






Article

# Improving the Accuracy of Automatic Facial Expression Recognition in Speaking Subjects with Deep Learning

Sathya Bursic <sup>1,2,\*</sup> , Giuseppe Boccignone <sup>1</sup> , Alfio Ferrara <sup>2</sup> , Alessandro D'Amelio <sup>1</sup>   
and Raffaella Lanzarotti <sup>1</sup> 

<sup>1</sup> PHuSe Lab, Department of Computer Science, University of Milan, Via Giovanni Celoria 18, 20133 Milano, Italy; giuseppe.boccignone@unimi.it (G.B.); alessandro.damelio@unimi.it (A.D.); raffaella.lanzarotti@unimi.it (R.L.)

<sup>2</sup> ISLab, Department of Computer Science, University of Milan, Via Giovanni Celoria 18, 20133 Milano; alfio.ferrara@unimi.it

\* Correspondence: sathya.bursic@unimi.it

Received: 12 May 2020; Accepted: 3 June 2020; Published: 9 June 2020



**Abstract:** When automatic facial expression recognition is applied to video sequences of speaking subjects, the recognition accuracy has been noted to be lower than with video sequences of still subjects. This effect known as the *speaking effect* arises during spontaneous conversations, and along with the affective expressions the speech articulation process influences facial configurations. In this work we question whether, aside from facial features, other cues relating to the articulation process would increase emotion recognition accuracy when added in input to a deep neural network model. We develop two neural networks that classify facial expressions in speaking subjects from the RAVDESS dataset, a spatio-temporal CNN and a GRU cell RNN. They are first trained on facial features only, and afterwards both on facial features and articulation related cues extracted from a model trained for lip reading, while varying the number of consecutive frames provided in input as well. We show that using DNNs the addition of features related to articulation increases classification accuracy up to 12%, the increase being greater with more consecutive frames provided in input to the model.

**Keywords:** facial expression recognition; speaking effect; emotion recognition; affective computing; deep learning

## 1. Introduction

In what Feldman Barrett et al. [1] name the “*common view*”, certain emotion categories are reliably signaled or revealed by specific configurations of facial-muscle movements. The latter are referred to as emotional expressions and facial expressions. We will not argue here on the plausibility of this widespread assumption, which pervades both research and the development of commercial applications; such ongoing and passionate debate is out of the scope of this paper (see [1] for details).

Yet, whatever the agreement with the pronouncements of this common view, it is surprising to notice how a straightforward fact has been hitherto overlooked to some extent: the facial apparatus, aside from being used in communicating affect, is much the same involved in articulation when subjects are speaking. This is allegedly occurring in spontaneous interactions and as a matter of fact, even a neutral expression of a speaking subject may be confused with emotional expressions [2]. This phenomenon was first described by Picard in [3] who coined the term “*speaking effect*”.

As an illustration of the argument, Figure 1 depicts an expression of *surprise* from the CK+ [4] dataset and a frame from the RAVDESS [5] dataset of a speaking subject expressing the *neutral*

expression and pronouncing the word “*talking*”. The action unit (AU) activation intensities are also shown, produced (by using the OpenFace toolbox [6]) according to the Facial Action Coding System (FACS) [7], a taxonomy of human facial movements by their appearance on the face. AUs 1, 2 and 5, belonging to the brow and eye areas of the face are strongly activated only in the true expression of *surprise*. However, AUs 25 and 26 which belong to the lower orofacial area are activated in both cases with similar intensities, corroborating that speech articulation may take the semblance of emotional expressions.



**Figure 1.** (Left) An expression of *surprise* from the CK+ [4] dataset. (Middle) A frame from the RAVDESS [5] dataset expressing the *neutral* emotion while speaking. (Right) A bar chart of action unit activation intensities for the CK+ image (in red) and the RAVDESS frame (in blue). The vertical axis shows the different action units (AUs), and to each are assigned two intensity bars for the two images. The horizontal axis shows the intensities of the activations on a scale from 0 to 5, cropped to the maximum value.

Limited research efforts have been devoted to the problem, though the *speaking effect* has gained attention by virtue of the growing efforts toward detecting affective states from a variety of modalities, cogently visual and auditory. Thus, work has been done on how to compensate for it and hence improve recognition accuracy. Indeed, such an effect is one of the hindrances in replicating accuracies of the acted datasets on spontaneous interactions.

Our chief concern here is the following research question. How can we account for the *speaking effect* in the design of an expression recognition system, rather than filtering it out as noise [8–10], in order to improve the system performance? Specifically, we attack the problem of whether and under what circumstances adding to a deep recognition model input features related to the speech articulation process will increase emotion recognition accuracy in speaking subjects. To the best of our knowledge, deep learning has not been explicitly applied to such effect, while generally it is widely applied to affective facial expression recognition.

This question was addressed from an empirical standpoint. To said end we developed two deep learning neural network models and tested their performances first on facial features only; subsequently, both facial and speech articulation features were combined. The models were furthermore trained multiple times on varying numbers of consecutive frames to determine the dependence between the temporal context and the accuracy increase.

Our main finding is that of a statistically significant increase in accuracy when features related to speech articulation are added in input, the increase being greater with more consecutive frames given in input.

In the next section we provide the necessary background and review the literature on the subject. In Section 3 we then describe the experimental setup, models, data and parameters used. Finally, in Section 4 we present the results, and in Section 5 the conclusions, respectively.

## 2. Background and Related Works

### 2.1. What is an Expression?

Facial expressions are a fundamental component of human interaction, along with prosody, body posture, hand gestures and other behavioral cues that carry non-verbal messages [11]. Facial expressions reveal the focal point of our attention; show signal comprehension or disagreement; indicate humorous or serious intent; and generally regulate and enrich interpersonal communication [12]. Consequently, automatic facial expression recognition (AFER) has consumed a lot of effort in the past two decades due to it being fundamental for human–computer interactions and having applications in a plethora of fields and problems, such as psychology, medicine, education, computer science, well-being assessment and ambient intelligence [13–16].

Darwin was the first to stipulate that certain facial configurations are expressions of certain emotion categories [17]. This idea of the universality of facial expressions across gender, race, ethnicity, culture, etc., later influenced Ekman [18] who proposed the discrete emotional model using six universal emotions—*happiness, surprise, anger, disgust, sadness* and *fear*, together with a *neutral* emotion. While critiques of Ekman’s basic emotions theory are aplenty and alternative theories of affect do exist [1], it stands that most research in AFER, as well as datasets, stand directly or indirectly on the shoulders of this theory aiming at classifying these (or a similar set of) discrete categories of emotion.

There are two main modalities in AFER research: *static* and *dynamic*. In static AFER, feature representation is obtained only from spatial information from a single image, while in dynamic AFER the temporal dynamic of an expression is also modeled [19]. Recent years have also seen a surge in multimodal systems where additional modalities are considered, such as speech (prosody) and physiological signals (see [20,21] for in-depth reviews), so that more general simulation or prediction-based approaches become affordable [22].

Early efforts in AFER research were driven by hand crafted features that were finally fed into a classifier [23]. However, with the advent of deep neural networks in recent years and the increase in data availability for AFER, deep learning became the state of the art for the problem by exceeding the performances of other methodologies by a large margin [21].

Alongside the rise in the trend of using deep neural nets, the datasets in AFER evolved to be ever more challenging. AFER is a problem a lot of attention has been dedicated to; there are plenty of publicly available datasets such as CK+, FER+, MMI, AMHUSE [4,24–26] and others (see [21,27] for a comprehensive list). However, a large number of these datasets have been collected in controlled laboratory conditions with minimum variations in head pose, illumination and occlusions. Thus, they lack the uncertainty of a real-world setting, making any machine learning model trained on them have unrealistic biases and questionable robustness. Further, the size of these datasets is seldom sufficient to train the data-hungry deep models well. This has spurred on efforts toward establishing the so called datasets of facial expressions “*in-the-wild*”, which consist of spontaneous expressions collected in diverse and realistic settings, such as the AFEW dataset [28]. The performances of state-of-the-art models differ greatly when applied to acted data and in-the-wild data. For example, a recent method described in [29] obtained 98.06%, 82.74% and 61.52% accuracy on CK+, MMI and AffectNet datasets, respectively, of which a considerably lower result was obtained on AffectNet, the only in-the-wild dataset of the three.

### 2.2. Problems with the Speaker

As stated from the beginning, one of the hindrances in replicating accuracies of the acted datasets on spontaneous interactions lays in the *speaking effect* [3]. One approach in the literature to alleviate the speaking effect, i.e., to increase accuracy of AFER on speaking subjects, rests on the assumption that emotional states and the pertaining facial cues are temporally more persistent than facial movements related to speech articulation. This approach has been adopted both by traditional approaches [8,9],

and by deep learning models that take as input single frames and output either vectors of features, or discrete emotion classes. The speech articulation-related facial movements are considered noise, and a statistic such as the mean or the mode is calculated on the features or classes seeking to cancel, smooth or filter out the “noise”. In [30] the authors fine-tuned on the FERPlus dataset a ResNet50 variant pretrained on the VGG-Face2 dataset. On the resulting classifications they calculated the mode of the predictions while ignoring the *neutral* expression and achieved 49.3% accuracy on the AFEW 6.0 dataset, thereby outperforming the baseline. Similarly, as an example of this approach applied to features, in [10] the authors extracted frame-level features with three diverse networks—VGG13, VGG16 and ResNet; and normalized the features by calculating the signed square root and performing  $\ell_2$  normalization and afterwards concatenating them. Then, over all the feature dimensions, the mean, the variance, the minimum and the maximum were calculated and fed into a support vector machine (SVM) classifier obtaining 59.42% on the AFEW 6.0 dataset.

While the above filtering approaches provide a certain gain in classification accuracy, they also filter out potentially useful information for emotion classification, thereby leaving room for further improvements in accuracy.

On the other hand, a different approach would be to directly model the information relating to expressing emotion and that relating to speech articulation. However, works pursuing this approach are few and far between. In [31] Mariooryad et al. used asymmetric bilinear factorization to perform the decoupling of linguistic and affective information. They build on their previous work [32] that models facial expressions as having three fundamental components: speaker, lexical and emotional. In [32], by using factor analysis, they demonstrate the dominant influence of speech articulation on the orofacial features. However, their model depends on having the speech transcription to compensate for the lexical information. In [31] they rectify that dependency, achieving 60.25% on the IEMOCAP [33] dataset with four discrete emotions on a baseline of 53.8%. Wu et al. [34] developed an approach for removing the speaking effect using eigenface conversion, whereby they convert speaking faces in non-speaking faces by compensating for the articulation effect in the mouth area.

Finally, a somewhat related problem is lip reading, the connection being that a model that performs visual-only lip reading will model speech articulation itself by definition. Prior to the advent of deep learning, most lip reading techniques were based on hand crafted features and were modeled by hidden Markov models [35]. However, lip reading, as did many other problems, experienced a quantum leap in recent years thanks to both the emergence of deep learning techniques and the availability of large datasets [36]. Deep learning is employed in lip reading for either feature extraction, upon which a classifier is applied, or as an end-to-end system. The models also differ in whether they model only visual information, or both audio and visual, with the latter providing greater accuracy [37].

A particularly interesting model is LipNet [38] by Assael et al., an end-to-end, visual information only model that maps a variable-length sequence of video frames to text, making use of spatio-temporal convolutions followed by recurrent layers. The accuracy obtained by the model outperformed both experienced human lip readers and the previous state of the art by nearly 10%. The authors also analyzed the learned representations of LipNet from a phonological perspective by generating a saliency map for utterances, which was then found to correspond to phonologically important regions in the video. They also analyzed how well the learned representations map to *visemes*, a visual equivalent of the phoneme, first by predicting phonemes, and then by mapping them to visemes through a predefined mapping. They found the confusion matrix of visemes on the GRID corpus to be accurate.

### 3. Method and Experimental Setting

Our chief purpose is to investigate a method that exploits all the available information to improve the emotion recognition accuracy for a speaking subject, rather than trying to clean up the data. In particular, we assess the effectiveness of including information relating to visemes in input to a deep learning model. Intuitively, one would surmise that if the model has information on how

the mouth area movements relate to language articulation, the model can learn better than otherwise to compensate for it when classifying emotion.

### 3.1. Models

The choice of using a deep learning model lays in the fact that state of the art for AFER is achieved by this approach. Most of the deep models used in research adopt convolutional layers, recurrent layers, or both. In order to investigate both these architectures, we develop two models, a convolutional neural network (CNN) and a gated recurrent unit (GRU) recurrent neural network (RNN), to classify the emotion expressed by a speaking subject in a video sequence. Each of these models has two modalities: the first is trained only on data relating to facial features which we will call the *face model*, and the second is trained both on data relating to facial features and viseme-related features which we will call the *lip model*. Finally, each of the models and the modalities is trained  $n$  times, each time on a different number of consecutive video frames provided in input. This approach provides an opportunity to answer not only the main question, whether the inclusion of viseme-related features in input improves accuracy for a speaking subject, but also to see the relationship between the accuracy and the temporal context provided as input to the model.

In addition to the above, we develop a simple multilayer perceptron (MLP), taking as input only facial features for a single video frame and classifying the expressed emotion, which we will call the *frame model*. When presented with a video sequence in input, we evaluate the expression with the model frame-by-frame, and then get the mode of the predictions as the predicted class for the whole sequence. With this approach we replicate what is done in [30] by Albanie et al. with the purpose of using it as a baseline for the evaluation of the other models. The only difference to [30] is that we do not treat the *neutral* expression as noise by leaving it out of the mode calculation but instead for completeness consider it a first-class citizen.

To extract emotion-related facial features from video sequences to feed into our models, we adopt a pretrained model based on the VGG19 architecture (available at <https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch.git>). The model was trained to classify the seven basic emotions: *anger* (ANG), *disgust* (DIS), *fear* (FEA), *happiness* (HAP), *sadness* (SAD), *surprise* (SUR) and *neutral* (NEU), achieving 73.11% on the FER2013 dataset, and 94.64% on the CK+ dataset. We discarded the last layer of the model and took the resulting 512-dimensional feature vector as facial features to input in our models. On the other hand, we took the articulation features from LipNet [38]. Visual-only lip reading as a task is intimately tied to mouth configurations and dynamics in language articulation, and a model trained for lip reading will have learned their representations. We support this claim with the high accuracy of the mapping of LipNet outputs to visemes in [38]. While it would be possible to consider a different set of features to represent speech-related mouth configurations, e.g., including phoneme information, it would deviate from our visual-only approach, and we retain it would be a more indirect way of representing them. We discarded the bidirectional GRU layers and took the output of the last spatio-temporal CNN (STCNN) + spatial pooling layers, yielding a 96-dimensional feature vector representing mouth movement features. In line with the emotions that the facial features model is trained to predict, and in line with the available emotions in the dataset we use, all of the models we train predict the same seven basic emotions.

### 3.2. Dataset

There are several datasets which feature speaking subjects annotated with emotions that could be considered for our experiments. IEMOCAP [33] features 12 h of dyadic interactions, both acted and staged, annotated with the seven discrete emotional classes. The state of the art on IEMOCAP for visual AFER stands at 64% and is achieved by deep modelling of interactions via multiple modalities [39]. SEMAINE [40] and its modified counterpart AVEC [41] contain a total of 959 interactions between humans and artificially intelligent agents. While AVEC does not provide discrete emotions, SEMAINE annotations are provided in two flavors, real valued affective dimensions

and discrete emotions. However, the discrete emotions do not contain the full set of Ekman's basic emotions but only a subset, lacking *surprise* and adding *contempt* and *amusement*, which reflects its purpose of being a dataset of dyadic interactions. The MELD [42] dataset contains about 13,000 utterances from multi-party dialogues from the TV-series Friends, with real valued annotations provided for the available modalities: audio, visual and textual. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [5] consists of 24 professional actors vocalizing lexically matched statements. Speech includes all of Ekman's basic emotions, namely, *calm*, *happy*, *sad*, *angry*, *fearful*, *surprise* and *disgust* expressions, with each expression produced at two levels of emotional intensity, with an additional *neutral* expression at only a single intensity. RAVDESS is a very recent dataset, and has been adopted in research so far mainly in the field of emotion detection from speech. State of the art accuracy on the dataset for the problem reaches over 70% (see [43–47] for an overview). However, to our knowledge there is no work on visual only emotion recognition on the RAVDESS dataset.

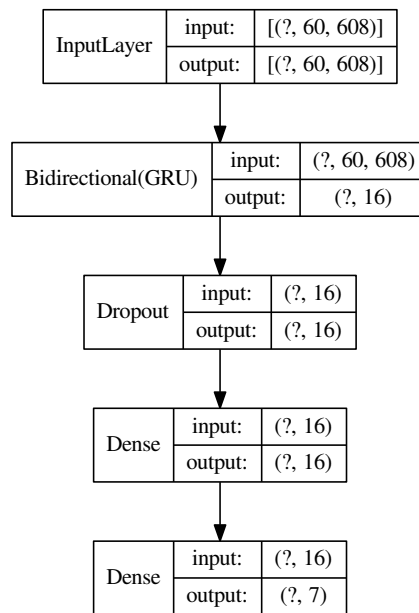
We turn to RAVDESS, it being particularly suitable for our experiments as it covers all the basic emotions, with the addition of *calm* which we ignore, and it has multiple subjects which introduces diversity in the data, among which there is minimal variation in external conditions. In fact, the actors are directly facing the camera with consistent illumination and no occlusions, and all are uttering the same phrases, making the only variations in the data the visual aspects of the actors and their performances. The face and the mouth being fully visible in consistent conditions makes for optimal inputs to both visual AFER and lip reading tasks. This reduces any noise and variation fundamentally unrelated to the problem at hand, allowing one to discover to a greater extent than otherwise the potential of our approach in increasing AFER accuracy. Furthermore, the other mentioned datasets are mainly captured in a setting of dyadic interactions and geared toward modeling the same through multimodal approaches, or they feature a non standard subset of Ekman's emotions as annotations. For these reasons we opted for RAVDESS, thereby allowing us to focus on Ekman's emotions and evade the additional complexity introduced by an interaction setting.

### 3.3. Implementation Details

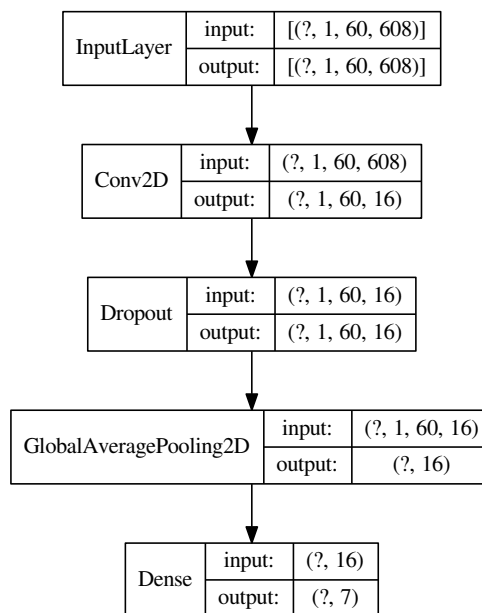
The RNN, the CNN and the MLP were all chosen with simple architectures due to the fact that the features in input are of low dimensionality, the dataset is not of vast size and the inputs themselves are the outputs of two relatively more complex deep models. Be that as it may, more complex architectures were attempted but with no discernible improvements in accuracy and worsening overfitting. The RNN network, shown in Figure 2, is composed of a layer of eight GRU bidirectional cells, followed by a dense layer with 16 neurons and a softmax output layer. The CNN network, shown in Figure 3, is composed of 16 filters, a kernel size of 8 and a stride of 1 in both the dimensions, and the inputs were zero padded so as to preserve input dimensionality at layer output. Following the convolutional layer are a 2D global average pooling layer and a softmax output layer. The MLP had a single hidden layer with 256 hidden units. Dropout was applied on the hidden, convolutional and recurrent layers in the MLP, CNN and RNN, respectively.

The dataset was split into test and train layers by actors: actors 5, 15 and 18 were randomly chosen for testing, and the rest were used for training and validation. Splitting by actors instead of sampling from the combined data offers a more objective measure of performance. Had parts of the data of a particular actor been present in both the test and train datasets, it would have provided the opportunity for the models to learn actor-specific attributes and potentially display higher classification accuracy on the test data.

All the videos are converted into 25 fps from the original 30 due to LipNet having been trained on that frame rate, and the frames at which the models are trained are in the range of 5–65 in 5 frame increments. This corresponds to a range from 0.2 s to 2.6 s of input duration in increments of 0.2 s.



**Figure 2.** Schematic of the RNN lip model for 60 frames. The input dimensions represent (*batch size, number of frames, feature space dimension*). The feature space dimension is the sum of 512 dimensions of facial features and 96 dimensions of LipNet features.



**Figure 3.** Schematic of the CNN lip model for 60 frames. The input dimensions represent (*batch size, channels, number of frames, feature space dimension*). The feature space dimension is the sum of 512 dimensions of facial features and 96 dimensions of LipNet features.

All the learning parameters were shared across all the models: learning rate 0.001, dropout probability 0.5 and batch size 32. All models were trained for 50 epochs. The loss function was categorical cross entropy, and the optimizer was Adam [48].

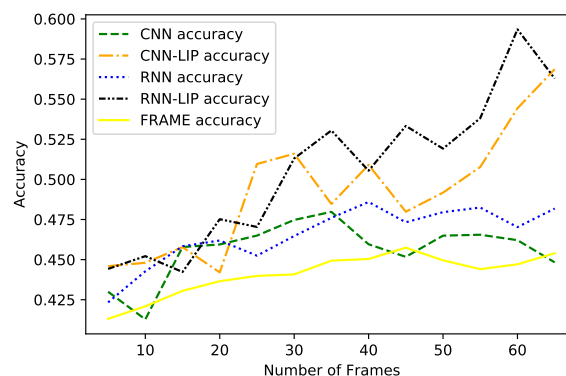
#### 4. Results and Discussion

Table 1 and Figure 4 present the accuracy of the models on test data varying the considered sequence duration. A general rising trend in accuracy can be noticed, indicating that providing more temporal context to the model increases the classification accuracy. The trend plateaus between 35 and

45 frames for all but the two lip models, which boasts the trend all the way to the maximum number of frames. The worst performing model is confidently the frame model, while the two best performing models are the CNN lip and RNN lip models, with the RNN lip model slightly outperforming the CNN lip model at higher frames. Most importantly, however, we see that there is, at frames greater than 25, a clear and persistent accuracy gain of the lip models with respect the corresponding face models.

**Table 1.** Test data accuracy of the models at different frame values. Models with highest accuracy are indicated in bold.

Frames	CNN	CNN Lip	RNN	RNN Lip	Frame
5	0.429	0.446	0.423	0.444	0.413
10	0.413	0.448	0.443	0.452	0.421
15	0.458	0.458	0.458	0.442	0.431
20	0.459	0.442	0.462	0.475	0.437
25	0.465	0.509	0.452	0.470	0.439
30	0.475	0.516	0.465	0.513	0.441
35	<b>0.479</b>	0.485	0.476	0.530	0.449
40	0.459	0.509	<b>0.485</b>	0.505	0.450
45	0.452	0.479	0.473	0.533	<b>0.457</b>
50	0.465	0.492	0.479	0.519	0.449
55	0.466	0.508	0.482	0.538	0.444
60	0.462	0.544	0.470	<b>0.594</b>	0.447
65	0.448	<b>0.569</b>	0.482	0.563	0.454



**Figure 4.** A plot of accuracy of the models on test data.

Looking at only the difference in accuracy between the lip models and the corresponding face models, as depicted in Figure 5, we can confirm both the increase in accuracy starting at 25 frames, and that it follows a seemingly linear rising trend, as we increase the number of frames in input to the model. Running linear regression on the data in Figure 5 reveals a statistically significant slope of 0.0014, representing the increase in accuracy per additional frame in input. However, the real trend is likely not linear due to there being an upper limit to the accuracy metric, and increasing further the number of frames would raise questions as to the persistence of expressions of base emotions in spontaneous conversations. Nonetheless, the accuracy gains for the data at hand seem to rise linearly with positive values starting from 1 s in input sequence duration. While attributing interpretation to the workings of a black box system, such as a neural network, is a slippery endeavour, we retain the reason behind the lip models not plateauing and the increases rising linearly: the information-related to articulation boosts the utility of the information-related to the face. Under the assumption that articulation-related features in speaking subjects carry no information on affect, they induce the network to become invariant to articulation. Thus, when providing more information in input with a greater number of frames, the lip models are able to classify better.



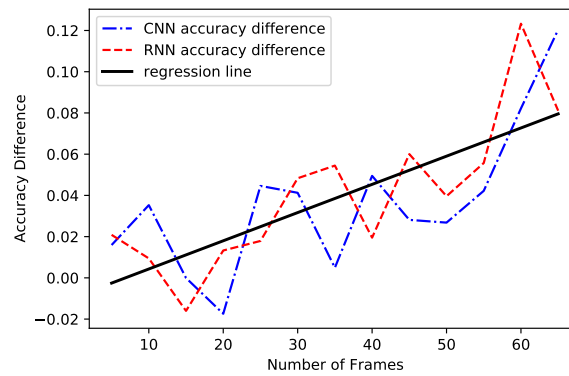


Figure 5. Differences in accuracy between lip and facial only models on test data.

On Figure 6 a representative example of the training behavior of the models can be seen. It depicts the training set and validation set accuracy values for different epochs for two models: the best performing RNN lip and the corresponding RNN face model. A sharp difference in validation and training accuracy can be noticed in both the models revealing a case of overfitting, as the training accuracy increases while the validation accuracy stays approximately the same with oscillations. This discrepancy in accuracy could not be alleviated by increasing dropout probability, and has been described previously in [21] by Li et al. as a general problem of AFER. They attribute this phenomenon to deep neural networks requiring a large amount of training data to avoid overfitting, and the existing facial expression databases not being sufficiently large for the task. Additionally, they point out that high inter-subject variations due to different personal attributes and levels of expressiveness, variations in pose, illumination and occlusions, all common in spontaneous interactions, contribute to the overfitting issue.

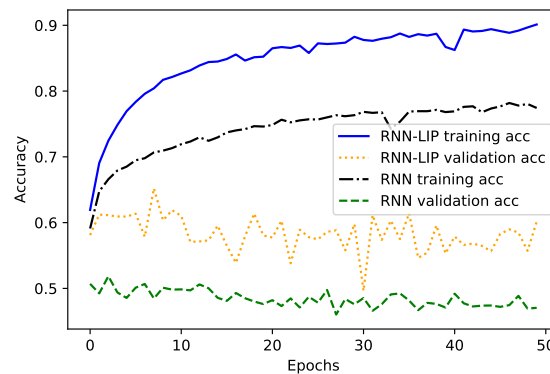


Figure 6. Training and validation accuracy for the best performing RNN models.

To evaluate in more detail how the accuracy is distributed over classes, we can examine the confusion matrices of the models evaluated on the test dataset. Figures 7–10 depict these for the best performing frame, CNN lip and RNN lip models. For the CNN lip and RNN lip models, respectively, the face model trained on the same number of frames was also included. In addition to the confusion matrices, heatmaps of the differences in the confusion matrices between the lip models and the corresponding face models are included. Obtained by subtracting the face model matrix from the lip model matrix, positive values signify increases in accuracy for the lip model, and vice versa for negative values. The horizontal axis represents the predicted classes, while the vertical classes represent ground truth. The rows of the confusion matrices are individually normalized to a range from 0 to 1.

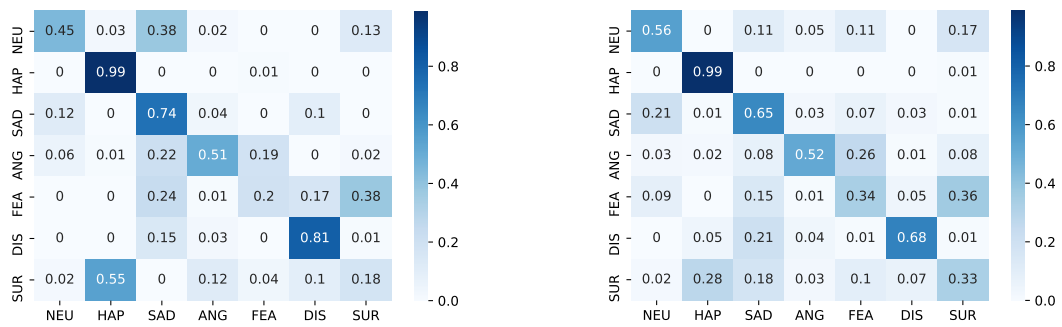


Figure 7. (Left) Confusion matrix of the best performing CNN lip model. (Right) Confusion matrix of the best performing RNN lip model.

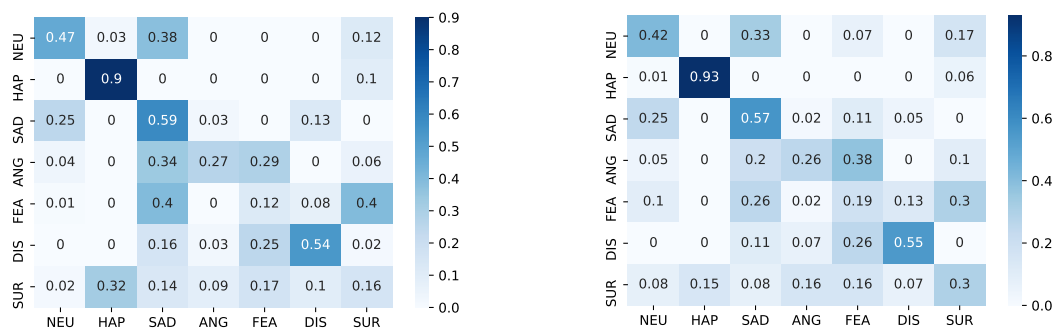


Figure 8. (Left) Confusion matrix of the CNN model with the same number of frames as the best performing CNN lip model. (Right) Confusion matrix of the RNN model with the same number of frames as the best performing RNN lip model.

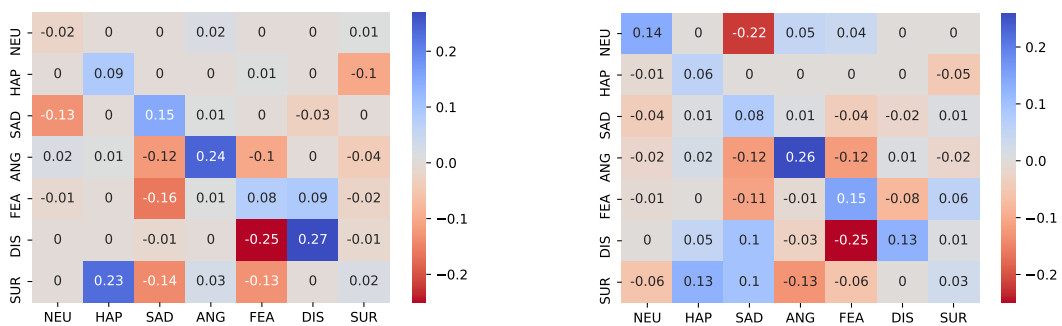


Figure 9. (Left) Differences between the confusion matrices of the CNN lip and the CNN model. (Right) Differences between the confusion matrices of the RNN lip and the RNN model.

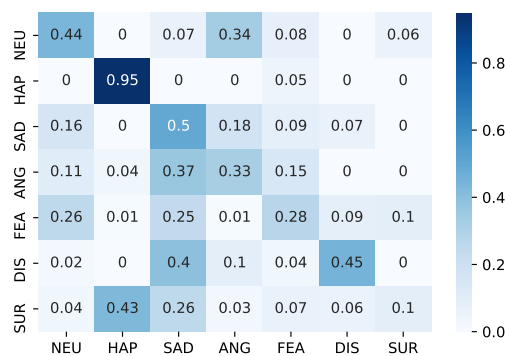


Figure 10. Confusion matrix of the best performing Frame model.

Patterns common to all of the models can be observed: *neutral* and *sad* appear to be somewhat ambiguous for the models; *happiness* has a very high accuracy; and *fear* and *surprise* have very low

accuracy. *Surprise* is in part confused for *happiness*, and *fear* for *surprise*. The frame model displays some deviations from these patterns and generally has lower accuracy, but fares well. Its biggest weak spot is the poor performance on *surprise*. If we confront CNN models with the RNN models we can also notice certain patterns, such as the RNN models being better at classifying *surprise*. We found these curious dissimilarities between the CNN and RNN models consistently across models with differing numbers of frames, which might suggest a hybrid model would achieve superior performance.

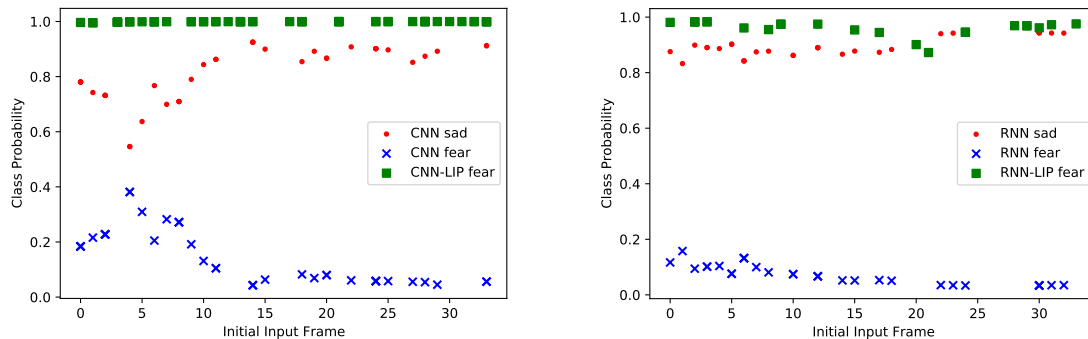
The improvements obtained by the lip models over the corresponding face models are visible in Figure 9 for the RNN and CNN models, respectively, depicted as heatmaps of values obtained by subtracting the confusion matrix of the face model from the confusion matrix of the lip model. Recall that a desirable outcome for such a representation is to have positive values on the diagonal, meaning that classification accuracy has increased, and having negative values elsewhere, meaning that misclassification has decreased. We observe in Figure 9 that nearly all values on the diagonals are positive, meaning that nearly all classes improved in accuracy. The largest improvements can be noted for the classes *anger*, *fear* and *disgust*. In fact, for both the RNN and the CNN models the *anger* class gained around 25% in accuracy, with the biggest increase of 27% being *disgust* for the CNN model. Similarly, the misclassification values were generally diminished with the biggest being misclassifying *disgust* as *fear*, in both of the models. The RNN lip model also diminished the misclassification of *neutral* as *sad* by 22%. The exception is misclassifying *surprise* as *happiness* that increased for both of the models.

Figure 11 shows some frames from RAVDESS of Actor 18 expressing *fear* and pronouncing the phrase “*The kids are talking by the door.*” The video sequence was run through the 60 frame models and Figure 12 shows the attribution of probability to the classes *fear* and *sadness*, respectively, for different initial frames provided in input. The CNN lip model and the RNN lip model respectively assign a very high probability of the correct emotion to the sequence regardless of the number of frames in input. However, both the CNN and the RNN models assigned a much higher probability to *sadness* irrespective of the starting frame, aside from the CNN model at initial frame 5, where it shows indecisiveness among the two emotions. The difference in performance among the lip models and their respective face models can be explained by mouth configurations of the actor’s performance influencing the face models to classify the expressed emotion incorrectly. In this sequence especially, both the expression of emotion and the mouth movements are particularly exaggerated by the actor, making it more likely to confuse the face models. This, however, occurs to a far lesser extent with the lip models due to them being able to compensate somewhat for the mouth movements related to articulation.

From the above considerations, we can claim that the inclusion of lip-related features to a deep learning model classifying basic emotions in speaking subjects improves recognition accuracy, but that the extent of the improvement is highly dependent on the number of frames provided in input. However, while the RAVDESS dataset is high by definition, the pose of the speakers stays the same, the utterances are the same across speakers, there are no occlusions and the illumination is consistent; datasets in-the-wild and videos of spontaneous interactions will most likely not satisfy most of the conditions listed above. Furthermore, while usual AFER models are known to display lesser accuracy when provided with such input, the lip model introduces an additional component sensitive to all of the factors above, the model extracting lip features. If any of the factors above influence lip reading performance—among which we suspect head pose variations to be the most deleterious to accuracy—the same will propagate to our model. This gives way to the expectation that in cases of datasets in-the-wild and videos of spontaneous interactions, the increases in accuracy will be lower than the ones reported in our experiment.



**Figure 11.** From left to right, top to bottom, frames sequentially extracted in seven frame increments from the RAVDESS dataset of Actor 18 expressing *fear* and pronouncing “The kids are talking by the door.”



**Figure 12.** Plots of the performance of the 60 frame models on the test video illustrated in Figure 11. The horizontal axis shows the initial frame provided in inputs to the models, while the vertical axis measures the probability assigned by the models to a particular emotion class. On the left are the probabilities of *fear* of the CNN lip model and the probabilities of *fear* and *sadness* of the CNN model. On the right are the probabilities of *fear* of the RNN lip model and the probabilities of *fear* and *sadness* of the RNN model.

### 5. Conclusions

We explore the problem of automatic facial expression recognition in speaking subjects. As language articulation lessens facial expression recognition accuracy we propose adding articulation-related features, extracted from a neural network trained to lip read, in input to a deep learning model.

We devise two models: a spatio-temporal CNN and a GRU cell RNN. They are first trained on facial features only, and afterwards both on facial features and articulation features on the RAVDESS dataset. We find that including information related to language articulation improves accuracy up to 12%, but that the increase in accuracy is highly dependent on the number of consecutive frames

provided in input. Accuracy gains were positive and linearly rose for input segments from 1 to 2.6 s in length.

While the results are promising, RAVDESS is an acted dataset and in future works we intend to evaluate our approach on benchmark in-the-wild datasets, such as the AFEW 6.0, to see the accuracy improvement when confronted with spontaneous interactions, occlusions, and other challenges non-acted scenarios bring. In general, the still low accuracy of AFER in speaking subjects offers ample opportunity for research and improvement. It would be interesting to see how the upper area of the face is used for communicating nonverbal cues unrelated to affect, which to the best of our knowledge hasn't been addressed before. Also, human communication being a multimodal phenomenon, we consider the greatest potential for improvement in emotion recognition in speaking subjects laying in modelling multiple modalities, in addition to visual. Thus, modelling prosody, semantic, or textual information in addition to visual might further improve performance.

**Author Contributions:** Conceptualization, G.B. and A.F.; data curation, S.B. and A.D.; formal analysis, S.B., G.B. and A.F.; investigation, S.B. and A.D.; methodology, G.B., A.F. and R.L.; project administration, G.B. and A.F.; resources, G.B.; software, S.B. and A.D.; supervision, G.B., A.F. and R.L.; validation, S.B., A.D. and R.L.; visualization, A.D.; writing—original draft, S.B., G.B. and A.F.; writing—review and editing, S.B., G.B., A.F., A.D. and R.L. All authors have read and agree to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** This work has been partially supported by Fondazione Cariplo, through the project “Stairway to elders: bridging space, time and emotions in their social environment for wellbeing”, grant no. 2018-0858.

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

## References

- Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Int.* **2019**, *20*, 1–68. [[CrossRef](#)] [[PubMed](#)]
- Kim, Y.; Mower Provost, E. Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 27–36.
- Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
- Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2010; pp. 94–101.
- 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](#)] [[PubMed](#)]
- Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: An open source facial behavior analysis toolkit. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–10.
- Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, MS, USA, 1997.
- Zeng, Z.; Tu, J.; Liu, M.; Zhang, T.; Rizzolo, N.; Zhang, Z.; Huang, T.S.; Roth, D.; Levinson, S. Bimodal HCI-related affect recognition. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 14–15 October 2004; pp. 137–143.
- Zeng, Z.; Tu, J.; Liu, M.; Huang, T.S.; Pianfetti, B.; Roth, D.; Levinson, S. Audio-visual affect recognition. *IEEE Trans. Multimed.* **2007**, *9*, 424–428. [[CrossRef](#)]
- Bargal, S.A.; Barsoum, E.; Ferrer, C.C.; Zhang, C. Emotion recognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 433–436.
- Mehrabian, A. *Nonverbal Communication*; Transaction Publishers: Piscataway, NJ, USA, 1972.

12. Pantic, M.; Patras, I. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Syst. Man Cybern. Part B* **2006**, *36*, 433–449. [[CrossRef](#)]
13. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697. [[CrossRef](#)]
14. Pantic, M.; Nijholt, A.; Pentland, A.; Huanag, T.S. Human-Centred Intelligent Human? Computer Interaction (HCI<sup>2</sup>): How far are we from attaining it? *Int. J. Auton. Adapt. Commun. Syst.* **2008**, *1*, 168–187. [[CrossRef](#)]
15. Grossi, G.; Lanzarotti, R.; Napoletano, P.; Noceti, N.; Odone, F. Positive technology for elderly well-being: A review. *Patt. Recognit. Lett.* **2019**, in press. [[CrossRef](#)]
16. Boccignone, G.; de'Sperati, C.; Granato, M.; Grossi, G.; Lanzarotti, R.; Noceti, N.; Odone, F. Stairway to Elders: Bridging Space, Time and Emotions in Their Social Environment for Wellbeing. In *International Conference on Pattern Recognition Applications and Methods*; SciTePress: Setubal, Portugal, 2020; pp. 548–554.
17. Ekman, P. *Darwin and Facial Expression: A Century of Research in Review*; Ishk: Los Altos, CA, USA, 2006.
18. Ekman, P. Basic emotions. *Handb. Cogn. Emot.* **1999**, *98*, 16.
19. Sariyanidi, E.; Gunes, H.; Cavallaro, A. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Patt. Anal. Mach. Intell.* **2015**, *37*, 1113–1133. [[CrossRef](#)]
20. D'Mello, S.K.; Kory, J. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv. CSUR* **2015**, *47*, 43. [[CrossRef](#)]
21. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**. [[CrossRef](#)]
22. Boccignone, G.; Conte, D.; Cuculo, V.; D'Amelio, A.; Grossi, G.; Lanzarotti, R. Deep Construction of an Affective Latent Space via Multimodal Enactment. *IEEE Trans. Cogn. Dev. Syst.* **2018**, *10*, 865–880. [[CrossRef](#)]
23. Kumari, J.; Rajesh, R.; Pooja, K. Facial expression recognition: A survey. *Proc. Comput. Sci.* **2015**, *58*, 486–491. [[CrossRef](#)]
24. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 12–16 November 2016; pp. 279–283.
25. Valstar, M.; Pantic, M. Induced disgust, happiness and surprise: An addition to the MMI facial expression database. In *Proc. Int'l Conf. Language Resources and Evaluation; EMOTION*: Paris, France, 2010; pp. 65–70. Available online: [http://www.cs.nott.ac.uk/~pszm/~/Documents/MMI\\_spontaneous.pdf](http://www.cs.nott.ac.uk/~pszm/~/Documents/MMI_spontaneous.pdf) (accessed on 1 January 2010).
26. Boccignone, G.; Conte, D.; Cuculo, V.; Lanzarotti, R. AMHUSE: A multimodal dataset for HUMour SENSing. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, Skotland, 13–17 November 2017; pp. 438–445.
27. Haamer, R.E.; Rusadze, E.; Lsi, I.; Ahmed, T.; Escalera, S.; Anbarjafari, G. Review on Emotion Recognition Databases. *Hum. Robot Interact. Theor. Appl.* **2017**, *3*, 39–63.
28. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia* **2012**, *19*, 34–41. [[CrossRef](#)]
29. Chen, Y.; Wang, J.; Chen, S.; Shi, Z.; Cai, J. Facial Motion Prior Networks for Facial Expression Recognition. *arXiv* **2019**, arXiv:1902.08788.
30. Albanie, S.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Emotion recognition in speech using cross-modal transfer in the wild. In *Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Korea, 22–26 October 2018; pp. 292–301.
31. Mariooryad, S.; Busso, C. Facial expression recognition in the presence of speech using blind lexical compensation. *IEEE Trans. Affect. Comput.* **2015**, *7*, 346–359. [[CrossRef](#)]
32. Mariooryad, S.; Busso, C. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Commun.* **2014**, *57*, 1–12. [[CrossRef](#)]
33. 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](#)]
34. Wu, C.H.; Wei, W.L.; Lin, J.C.; Lee, W.Y. Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion. *IEEE Trans. Multimed.* **2013**, *15*, 1732–1744. [[CrossRef](#)]

35. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Proc. Mag.* **2012**, *29*, 82–97. [[CrossRef](#)]
36. Afouras, T.; Chung, J.S.; Zisserman, A. Deep lip reading: A comparison of models and an online application. *arXiv* **2018**, arXiv:1806.06053 .
37. Fernandez-Lopez, A.; Sukno, F.M. Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput.* **2018**, *78*, 53–72. [[CrossRef](#)]
38. Assael, Y.M.; Shillingford, B.; Whiteson, S.; De Freitas, N. Lipnet: End-to-end sentence-level lipreading. *arXiv* **2016**, arXiv:1611.01599.
39. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguecn: A graph convolutional neural network for emotion recognition in conversation. *arXiv* **2019**, arXiv:1908.11540.
40. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* **2011**, *3*, 5–17. [[CrossRef](#)]
41. Schuller, B.; Valstar, M.; Eyben, F.; McKeown, G.; Cowie, R.; Pantic, M. Avec 2011—the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*; Springer: Berlin, Germany, 2011; pp. 415–424.
42. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
43. Bhavan, A.; Chauhan, P.; Hitkul; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl. Based Syst.* **2019**, *184*, 104886. [[CrossRef](#)]
44. Zamil, A.A.A.; Hasan, S.; Baki, S.M.J.; Adam, J.M.; Zaman, I. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*; IEEE: Piscataway, NJ, USA, 2019; pp. 281–285.
45. Jalal, M.A.; Loweimi, E.; Moore, R.K.; Hain, T. Learning temporal clusters using capsule routing for speech emotion recognition. *Proc. Intersp.* **2019**, *2019*, 1701–1705.
46. Haque, A.; Guo, M.; Verma, P.; Fei-Fei, L. Audio-linguistic embeddings for spoken sentences. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2019; pp. 7355–7359.
47. Jannat, R.; Tynes, I.; Lime, L.L.; Adorno, J.; Canavan, S. Ubiquitous emotion recognition using audio and video data. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, Singapore, 8–12 October 2018; pp. 956–959.
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).