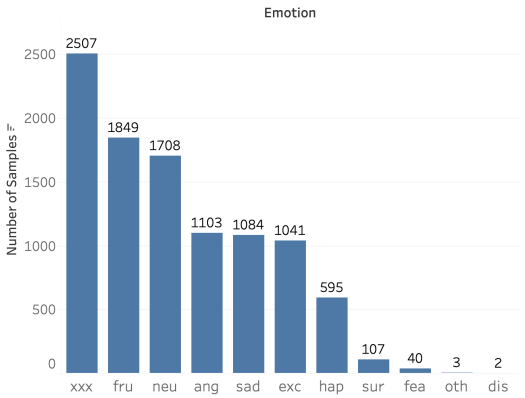

Figure 4. Emotion label distribution in the IEMOCAP dataset.

4.2. Experimental Setup

The proposed model and explorations were conducted on the IEMOCAP dataset. Figure 4 shows the distribution of the emotion labels and train/test samples. We carried out the experiments for six emotion categories by combining similar emotions together. We also conducted experiments on various confidence thresholds (0.6, 0.65, 0.7, 0.75, and 0.8) as calculated with Equation (1). Table 1 shows the number of samples in the training data for each emotion category and each confidence level. It is observed that very few samples exceed 0.8 confidence for the happy emotion category. Experiments were conducted by splitting the dataset into five folds by the session number; one fold is configured as the test set and the other four folds are used as the training set. This process is repeated for every fold, and the performance metrics are collected and averaged to obtain the model's generalization performance.

Table 1. Number of samples in training data for each confidence level.

Confidence	Happy	Excited	Neutral	Sad	Angry	Frustrated
0.6	361	619	1164	668	747	1238
0.65	354	583	1137	624	699	1187
0.7	85	258	488	366	300	498
0.75	85	258	488	366	299	497
0.8	34	179	259	239	273	337
0.85–1.0	34	150	254	224	272	313

We directly used the audio files of the given utterances as input to our proposed pipeline. We converted them to the required audio format (16-bit PCM WAV files) using ffmpeg, and then, to extract features, we utilised the openSMILE configuration file ComParE_2016, extracting 130 LLDs for each 25-millisecond speech frame at a frame rate of 10ms, assuming stationary audio-related properties. The openXBOW toolkit, a Java

application, was employed to create a BoAW representation from the acoustic LLDs. We split the LLD feature vectors of the train partition into two sets of 65 features: 55 spectral features, 6 voice related low-level descriptors, 4 energy-related features, and derivatives of first 65 features, forming two codebooks with 1000 generated indices each, resulting in 2000 indices in total, as used in [12].

Inside the RNN, the dimensions consist of the input embedding $dim(u_t) = 2000$, the global state of the conversation $dim(g_t) = 500$, the speaker's state $dim(p_t) = 500$, and the emotional state of the speaker $dim(e_t) = 500$. The emotion detection model employs negative log-likelihood loss with L2 regularization, assigning weights to each class to address data imbalance. Optimisation is carried out using the Adam optimiser with a learning rate of 0.0001 and weight decay of 0.0001. The model is trained for 40 epochs with a batch size of 30 dialogue sequences.

4.3. Baseline Models

Models that have used only audio input in experiments are considered as baseline models to compare the results with our model, as shown in Table 2.

As outlined in the previous section, our methodology involves employing a BoAW representation technique with two codebooks derived exclusively from audio segments accompanied by high-confidence emotion labels (Con-BoAW). This approach offers enhanced interpretability when contrasted with DL feature extraction methods employed in BC-LSTM, MMGCN, and EmoCaps. Moreover, our strategy endeavours to achieve a more refined codebook that captures emotion-related audio features more effectively than the codebooks in [12].

Subsequently, we assessed the implications of integrating Con-BoAW as the audio features into both the BiDialogueRNN and EmoCaps multi-model SER models. Consistent with the experiments detailed in [14], the conventional audio vector was substituted with Con-BoAW, and concatenation was employed to merge the text and visual modality data.

As shown in Table 3, we used identical configurations for text and visual features as outlined in [39] to facilitate a comparison of the effects of our audio feature set. To ensure alignment with the audio feature vector dimension employed in [39], a linear neural network was utilised to reduce the dimensionality of the audio feature vector to 100. Subsequently, we conducted tests with dimensions of 128, 256, 512, 1024, and 2000 to assess the optimal performance configuration.

Table 2. Baseline models.

Model	Description
BC-LSTM	BC-LSTM [40] uses Bi-directional LSTM structure to encode contextual semantic information; it does not recognize the speaker relationship.
MMGCN	MMGCN [41] uses GCN network to obtain contextual information, which not only makes use of multimodal dependencies effectively, but also leverages speaker information.
EmoCaps	EmoCaps [39] is a state-of-the-art multi-modal emotion capturing model that contains an Emoformer architecture to extract emotion vectors from multimodal features and splice them with sentence vectors to form an emotion capsule.
openSMILE + BiDialogueRNN	We established a baseline emotion detection pipeline by extracting 6373 features from each utterance using the IS13 ComParE configuration script from openSMILE, which we then inputted into the BiDialogueRNN emotion detection model.
BoAW + BiDialogueRNN	As in [12], BoAW feature vector was used for the audio features with a BiDialogueRNN model, which achieved good accuracy for SER in audio modality.

Table 3. EmoCaps multi-modal SER model parameter settings.

Parameter Setting	Value
Epochs	80
Learning rate	0.0001
Drop out	0.1
Batch size	30
Dim-T	100
Dim-A	100
Dim-V	256

4.4. Evaluation Metrics

Precision measures the accuracy of positive predictions, representing the ratio of true positive predictions to the total number of predicted positives. It is particularly useful when the cost of false positives is high. Recall, on the other hand, gauges the model's ability to capture all relevant instances of the positive class, indicating the ratio of true positives to the total number of actual positives. It is crucial when missing positive cases is costly. The F1 score, a harmonic mean of precision and recall, provides a balanced assessment of a model's performance, offering a single metric that considers both false positives and false negatives. It is especially valuable in scenarios where achieving a balance between precision and recall is essential for a successful classification outcome.

4.5. Label Adjustment for Audio Segments with Confused Emotions

The target emotion of an audio segment in the dataset is determined considering the labels given by the majority of the annotators who are observing the conversation. However, this could be inaccurate, as a person can misinterpret the emotion of someone else, and it is subjective as well. For example, anger and frustration can be seen as overlapping emotions on certain occasions: 55.77% segments labelled as anger had tags as frustration. A speaker could be both angry and frustrated, which led some annotators to label the emotion as angry while others labelled it as frustration. Both are acceptable labels for the audio segment. Therefore, in order to include the audio segments with confused labels, evaluation criteria were adjusted to accept the prediction as successful if even one annotator labelled the audio segment with the predicted emotion, as shown in Equation (7).

$$\text{Success}(E_p, E_l) = (\Sigma E_p)/(\Sigma_{l \epsilon L} E_l) > 0 \tag{7}$$

5. Results and Analysis

In this section, we present the results of our audio feature performance with different experiments carried out with the IEMOCAP dataset, and then we compare with both audio-only and multi-modal state-of-the-art SER models. Precision, recall, and F1 score are key metrics used to evaluate the performance of classification models, which are used to classify input segments into six basic emotions. (Random classification accuracy for six emotion classes is 16.67%.)

5.1. Experiment Results Based on the IEMOCAP Dataset

Our method achieved a weighted accuracy of 61.09% in classifying the six basic emotions using audio data from the IEMOCAP dataset. Table 4 illustrates the accuracy results for Con-BoAW codebooks with varying confidence thresholds utilised as inputs for the BiDialogueRNN audio SER classification model. It is evident that the accuracy values demonstrate an increase as the confidence threshold rises from 0.6 to 0.75. However, the accuracy begins to decrease beyond 0.75, likely due to the reduced number of samples in the training set beyond this threshold. The highest accuracy was attained with a confidence threshold of 0.75. These results show the significant impact of confidence intervals on the accuracy of the final speech emotion recognition (SER) model. Having more distinguishable

segments for emotion in the training data, which are used to generate the audio representation as BoAW feature vectors, enhances the performance of downstream classification models. Ideally, it would be preferable to label emotions solely based on audio cues without relying on visual information. The experiment results with confused emotion tags show 78.81% accuracy with the updated evaluation criteria, which suggests our model is able to determine the emotions expressed in utterances with low-confident emotion tags even it is trained only using high-confident segments.

Table 4. Accuracy (%) of the audio SER model with codebooks generated with different confidence threshold values. * BiDialogueRNN without confused labels. ** BiDialogueRNN with confused labels and adjusted evaluation criteria.

Confidence	Without Confused Labels *		Including Confused Labels **	
Threshold	Weighted F1	F1	Weighted F1	F1
0.6	54.13	54.07	73.36	73.18
0.65	55.02	54.32	74.51	74.36
0.7	59.69	59.63	76.24	75.99
0.75	61.09	60.80	78.81	78.58
0.8–1.0	59.15	59.94	74.62	75.12

Figure 5 shows the confusion matrix output of the emotion prediction for the 0.75 confidence interval. The confusion matrices provide insights into the performance (exited—74.2%, happy—18.7%, neutral—70.4%, sad—57.2%, angry—67.5%, and frustrated—59.5%) of our audio feature pipeline in emotion detection. They reveal that misclassification often occurs between similar emotion categories. For instance, frustrated emotions are sometimes misinterpreted as angry due to their close resemblance in nature. Moreover, happy emotions may be misclassified into other emotions, possibly because the dataset lacks high-confidence happy emotion labels, as shown in Table 1. Additionally, other emotion categories are frequently misidentified and labelled as neutral. Study [8] shows the performance values of human evaluators on a SER task for five emotions (happy, normal, sad, angry, and afraid) with accuracy scores of happy—61.4%, normal—66.3%, sad—68.3%, angry—72.2%, and afraid—49.5%, which are also similar to our model's values (apart from the happy emotion), even with six categories.

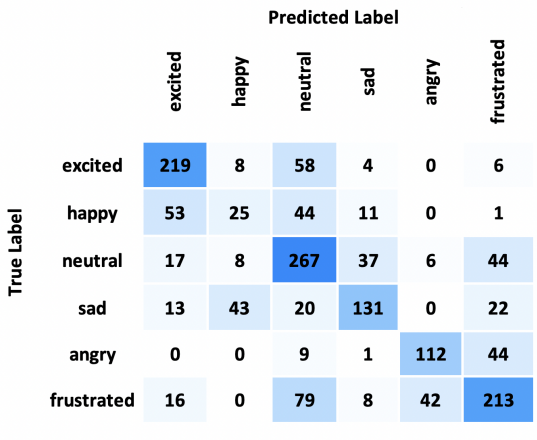

Figure 5. Confusion matrix output for Con-BoAW with 0.75 confidence threshold.

Figures 6 and 7 show the 50th percentile values of the first codebook LLD vector's Uniform Manifold Approximation and Projection (UMAP) projection for different speakers. Figure 6 shows the happy emotion and Figure 7 shows the frustration emotion.

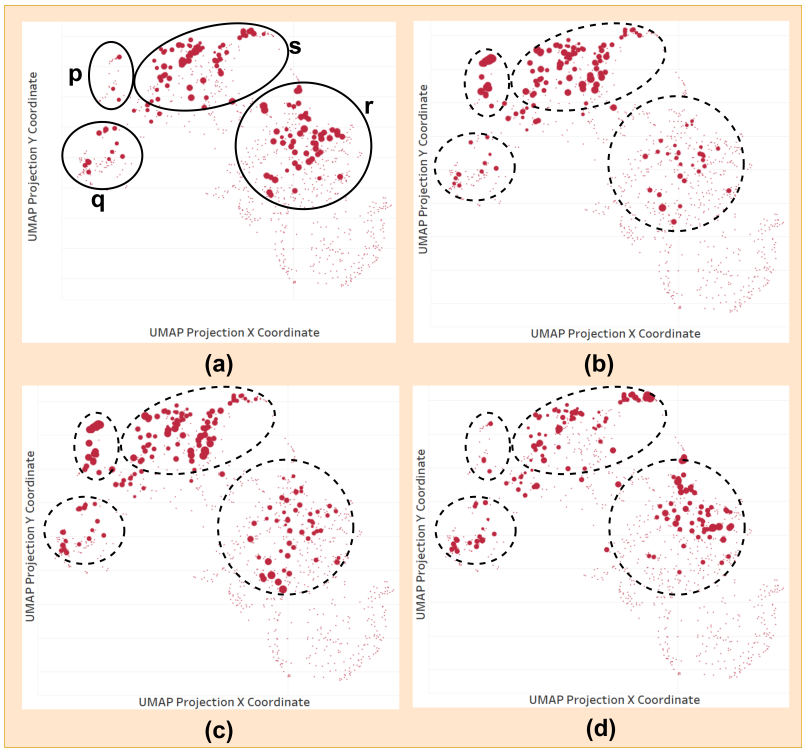

Figure 6. Happy emotion's 50th percentile values of the first codebook LLD vector's UMAP projection for different speakers: (**a**) session 2 male speaker, (**b**) session 3 male speaker, (**c**) session 4 female speaker, (**d**) session 5 male speaker.

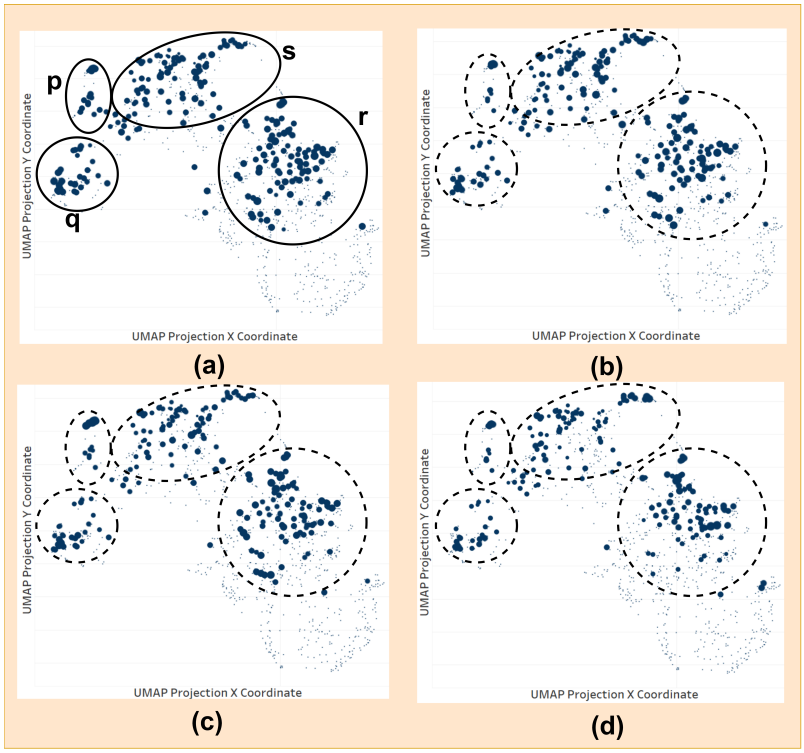

Figure 7. Frustration emotion's 50th percentile values of the first codebook LLD vector's UMAP projection for different speakers: (**a**) session 2 male speaker, (**b**) session 3 male speaker, (**c**) session 4 female speaker, (**d**) session 5 male speaker.

It is observable in Figure 6 that cluster p is significantly different in Figure 6a and Figure 6d compared to Figure 6b and Figure 6c. Additionally, cluster r is slightly different

in each chart, where Figure 6a shows significant values overall, Figure 6b shows low values, and Figure 6c and Figure 6d show skewed distribution. However, codebook vectors in Figure 7 sub-charts show similar patterns overall in each p, q, r, and s clusters compared to Figure 6. It suggests that triggered codebook vectors are slightly different for different people in happy emotions compared to frustration, which might be the reason for the low accuracy shown in the confusion matrix. Having the intermediate layer of audio vector representation support allows us to analyse the results and improve accordingly.

5.2. Study with Confused Emotion Segments

As shown in Figure 4, a large proportion of the segments are labelled as 'xxx' (24.97% of the samples in the dataset), which suggests human evaluators did not agree on the emotion label one quarter of the time. As shown in Figure 8, our model matched the individual evaluator's labelling on average better than that of human evaluators.

	E1	E2	E3	E4	E5	E6	MODEL
E1	N/A	38.22	34.45	33.69	24.13	37	50.33
E2	37.33	N/A	44.01	44.13	N/D	45.88	53.89
E3	45.55	46.15	N/A	53.22	N/D	N/D	57.75
E4	44.4	39.66	47.64	N/A	55.78	N/D	52.82
E5	31.19	N/D	N/D	54.11	N/A	N/D	42.8
E6	47.28	50.41	N/D	N/D	N/D	N/A	49.12
AVG	41.15	43.61	42.03	46.29	39.96	41.44	51.12

Figure 8. Evaluator and model labelling comparison (weighted average F1 score in percentage, N/A—not applicable, N/D—no data points).

Further analysis revealed that human evaluators have superior sentiment-wise agreement compared to emotion level, as shown in Figure 9. This suggests that the confusion is mainly present between positive emotions (happy, excited), negative emotions (sad, angry, frustrated), and a neutral state. Therefore, these data indicate that human judgment agrees on sentiments but shows confusion and disagreement with specific emotions. Emotions could be considered as at a more detailed level of granularity, where the observer has to distinguish between a number of distinct emotions compared to only three sentiments: positive, negative, and neutral. As different emotions can be expressed together, interpreting the emotion state of the speaker could be subjective to the observer, which results in lesser agreement between different observers compared to sentiment. As illustrated in Figure 10, our model also shows similar behaviour of matching the emotion prediction of the utterance sentiment-wise. Based on this analysis, it is evident that our model demonstrates and matches the basic behaviour of a human evaluator.

	excited	happy	neutral	sad	angry	frustrated
excited	4137	2529	1586	151	114	373
happy	2754	1635	1351	184	14	214
neutral	1699	1319	5975	1321	561	3176
sad	145	169	1253	4337	181	1333
angry	76	15	532	192	3622	3288
frustrated	378	226	3137	1440	3307	7038

Figure 9. Evaluator emotion confusion matrix.

Figure 10. Evaluators vs. model emotion confusion matrix.

5.3. Experiments with Multi-Modal SER Models

BiDialogueRNN and EmoCaps multi-modal SER models with Con-BoAW also improved the accuracy, as shown in Tables 5 and 6. (These accuracy values were calculated by running in our servers.)

Table 5. Multi-modal BiDialogueRNN model with BoAW and Con-BoAW audio feature vecotors.

Model	F1 Accuracy %	F1 Weighted Average %
BoAW + BiDialogueRNN	60.52	60.80
Con-BoAW + BiDialogueRNN	**62.24**	**62.62**

Table 6. Multi-modal Con-BoAW + EmoCaps model with different audio feature vector dimension transformations.

Audio Feature Vector Dimension Transformation	F1 Accuracy % (EmoCaps)	F1 Weighted Average % (EmoCaps)
2000 → 100	72.09	72.14
2000 → 128	72.26	72.23
2000 → 256	72.79	72.76
2000 → 512	**73.49**	**73.51**
2000 → 1024	73.43	73.47
2000	70.92	70.94

Table 7 presents a comparison of results with baseline models in both audio-only and multi-modal scenarios. Accuracy values of all baseline models listed in Table 2 were published with one-fold train/test data split criteria, where session 1–4 data were considered as the training set and session 5 data were taken as the test set. In order to compare with baseline models, we evaluated the accuracy values in the same train/test split and listed with the tag "(one-fold)" in Table 7. The Con-BoAW + BiDialogueRNN + Attn model demonstrates the highest accuracy of 61.09% among audio models. In the multi-modal SER model scenario, the Con-BoAW + EmoCaps combination achieves the highest accuracy of 73.49%.

Table 7. Audio only and multi-modal performance comparison with baseline models. The * sign indicates models for which the accuracy values were calculated by running the model in our server and configurations, as they were not available with the published results.

	Accuracy % (Audio)	**Accuracy % (Multi-Modal)**
EmoCaps (one-fold)	33.06	71.77
BoAW + Classification (one-fold)	43.23	-
Bc-LSTM (one-fold)	47.4	57.5
MMGCN (one-fold)	54.66	66.22
BiDialogueRNN (one-fold)	58.41 *	60.28
BiDialogueRNN+Attn (one-fold)	59.32 *	62.9
BoAW+BiDialogueRNN+Attn (one-fold)	60.87	60.80 *
Con-BoAW +BiDialogueRNN+Attn (one-fold)	**61.09**	62.62
Con-BoAW + EmoCaps (one-fold)	50.42	**73.49**
BiDialogueRNN+Attn (5-fold)	56.44 $\pm$ 2.04 *	-
BoAW+BiDialogueRNN+Attn (5-fold)	57.27 $\pm$ 3.1 *	-
Con-BoAW +BiDialogueRNN+Attn (5-fold)	**59.02** $\pm$ **1.78**	-

To further assess how our model generalizes to new, unseen data, we performed five-fold cross-validation considering data in each session as the test set for each fold. Accuracy values listed in Table 7 with "(5-fold)" show that our model outperformed both BiDialogueRNN + Attn and BoAW + BiDialogueRNN + Attn audio models by +2.58% and +1.75%, respectively.

5.4. Ablation Study

We carried out experiments utilizing Con-BoAW embeddings, BoAW embeddings, and openSMILE low-level descriptors (LLDs) as audio features to assess the efficacy of our proposed Con-BoAW features. We employed two RNN-based classification models for each type of audio feature. As depicted in Table 8, the accuracy exhibits a consistent increase from LLDs to BoAW and from BoAW to Con-BoAW across all models.

Table 8. F1 accuracy values for different audio feature vectors and models.

	BiDialogueRNN	**EmoCaps**
LLDs	52.69	33.0
BoAW	60.87	44.34
Con-BoAW	**61.09**	**50.42**

6. Discussion and Conclusions

Overall, the audio model with Con-BoAW + Bi-DialogueRNN achieved an accuracy of 61.09%, which stands as the highest reported in the literature. The performance of multi-modal SER models further suggests that our Con-BoAW approach is more effective compared to basic BoAW, LLDs, or deep learning feature extraction mechanisms reported in the literature. Specifically, the accuracy of the EmoCaps model improved to 73.49%, representing a 1.72% increase, and the Bi-DialogueRNN model achieved an accuracy of 62.62%. Our model showed an accuracy of 78.81% with the adjusted evaluation criteria for utterances with confused emotion labels, which were not investigated in previous studies in the literature. Furthermore, five-fold cross-validation results showed our model's generalization capability and achieved 59.02% accuracy with 1.75% improvement compared to the BoAW + Bi-DialogueRNN model. These experimental results validate that audio features alone contain clues about emotions, and speech segments tagged with high-confident emotion labels contain more relevant patterns, which would help to predict

the expressed emotion more accurately even for segments with low-confident emotion expressions. Furthermore, the proposed audio model performed similarly to the human annotators, who utilised multi-modal content for emotion annotation. This indicates that, despite other modalities, emotion cues are prominently embedded in audio features as well. Causes of misclassification could be identified with the intermediate representation of audio features with Con-BoAW vectors, which would help in further improving the accuracy in the next steps. Automated SER systems can handle larger volumes of data in near real-time and provide a quantitative measurement. Since humans tend to be subjective and qualitative in emotion interpretation, computer applications would be able to integrate SER models more effectively to facilitate human-like interactions.

A limitation identified during the research is the absence of a balanced, large corpus of annotated speech datasets with comprehensive details to precisely determine which segments contain high-confidence emotion labels. BoAW embeddings generated from a substantial training dataset could potentially offer greater reliability due to their high dimensionality and sparsity. Another limiting factor in implementing this research in real-world scenarios is the need for high-accuracy, real-time speaker diarization solutions. With the availability of a reliable real-time speaker diarization solution, a fully automated and dependable SER pipeline could be deployed based on the research contributions.

The proposed research holds promise for various practical innovations driven by new technical advancements. If machines become capable of effectively identifying human emotions, numerous systems could utilise this data to communicate more naturally, offering customized services for special cases such as individuals with aphasia where language clarity might hinder consistent emotion detection and visual cues may not always be accessible. Systems that are reliable in functioning with missing modalities are crucial for real-world applications. In this context, empowering machines to comprehend human emotions through audio conversations represents a significant advancement in the field of human–computer interaction, leveraging big audio data to enhance user experiences.

Author Contributions: All authors made equal contributions to: conceptualization, methodology and formal analysis. N.P. conducted data processing, experimental design and writing—original draft preparation. D.A., A.A., J.E.P., and M.L.R. carried out writing—review and editing and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Health and Medical Research Council Ideas Grant [APP2003210].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The IEMOCAP dataset can be downloaded from (accessed on 26 December 2022) https://sail.usc.edu/iemocap/. Source code can be found in https://github.com/nwnpallewela/conf-boaw.

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

Abbreviations

The following abbreviations are used in this manuscript:

BoAW	Bag of Audio Words
RNN	Recurrent Neural Network
HCI	Human Computer Interaction
AI	Artificial Inteligence
SER	Speech Emotion Recognition
DL	Deep Learning
LLD	Low-Level Descriptors
MFCC	Mel-Frequency Cepstral Coefficients
LPCC	Linear Prediction Cepstral Coefficients

F0	Fundamental Frequency
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DCNN	Deep Convolutional Neural Network
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
SVM	Support Vector Machines
LSTM	Long Short Term Memory
GRU	Gated Recurrent Units
NLP	Natural Languge Processing
BoW	Bag of Words
TF	Term Frequency
IDF	Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

References

1. Abeysinghe, S.; Manchanayake, I.; Samarajeewa, C.; Rathnayaka, P.; Walpola, M.; Nawaratne, R.; Bandaragoda, T.; Alahakoon, D. Enhancing Decision Making Capacity in Tourism Domain Using Social Media Analytics. In Proceedings of the 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 26–29 September 2018; Volume 9, pp. 369–375.
2. Adikari, A.; Burnett, D.; Sedera, D.; De Silva, D.; Alahakoon, D. Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100022. [CrossRef]
3. Adikari, A.; Nawaratne, R.; De Silva, D.; Ranasinghe, S.; Alahakoon, O.; Alahakoon, D. Emotions of COVID-19: Content Analysis of Self-Reported Information Using Artificial Intelligence. *J. Med. Internet Res.* **2021**, *23*, e27341. [CrossRef]
4. Adikari, A.; Alahakoon, D. Understanding citizens' emotional pulse in a smart city using artificial intelligence. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2743–2751. [CrossRef]
5. Alahakoon, D.; Nawaratne, R.; Xu, Y.; De Silva, D.; Sivarajah, U.; Gupta, B. Self-Building artificial intelligence and machine learning to empower big data analytics in smart cities. *Inf. Syst. Front.* **2020**, *25*, 221–240. [CrossRef]
6. Xu, M.; Zhang, F.; Zhang, W. Head Fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access* **2021**, *9*, 74539–74549. [CrossRef]
7. Nediyanchath, A.; Paramasivam, P.; Yenigalla, P. Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7179–7183. [CrossRef]
8. Petrushin, V.A. Emotion recognition in speech signal: Experimental study, development, and application. In Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China, 16–20 October 2000. [CrossRef]
9. 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*, 335–359. [CrossRef]
10. Singh, R. *Profiling Humans from Their Voice: Profiling and Its Facets*; Springer: Singapore, 2019; pp. 3–26. [CrossRef]
11. 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. [CrossRef]
12. Chamishka, S.; Madhavi, I.; Nawaratne, R.; Alahakoon, D.; De Silva, D.; Chilamkurti, N.; Nanayakkara, V. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. *Multimed. Tools Appl.* **2022**, *81*, 35173–35194. [CrossRef]
13. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
14. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6818–6825. [CrossRef]
15. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093. [CrossRef]
16. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
17. Anagnostopoulos, C.N.; Iliou, T.; Giannoukos, I. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.* **2012**, *43*, 155–177. [CrossRef]
18. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; Andre, 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]
19. 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]

20. 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. [CrossRef]
21. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef]
22. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
23. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Mohamed, A.-R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
24. 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]
25. Spyrou, E.; Nikopoulou, R.; Vernikos, I.; Mylonas, P. Emotion Recognition from Speech Using the Bag-of-Visual Words on Audio Segment Spectrograms. *Technologies* **2019**, *7*, 20. [CrossRef]
26. 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. [CrossRef]
27. 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. [CrossRef]
28. 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. [CrossRef]
29. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
30. 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. [CrossRef]
31. 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. [CrossRef]
32. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. *arXiv* **2019**. [CrossRef]
33. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.. [CrossRef]
34. Ruusuvuori, J. Emotion, affect and conversation. In *The Handbook of Conversation Analysis*; Blackwell Publishing Ltd.: Hoboken, NJ, USA, 2013; pp. 330–349.
35. Adolphs, R.; Damasio, H.; Tranel, D. Neural systems for recognition of emotional prosody: A 3-D lesion study. *Emotion* **2002**, *2*, 23–51. [CrossRef] [PubMed]
36. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010. [CrossRef]
37. Pancoast, S.; Akbacak, M. Bag-of-audio-words approach for multimedia event classification. In Proceedings of the Interspeech, Portland, OR, USA, 9–13 September 2012; pp. 2105–2108.
38. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**. [CrossRef]
39. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. *arXiv* **2022**, arXiv:2203.13504.
40. Asokan, A.R.; Kumar, N.; Ragam, A.V.; Sharath, S.S. Interpretability for Multimodal Emotion Recognition Using Concept Activation Vectors. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022. [CrossRef]
41. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. *arXiv* **2021**. [CrossRef]