EMOLIPS: Towards Reliable Emotional Speech Lip-Reading

Dmitry Ryumin *, Elena Ryumina and Denis Ivanko

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 199178 St. Petersburg, Russia; ryumina_ev@mail.ru (E.R.); ivanko.d@iias.spb.su (D.I.)
* Correspondence: ryumin.d@iias.spb.su

Abstract: In this article, we present a novel approach for emotional speech lip-reading (EMOLIPS). This two-level approach to emotional speech to text recognition based on visual data processing is motivated by human perception and the recent developments in multimodal deep learning. The proposed approach uses visual speech data to determine the type of speech emotion. The speech data are then processed using one of the emotional lip-reading models trained from scratch. This essentially resolves the multi-emotional lip-reading issue associated with most real-life scenarios. We implemented these models as a combination of EMO-3DCNN-GRU architecture for emotion recognition and 3DCNN-BiLSTM architecture for automatic lip-reading. We evaluated the models on the CREMA-D and RAVDESS emotional speech corpora. In addition, this article provides a detailed review of recent advances in automated lip-reading and emotion recognition that have been developed over the last 5 years (2018–2023). In comparison to existing research, we mainly focus on the valuable progress brought with the introduction of deep learning to the field and skip the description of traditional approaches. The EMOLIPS approach significantly improves the state-of-the-art accuracy for phrase recognition due to considering emotional features of the pronounced audio-visual speech up to 91.9% and 90.9% for RAVDESS and CREMA-D, respectively. Moreover, we present an extensive experimental investigation that demonstrates how different emotions (happiness, anger, disgust, fear, sadness, and neutral), valence (positive, neutral, and negative) and binary (emotional and neutral) affect automatic lip-reading.

Keywords: lip-reading; visual speech recognition; emotional speech; speech to text; affective computing; deep learning; machine learning; human–computer interaction

MSC: 68T10

1. Introduction

When speaking, both audio and visual speech cues are generated simultaneously by human articulators [1–5]. Consequently, there is a strong correlation between lip movements and audio signals, allowing us to extract linguistic information not only from the audio but also from the visual modality. This phenomenon is also evident in human communication. When there is acoustic noise present, individuals often shift their focus to the movement of their interlocutor’s lips in order to enhance their comprehension of the speech. People begin to “watch” speech, not just listen to it.

This fact has led to the development of the Visual Speech Recognition (VSR) field, also known as lip-reading [6–9]. Its main idea is to understand human speech based solely on lip movements. However, until the recent development of Neutral Network (NN) and Deep Learning (DL), automatic systems were not able to model this feature of human speech perception with sufficient Accuracy (Acc). Nevertheless, at the moment, automatic VSR has achieved tremendous success, especially in limited-vocabulary tasks, far surpassing human abilities to lip-read. VSR is useful when the acoustic signal is corrupted by background noise or when the speaker is too far away to hear distinctly [10].
While the performance of current State-of-the-Art (SOTA) lip-reading methods [8,11–15] has improved significantly in emotionally neutral Speech Recognition (SR) tasks, they still suffer with crucial performance degradations in real application scenarios, in particular, when uttered speech is affected by human emotions. However, in real applications it is often crucially important for a person that an automatic system understands his/her requests in an emotional state. Emotions are expressed differently in acoustic and visual modalities, and there is no “out of the box” solution for emotion-robust VSR to date. It changes depending on a speaker’s emotional state: the timbre, pitch, and loudness of the voice; the duration of sounds, pauses, and articulation. Therefore, the task of recognizing emotional visual Speech-to-Text (S2T) is extremely challenging.

Modern Speech Emotion Recognition (SER) focuses on enabling devices to perform emotion detection using Deep Neutral Network (DNN) models [16–19]. Currently, they have a wide range of applications in healthcare, education, media, etc. However, the emotional VSR performance is limited by the lack of methodology and the unavailability of labeled corpora suitable for model training.

Motivated by human perception and the recent developments in multimodal DL, we propose a two-level approach to emotional S2T recognition based on visual data processing (EMOLIPS). The approach is built on the results of previous work in emotion recognition [20] and automated lip-reading [8] and leverages recent advances in DL for both. The main advantage of the proposed approach is improving automatic lip-reading Acc by genuinely tackling emotion recognition first, followed by emotion-specific VSR. We make the source code and models free for research use, as we intend to establish a baseline for future research in emotional S2T recognition. The source code and trained models are available in our GitHub (https://github.com/SMIL-SPCRAS/EMOLIPS, accessed on 25 October 2023).

The remainder of this article is organized as follows: Section 2 provides an in-depth analysis of the recent literature pertaining to SOTA approaches in both Visual Emotion Recognition (VER) and lip-reading. In Section 3, we describe the corpora used for training, development, and testing our proposed method and model. In Section 4, we provide a two-level approach called EMOLIPS for emotional speech lip-reading. The experimental results of our approach are presented in Section 5. Finally, some concluding remarks are presented in Section 6.

2. Related Work

In this section we carefully analyze the recent relevant literature related to the SOTA approaches for VER as well as approaches for lip-reading.


The SOTA emotion recognition approaches have been improved tremendously over the last five years due to the rapid development of modern Machine Learning (ML) techniques and algorithms, as well as the release of large-scale corpora. Emotion recognition has drawn increasingly intense attention from researchers in a variety of scientific fields. Human emotions can be recognized and analyzed from visual data (facial expressions [21–23], behavior [24], gesture [25–27], pose [28]), acoustic data (speech) [29], physiological signals [30–33], or even text [34]. In this section, we will focus on SOTA approaches for VER based on facial expression analysis.

The visual information of facial expressions is a crucial indicator of a speaker’s emotion. According to the SOTA pipeline for VER, the first step is to localize the Region-of-Interest (ROI). Several open-source Computer Vision (CV) libraries allow the detection of facial variations and facial morphology (e.g., the Dlib (http://dlib.net/, accessed on 25 October 2023) or MediaPipe (https://google.github.io/mediapipe/, accessed on 25 October 2023) libraries). This step evaluates the position of the landmarks on a face, which contain meaningful information about a person’s facial expression that helps to perform automatic emotion recognition [35].
For any utterance, the underlying emotions are classified using SER. Classification of SER can be carried out in two ways:
(a) Traditional classifiers;
(b) DL classifiers.

The latter are an order of magnitude superior to traditional ones in terms of Acc of emotion recognition. Therefore, only approaches based on DL will be considered below.

DNN is a multilayered NN used for complex data processing. Many researchers have investigated different deep models, including Deep Belief Networks (DBNs) [36], Convolutional Neural Networks (CNNs) [37], and Long-Short Term Memory (LSTM) networks [38] to improve the performance of SER systems. However, nowadays the most promising approaches to emotion recognition usually rely on CNNs and Recurrent Neural Networks (RNNs) at their core, with attention mechanisms and some additional modifications [18].

A CNN is a NN that consists of various layers sequentially. Usually, the model contains several convolution layers, pooling layers, and fully connected layers, followed by a SoftMax unit. This sequential network forms a feature-extraction pipeline modeling the input in the form of an abstract. The basis of CNNs is convolutional layers, which constitute filters. These layers perform a convolutional operation and pass the output to the pooling layer. The pooling layer’s main aim is to reduce the resolution of the output of the convolutional layers, thereby reducing the computational load. The resulting outcome is fed to a fully connected layer, where the data are flattened and are finally classified by the SoftMax unit.

RNNs and their variations are usually designed for capturing information from sequence/time series data. RNNs work on the recursive equation:

$$ S_t = F_W(S_{t-1}, X_t), \quad (1) $$

where $X_t$ is the input at time $t$; $S_t$ (new state) and $S_{t-1}$ (previous state) are the state at time $t$ and $t - 1$, respectively; and $F_W$ is the recursive function. The recursive function is a $\tanh$ function. The equation is simplified as:

$$ S_t = \tanh(W_s S_{t-1} + W_x X_t), \quad \text{and} \quad Y_t = W_y S_t, \quad (2) $$

where $W_s$ and $W_x$ are the weights of the previous state and input, respectively, and $Y_t$ is the output. An efficient approach based on RNNs for emotion recognition was presented in Ref. [39]. In Ref. [40], it was shown that RNNs can learn temporal aggregation and frame-level characterization over a long period.

With recent developments in NN models, and, more specifically, with the introduction of DNN architectures such as Visual Geometry Group (VGG) [41] and ResNet-like [42] that are able to consume raw data without a feature-extraction phase, modern emotion SR approaches started to shine. For the last five years, numerous research studies have been published, e.g., [43]. In existing DL emotion recognition models for recognizing spatial-temporal input, there are three common topologies: CNN-RNN [44], 3-Dimensional (3D) CNN [45], and Two-Stream Network (TSN) [46].

This enables them to capture both macro- and micro-motions. However, it cannot incorporate transferred knowledge as conveniently as CNNs–RNNs. TSNs contain two parallel CNNs: a network that processes images and a temporal network that processes motions [47].

There are also numerous studies dedicated to multimodal Audio-Visual (AV) emotion recognition [48–55]. Moreover, visual information such NNs consider different information channels. Information fusion is commonly performed using DBNs [56] or attention mechanisms [57].
A notable set of approaches transform each audio sample into a 2-Dimensional (2D) visual representation in order to be able to employ SOTA CNNs based on 2D/3D CNN [58,59]. To this end, the researchers usually compute the discrete Short Time Fourier Transform (STFT), as follows:

\[
\text{STFT} \{x[n]\}(m,k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n-mR] \cdot e^{-j \frac{2\pi}{N_x} kn},
\]

where \(x[n]\) is the discrete input signal, \(w[n]\) is a window function, \(N_x\) is the STFT length, and \(R\) is the hop (step) size. Following computation of the spectrograms as the squared magnitude of the STFT, the resulting values are then converted to a logarithmic scale (decibels) and normalized to the interval \([-1, 1]\), generating single-channel images suitable for further image processing.

Furthermore, the authors of Ref. [60] proposed an audio spectrogram Transformer method based on the well-known Transformer [61], the first convolution-free, purely attention-based model for emotion classification.

Since the main goal is to evaluate emotion in videos rather than in images (frames), the researchers in Ref. [62] focused on video analysis and evaluated two strategies to provide a final prediction for the video. The first collapses the sequence of action units generated in each timestamp into a vector that is the average of all the temporal steps. These features fed three static models: a Support Vector Machine (SVM), a k-NN, and a Multilayer Perceptron (MLP). The second strategy employed a sequential Bidirectional (Bi)LSTM with an attention mechanism to extract the video prediction from the sequence of features retrieved from each frame of the video.

As an alternative to static models, the authors adopted a sequential model with an assumption that there is relevant information not only in frames themselves but also in the order of frames. The authors used a BiLSTM network as a sequential model.

The BiLSTM layer operates in a Bi manner, which allows them to collect sequential information in two directions of the hidden states \(h_1, h_2, \ldots, h_N\) of the LSTMs. To obtain the emotion prediction from the BiLSTM layers, the embeddings of the outputs of each specific direction are concatenated, according to the equation:

\[
h_i = h_i^\rightarrow \parallel h_i^\leftarrow, \quad \text{where } h_i \in \mathbb{R}^{2L},
\]

where \(\parallel\) denotes the concatenation operator and \(L\) the size of each LSTM.

The general goal of the following attention mechanism is to distinguish the most relevant active unit associated with a specific frame while detecting the emotion in the whole video. The researchers used the actual contribution of each embedding evaluated through MLP with a non-linear activation function, similar to [63].

The attention function is a probability distribution applied to hidden states that allows us to obtain attention weights that each frame of the video receives. The models estimate the linear combination of the LSTM outputs \(h_i\) with the weights \(a_i\). Finally, the resulting feature vector is fed to a final task-specific layer for emotion recognition. The authors used a Fully Connected (FC) layer of eight neurons, followed by a SoftMax activation that returns the probability distribution over the emotional class.

A notable set of approaches in VER rely on the implementation of self-attention [61]. Self-attention is formulated via estimating the dot-product similarity in the latent space, where queries \(q\), keys \(k\), and values \(v\) are learned from the input feature representation via the trainable projection matrices \(W_q, W_k, W_v\), and a resulting representation is calculated based on them:

\[
A_i = \text{softmax} \left( \frac{qk^T}{\sqrt{d}} \right) v,
\]

where \(d\) is the dimensionality of a latent space.

This idea has been widely employed for solving a variety of tasks in emotion/affect recognition [57,64,65]. In Ref. [64], the authors presented a Transformer-based methodology...
for unaligned sequences and performed fusion of three modalities of audio, video, and text. The data from each modality are first projected via a 1-Dimensional (1D) convolutional layer to the desired dimensionality, followed by a set of Transformer blocks. Specifically, two Transformer modules were added to each modality. These representations are subsequently concatenated and another Transformer module is applied in each modality branch on the joint representations. The resulting representations from each of the three branches are subsequently concatenated for final emotion classification.

The authors of Ref. [66] recently proposed the use of Emoformer blocks for VER tasks. Emoformer has an Encoder structure similar to Transformer, excluding the decoder part. The previously described self-attention layer is used to extract features that contain emotional information effectively, and the residual structure ensures the integrity of the original information. Finally, the mapping network decouples the features and reduces the feature dimensions.

For a given input feature \( U \), three matrices of query \( Q \in \mathbb{R}^{T_Q \times d_Q} \), key \( K \in \mathbb{R}^{T_K \times d_K} \), and value \( V \in \mathbb{R}^{T_V \times d_V} \) are calculated from \( U \):

\[
[Q, K, V] = U \begin{bmatrix} W_Q \, W_K \, W_V \end{bmatrix},
\]

where \( T_Q, T_K, T_V \) represent the sequence length of the \( Q, K, V \), respectively; \( d_Q, d_K, d_V \) represent the dimensions of the \( Q, K, V \), respectively; and \( W_Q \in \mathbb{R}^{d_Q \times d_m}, W_K \in \mathbb{R}^{d_K \times d_m}, W_V \in \mathbb{R}^{d_V \times d_m} \).

Multiple self-attention layers are concatenated to obtain a multi-head attention layer:

\[
\text{MultiHead}(A) = \text{Concat}(A_1, \ldots, A_h)W,
\]

where \( A_1, \ldots, A_h \) are the outputs of the self-attention layers, \( h \) is the number of layers, and \( W \) is the weight parameter.

A residual connection with the normalization layer is used to normalize the output of the multi-head attention layer, and a feed-forward layer is employed to obtain the output of the self-attention parts:

\[
A = \text{Norm}(A + \text{MultiHead}(A)),
F = \max(0, NW_1 + b_1)W_2 + b_2,
G = \text{Norm}(F + \text{MultiHead}(F)),
\]

where \( W_1, W_2 \) are the weight parameters and \( b_1, b_2 \) are the bias parameters.

Finally, the original features \( U \) and the output of the self-attention parts \( G \) are connected through the residual structure, and a mapping network is employed to obtain the final output \( E \):

\[
H = U \oplus G,
E = \text{Map}(H),
\]

where \( \text{Map} \) represents the mapping network, which consists of five fully connected layers. Combining the above equations, we can obtain different modality emotion vectors from different input channels with Emoformer:

\[
[E_a, E_v, E_t] = \text{Emoformer}(U_a, U_v, U_t),
\]

where \( U_a, U_v, U_t \) represent the original input of audio, visuals and textual features, respectively, and \( E_a, E_v, E_t \) represent the emotion vectors of the respective modalities.

The authors of Ref. [67] proposed a method to tackle emotion recognition in dialogue videos. The researchers proposed the Recurrence based Uni-Modality Encoder (RUME), inspired by the famous Transformer architecture [61]. The authors added fully connected networks and residual operations to improve the expressiveness and stability of the emotion recognition system. The proposed structure can be formalized as:
where $X_r$ denotes the feature matrix of the utterance; RNN($\cdot$), LN($\cdot$), and FF($\cdot$) denote the RNN, normalization, and feed-forward network layers, respectively. In this study, the RNN($\cdot$) and LN($\cdot$) default to Bi Gated Recurrent Unit (GRU) and layer normalization, while the feed-forward layer consists of two fully connected networks, which can be formulated as:

$$\text{FF}(X_r) = \text{DP}(\text{FC}(\alpha(\text{FC}(X_r))))),$$

where FC($\cdot$) and DP($\cdot$) denote the fully connected network and dropout operation, respectively, and $\alpha(\cdot)$ denotes the activation function.

In this subsection, we have analyzed the SOTA SER systems based on visual data processing. DL classifiers are usually used to recognize emotions. A framework of DNN, CNN, and RNN and modifications along with an attention mechanism is usually used to achieve SOTA performance.

2.2. State-of-the-Art Approaches for Lip-Reading

In this subsection, we evaluate recent progress in VSR methodology. More detailed research on these issues can be found in related studies [68,69].

Traditionally, VSR systems consist of two processing stages: feature extraction from video data followed by lip-reading. For traditional methods, features are usually extracted around the mouth ROI. However, in recent years, with the development of DL technology, the feature-extraction step has been replaced with deep bottleneck architectures.

The first CNN image classifier to discriminate visemes was trained by the researchers in Ref. [70]. In Ref. [71], the deep bottleneck features were used for word recognition in order to take full advantage of deep convolutional layers and explore highly abstract features. Similarly, it was applied to every frame of the video in Ref. [72].

The researchers in Ref. [73] proposed using 3D convolutional filters to process spatio-temporal information of the lips. A basic 2D convolution layer from C channels to C' channels (without a bias and with unit stride) computes [74]:

$$|c = \ast \text{conv}(x, w)|_{c+ij} = \sum_{c=1}^{C} \sum_{j=1}^{k_y} \sum_{j=1}^{k_y} w_{c'lij} x_{c_{ij}l'},$$

for input $x$ and weights $w \in \mathbb{R}^{C \times C \times k_y \times k_y}$, where we define $x_{ij} = 0$ for $i, j$ out of bounds. Spatio-Temporal Convolutional Neural Networks (STCNNs) can process video data by applying convolution across both time and spatial dimensions [75], thus:

$$|c = \ast \text{stconv}(x, w)|_{c+ij} = \sum_{c=1}^{C} \sum_{j=1}^{k_y} \sum_{j=1}^{k_y} w_{c'lij} x_{c_{ij}l'},$$

Then, the researchers in [76] applied an attention mechanism to the mouth ROI. The authors proposed the Watch, Listen, Attend, and Spell networks, which learn to predict characters in sentences being spoken from a video of a talking face without audio. The authors model each character $y_i$ in the output character sequence $y = (y_1, y_2, \ldots, y_l)$ as a conditional distribution of the previous characters in the input image sequence $x = (x_1, x_2, \ldots, x_n)$ for lip-reading. Hence, we model the output probability distribution as:

$$P(y_i \mid x, x') = \prod_i P(y_i \mid x, x', y_i),$$

The model consists of three key components: the image encoder Watch, the audio encoder Listen, and the character decoder Spell. Each encoder transforms the respective input sequence into a fixed dimensional state vector $s$ and sequences of encoder outputs.
Mathematics 2023, 11, 4787

\( o = (o_1, \ldots, o_p), p \in (n, m) \). The decoder ingests the state and the attention vectors from both encoders and produces a probability distribution over the output character sequence.

\[
\begin{align*}
    s^g, o^g &= \text{Watch}(x^g), \\
    s^o, o^o &= \text{Listen}(x^o), \\
    P(y | x^g, x^o) &= \text{Spell}(s^g, s^o, o^g, o^o).
\end{align*}
\]

The three modules in the model are trained jointly.

Several training strategies and DNNs have been recently proposed for lip-reading [77]. It has received a lot of attention due to the availability of large publicly available corpora, e.g., the Lip Reading in the Wild dataset (LRW) [78] and the Lip-Reading Sentences in-the-Wild dataset (LRS) [76].

The authors of Ref. [76] proposed an image encoder that consists of a convolutional module that generates image features \( f \) for every input time-step \( x \) and a recurrent module that produces the fixed dimensional state vector \( s \) and a set of output vectors \( o \):

\[
\begin{align*}
    f^v_i &= \text{CNN}(x^v_i), \\
    h^v_i, o^v_i &= \text{LSTM}(f^v_i, h^v_{i+1}), \\
    s^v &= h^v_1.
\end{align*}
\]

The majority of SOTA approaches follow a similar lip-reading strategy that consists of a visual encoder, followed by a temporal model and a classification layer. The authors of Ref. [11] proposed a modification of self-distillation. It is based on the idea of training a series of models with the same architecture using distillation and has been recently applied to lip-reading. The teacher network provides additional supervisory signals, including inter-class similarity information. The overall loss \( \mathcal{L} \) to be optimized is the weighted combination of the cross-entropy loss \( \mathcal{L}_{CE} \) for hard targets and the Kullback-Leibler (KL) divergence loss \( \mathcal{L}_{KD} \) for soft targets:

\[
\mathcal{L} = \mathcal{L}_{CE}(y, \delta(z_s; \theta_s)) + \alpha \mathcal{L}_{KD}(\delta(z_s; \theta_s), \delta(z_t; \theta_t)),
\]

where \( z_s \) and \( z_t \) represent the embedded representations from student and teacher networks, respectively; \( \theta_s \) and \( \theta_t \) denote the learnable parameters of the student and teacher models, respectively; \( y \) is the target label; \( \delta(\cdot) \) stands for the SoftMax function; and \( \alpha \) is the balancing weight between the two terms.

A visual encoder was initially proposed in Ref. [79], and since then has been widely used and improved in later studies [80]. At the same time, the most recent advances include the temporal model and the training strategy. Bi GRUs and LSTMs, as well as Multi-Scale Temporal Convolution Network (MSTCN)s, have been the most popular temporal models [14].

Recently, as an alternative to RNN for classification, Transformer models using attention mechanisms and temporal convolutional networks have begun to be used [81,82]. The Vision Transformer (ViT) [83] first converts an image into a sequence of patch tokens by dividing it with a certain patch size and then linearly projecting each patch into tokens.

A Transformer encoder is composed of a sequence of blocks, where each block contains Multi-head Self-Attention (MSA) with a feed-forward network. It contains a two-layer multilayer perceptron with an expanding ratio \( r \) at the hidden layer, and one Gaussian Error Linear Unit (GELU) non-linearity is applied after the first linear layer. Layer Normalization (LN) is applied before every block, and residual shortcuts after every block. The input of ViT, \( x_0 \), and the processing of the \( k \)-th block can be expressed as:

\[
\begin{align*}
    x_0 &= [x_{cls} \parallel x_{patch}] + x_{pos}, \\
    y_k &= x_{k-1} + \text{MSA}(\text{LN}(x_{k-1})), \\
    x_k &= y_k + \text{FFN}(\text{LN}(y_k)),
\end{align*}
\]
where $x_{\text{cls}} \in \mathbb{R}^{1 \times C}$ and $x_{\text{patch}} \in \mathbb{R}^{N \times C}$ are the CLS and patch tokens, respectively, and $x_{\text{pos}} \in \mathbb{R}^{(1+N) \times C}$ is the position embedding. $N$ and $C$ are the number of patch tokens and the dimension of the embedding, respectively.

However, training such models requires large computing power, as well as a significant amount of training data, so one of the popular approaches in such cases is transfer learning [84]—a method of using a pre-trained model to improve predictions within a different but similar task, allowing a reduction in training time and data requirements, as well as improvement of the performance of the NN.

In order to develop real-world lip-reading systems, high-quality training and testing corpora are essential. A modern trend that has appeared recently is web-based corpora: corpora collected from open sources such as YouTube or TV shows [85]. The most well-known of them are the LRW [78], LRS2-BBC, and LRS3-TED corpora [86], among others. The combination of SOTA DL methods and large-scale visual corpora has been highly successful, achieving significant recognition Acc results and even surpassing human performance.

3. Research Corpora

The selection of a proper AV speech corpus is eventually a key task in building SR systems because SOTA solutions are usually based on data-driven ML methodology. In order to assess the influence of emotions on the Acc of VSR, it is necessary to have an adequate corpus for training models. The corpus must meet the requirements of DL in terms of the number of speakers, vocabulary size, number of repetitions, etc. Compliance with all the necessary parameters is challenging; therefore, it is necessary to conduct a separate analysis of the existing AV corpora and emotional speech that are publicly available and select the most suitable ones to solve the task.

We have reviewed the 28 existing benchmarking corpora for VSR and Audio-Visual Speech Recognition (AVSR) suitable for training DL models in our related study [68]. However, all these existing publicly available corpora (both recorded in laboratory conditions and collected from the internet) contain speech recordings only with neutral emotions, or with a minimal amount of emotions/data not labeled with emotion classes. Therefore, it can be stated that the existing corpora designed for VSR are not suitable for assessing the influence of emotions on the Acc of SR.

Thus, we need to consider corpora containing emotional speech. However, most existing emotional speech corpora were created for recognizing different classes of emotions rather than for SR. Therefore, we need to analyze the existing emotional corpora for the possibility of using them for SR tasks, along with recognition of emotion classes. In Table 1, the most widely used benchmarking corpora of emotional speech are listed with their main characteristics.

### Table 1. Comparison of AV corpora of emotional speech. # means the number of data.

<table>
<thead>
<tr>
<th>No</th>
<th>Corpus</th>
<th>Language</th>
<th>Year</th>
<th>Condition</th>
<th># Speakers</th>
<th># Utterances</th>
<th># Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FABO [87]</td>
<td>–</td>
<td>2006</td>
<td>Lab</td>
<td>23</td>
<td>246</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>eNTERFACE’05 [88]</td>
<td>English</td>
<td>2008</td>
<td>Lab</td>
<td>42</td>
<td>1166</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>IEMOCAP [89]</td>
<td>English</td>
<td>2008</td>
<td>Lab</td>
<td>10</td>
<td>3060</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>SAVEE [90]</td>
<td>English</td>
<td>2009</td>
<td>Lab</td>
<td>4</td>
<td>480 (3 phrases)</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>RML</td>
<td>Multi</td>
<td>2010</td>
<td>Lab</td>
<td>–</td>
<td>720</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>SEMAINNE [91]</td>
<td>English</td>
<td>2012</td>
<td>Lab</td>
<td>150</td>
<td>959</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>AFEW [92]</td>
<td>English</td>
<td>2012</td>
<td>Wild</td>
<td>330</td>
<td>1809</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>RECOLA [93]</td>
<td>French</td>
<td>2013</td>
<td>Lab</td>
<td>46</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>9</td>
<td>CREMA-D [94]</td>
<td>English</td>
<td>2014</td>
<td>Lab</td>
<td>91</td>
<td>7442 (12 phrases)</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>NNIME [95]</td>
<td>Chinese</td>
<td>2017</td>
<td>Lab</td>
<td>44</td>
<td>–</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>RAVDESS [96]</td>
<td>English</td>
<td>2018</td>
<td>Lab</td>
<td>24</td>
<td>4904 (2 phrases)</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>RAMAS [97]</td>
<td>Russian</td>
<td>2018</td>
<td>Lab</td>
<td>10</td>
<td>581</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>PolishDB [98]</td>
<td>Polish</td>
<td>2018</td>
<td>Lab</td>
<td>16</td>
<td>560</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>MELD [99]</td>
<td>English</td>
<td>2018</td>
<td>Wild</td>
<td>–</td>
<td>1433</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>CHEAVD v1./v2 [100]</td>
<td>Chinese</td>
<td>2018</td>
<td>Wild</td>
<td>527</td>
<td>7030</td>
<td>6-8</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>No</th>
<th>Corpus</th>
<th>Language</th>
<th>Year</th>
<th>Condition</th>
<th># Speakers</th>
<th># Utterances</th>
<th># Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>HEU Emotion [101]</td>
<td>Multi</td>
<td>2020</td>
<td>Wild</td>
<td>8984</td>
<td>16,569</td>
<td>10</td>
</tr>
<tr>
<td>18</td>
<td>CelebV-HQ [103]</td>
<td>English</td>
<td>2022</td>
<td>Wild</td>
<td>15,653</td>
<td>35,666</td>
<td>8</td>
</tr>
<tr>
<td>19</td>
<td>MimicME [104]</td>
<td>Multi</td>
<td>2022</td>
<td>Wild</td>
<td>4700</td>
<td>280,000</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>EmoSet [105]</td>
<td>Multi</td>
<td>2023</td>
<td>Wild</td>
<td>–</td>
<td>118,102</td>
<td>6</td>
</tr>
</tbody>
</table>

The FABO [87] is a bimodal face and body gesture corpus for the automatic analysis of human nonverbal affective behavior. The corpus contains approximately 1900 videos of emotional facial expressions recorded with two cameras simultaneously. One is a facial camera and the second is a body camera. The corpus was the first to combine facial and body recordings in an organized manner. This AV corpus consists of recordings of 23 speakers, 11 males and 12 females. The age ranges between 18 and 50 years old. The authors recorded nine emotions in total: angry, scared, surprised, happy, disgusted, bored, worried, sad, and uncertain.

The eNTERFACE’05 [88] is an AV emotion corpus that can be used as a benchmarking corpus for testing and evaluating video, audio, or joint AV emotion recognition methods. It includes recordings of 42 speakers of 14 different nationalities uttering English phrases. Among the speakers, thirty-four were men and eight were women. The recordings include only the head area and were recorded on a dark background.

The IEMOCAP [89] corpus consists of 151 videos of recorded dialogues. It has two speakers per session, resulting in 302 videos in the corpus. Each video is annotated for the presence of nine emotions as well as valence, arousal, and dominance. It contains approximately 12 h of AV data, including video, voice, facial motion, and text transcriptions.

The SAVEE [90] is a corpus that consists of recordings of four speakers uttering three phrases with seven different emotions and twelve phrases with a neutral state. The sentences were chosen from the standard the TIMIT corpus [106] and phonetically balanced for each emotion. The emotions were recorded when speakers watched video clips and read text on the display.

The RML (http://www.rml.ryerson.ca/rml-emotion-database.html, accessed on 25 October 2023) is one of the few multilingual emotion corpora available to date. It contains 720 samples of six different emotions. Each participating speaker speaks several languages. Video recordings were collected from eight speakers, speaking six languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). The samples were recorded in AV format on a simple front-facing camera with an average duration of each sample of around 5 s.

The SEMAINE [91] corpus is an annotated multimodal record of emotionally colored conversations between a person and a limited agent. It includes recordings of around 150 speakers for a total of 959 conversations with an average duration of 5 min. In total, it has 27 different emotional categories. High-quality recordings were made with five cameras and four microphones.

The AFEW [92] is a movie-extracted emotional corpus representing real-world conditions. The corpus contains recordings demonstrating natural head poses and movements, real-world illumination, multiple speakers in the same frame, a large variety of occlusions, etc. It was collected from 54 Hollywood movies, including 1809 videos. However, due to copyrights, only a small part of the corpus is publicly available.

The RECOLA [93] is an AV corpus of spontaneous collaborative and affective interactions in French. Speakers were recorded during a video conference while completing a task requiring collaboration. In total, 46 speakers participated in the recordings, among them 27 females and 19 males. All speakers are French-speaking while having different mother tongues: thirty-three French, eight Italian, four German and one Portuguese.
The CREMA-D [94] is a corpus of 7442 video clips of 91 actors (48 males and 43 females) from a variety of races and ethnicities. The actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (anger, disgust, fear, happiness, neutral, and sadness) and four different emotion levels (low, medium, high, and unspecified).

The NNIME [95] is a Chinese interactive multimodal emotion corpus. It includes recordings of 44 speakers engaged in spontaneous interactions, among them 24 females and 20 males. Each pair of speakers was instructed to perform spontaneously a short scene with a duration of 3 min. The general goal was to demonstrate one of six emotions (anger, frustration, happiness, neutrality, sadness, and surprise). It is the first large-scale Chinese interaction corpus that has been systematically collected, organized, and publicly released.

The RAVDESS [96] corpus consists of 4904 AV recordings of actors uttering two lexically matched statements in a neutral North American accent. In total, 24 professional actors (12 male and 12 female) participated in the recordings, trying to demonstrate eight different emotions (fear, sadness, surprise, happiness, anger, disgust, calm, and neutral). Each expression was produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

The RAMAS [97] is the first Russian AV emotional corpus. Ten young actors aged between 18 and 28 y.o. participated in the recordings (five male and five female). They expressed one of the six basic emotions (anger, sadness, disgust, happiness, fear, and surprise). The corpus contains approximately 7 h of high-quality video recordings of speakers’ faces, speech, motion-capture data, and physiological signals. The records were labeled by 21 annotators.

The PolishDB [98] is probably the only corpus available for the Polish language. The recordings were made by sixteen professional actors (eight male and eight female). The corpus covers the following modalities: facial expressions, body movement and gestures, and speech. The data are labeled with six basic emotion categories, according to Ekman’s emotion categories. To check the quality of performance, all recordings were evaluated by experts and volunteers.

The MELD [99] is a multimodal multi-party corpus for emotion recognition in conversations collected from TV series. It contains about 13,000 utterances from 1433 dialogues from the TV series Friends. Each utterance is annotated with emotion and sentiment labels and encompasses audio, visual, and textual modalities. The number of participants was limited, and 84% of the sessions were obtained with six starring actors.

The CHEAVD [100] is a Chinese natural emotional AV corpus. It contains about 140 min of emotional speeches extracted from movies, TV series, and talk shows. In total, 238 speakers, aging from children to the elderly, participated in the collected corpus. This corpus is the first large-scale Chinese natural emotion corpus dealing with multimodal and natural emotions.

The HEU Emotion [101] is currently the largest video-based AV emotion recognition corpus in the wild. It consists of 19,004 video clips. The speakers in the video recordings include Asians, Africans, and Caucasians. The recordings of the HEU corpus were taken under uncontrollable natural conditions and included a variety changes in multi-view face and body posture, local occlusion, illumination, and expression intensity.

The AVSP-ESD [102] corpus consists of 13,285 AV files collected from movies, TV shows, and YouTube containing speech and non-speech. It covers twelve different natural emotions (boredom, neutral, happiness, sadness, anger, fear, surprise, disgust, excitement, pleasure, pain, disappointment) with two levels of intensity. It is also a multi-language corpus that includes recordings in Chinese, English, French, Russian, and other languages.

CelebV-HQ [103] is a large-scale video facial attributes corpus. It consists of 35,666 video clips with a resolution of 512 × 512 at least, involving 15,653 speakers. All videos are labeled manually with 83 facial attributes, covering appearance, action, and eight different emotions. The corpus has a variety of age, ethnicity, brightness stability, motion smoothness, head pose diversity, and data-quality parameters.
The MimicME [104] corpus contains recordings of 4700 speakers. The recordings are in the form of 4-Dimensional (4D) videos of subjects displaying a multitude of facial behaviors, resulting in over 280,000 video files. The corpus has great diversity in age, gender, and ethnicity. Not all the files have been categorized with the corresponding emotions.

The EmoSet [105] corpus aims at predicting people’s emotional responses to visual stimuli. EmoSet consists of 3.3 million files in total, with only 118,102 of these files labeled by human annotators. Nevertheless, it is by far the largest existing emotional corpus. EmoSet is well-balanced between different emotion categories.

Based on the conducted analysis, we can conclude that most of the existing corpora are not suitable for the task of emotional speech lip-reading. Many corpora have been recorded for the task of emotion recognition, rather than SR. Due to this, they do not have the required vocabulary and the amount of repetitions per phrase with target emotions. From this list, we selected two publicly available emotional corpora suitable for SER and VSR evaluations. Only these two emotional corpora among those available in the scientific literature have a sufficient amount of data to train NN-based lip-reading systems, as well as the number of required parameters, such as number of speakers, vocabulary size, number of repetitions, emotionally balanced data, etc. The details of the selected corpora are presented below.

### 3.1. RAVDESS Corpus

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) corpus [62] comprises 4904 AV recordings of 24 actors, evenly divided between genders (12 females and 12 males). The corpus was recorded in laboratory conditions. These actors articulate with neutral North American accent two phrases (see Table 2) that are lexically matched. The phrases are spoken by actors with two vocal channels (speech, song) and two emotional intensities (normal, strong). The speech vocal channel represents eight emotions (Neutral (NE), Happy (HA), Angry (AN), Sad (SA), Disgust (DI), Fearful (FE), Calm (CA), Surprised (SU)), while the song vocal channel provides six emotions (DI and SU are excluded).

<table>
<thead>
<tr>
<th>Label</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAT</td>
<td>Kids are talking by the door.</td>
</tr>
<tr>
<td>DAS</td>
<td>Dogs are sitting by the door.</td>
</tr>
</tbody>
</table>

We divided the recordings into three samples, taking into account speaker independence: Train (first eighteen speakers or 75%), Development (next two speakers or 8%), Test (remaining four speakers or 17%). The distribution of recordings per emotion in three subsets is presented in Figure 1, while Figure 2 illustrates the distribution of recordings per phrase.

![Figure 1](image-url). Distribution of recordings per emotion in the Train/Development/Test subsets of the RAVDESS corpus.
3.2. CREMA-D Corpus

The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [94] comprises 7442 AV recordings of 91 actors, divided between genders (43 females and 48 males). The corpus was recorded in laboratory conditions. These actors uttered twelve phrases (see Table 3). The phrases were spoken by actors with three emotional intensities (low, medium, high) and six emotions (NE, HA, AN, SA, DI, FE). Neutral phrase recordings are not presented with different intensities.

![Distribution of recordings per phrase in the Train/Development/Test subsets of the RAVDESS corpus.](image1)

**Figure 2.** Distribution of recordings per phrase in the Train/Development/Test subsets of the RAVDESS corpus.

<table>
<thead>
<tr>
<th>Label</th>
<th>Phrase</th>
<th>Label</th>
<th>Phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFA</td>
<td>Don’t forget a jacket.</td>
<td>IWW</td>
<td>I wonder what this is about.</td>
</tr>
<tr>
<td>IEO</td>
<td>It’s eleven o’clock.</td>
<td>MTI</td>
<td>Maybe tomorrow it will be cold.</td>
</tr>
<tr>
<td>IOM</td>
<td>I’m on my way to the meeting.</td>
<td>TAI</td>
<td>The airplane is almost full.</td>
</tr>
<tr>
<td>ITH</td>
<td>I think I have a doctor’s appointment.</td>
<td>TIE</td>
<td>That is exactly what happened.</td>
</tr>
<tr>
<td>ITS</td>
<td>I think I’ve seen this before.</td>
<td>TSI</td>
<td>The surface is slick.</td>
</tr>
<tr>
<td>IWL</td>
<td>I would like a new alarm clock.</td>
<td>WSI</td>
<td>We’ll stop in a couple of minutes.</td>
</tr>
</tbody>
</table>

Table 3. Phrases and their labels in the CREMA-D corpus.

![Distribution of recordings per emotion in the Train/Development/Test subsets of the CREMA-D corpus.](image2)

**Figure 3.** Distribution of recordings per emotion in the Train/Development/Test subsets of the CREMA-D corpus.

We divided the data into three subsets, taking into account the even distribution by speakers’ age and gender, as well as speaker independence: Train (70%), Development (10%), and Test (20%). The distribution of recordings per emotion in the three subsets is presented in Figure 3, while Figure 4 illustrates the the distribution of recordings per phrase.
4. Methodology

We consider three different strategies for emotional speech lip-reading and compare them to the classical one (lip-reading model trained on emotionally neutral recordings). Our approach consists of visual speech data pre-processing followed by two main levels (see Figure 5). At the first level, the emotion recognition task is solved; at the second level, depending on the emotion, a specific model is used for VSR. The emotional strategies are divided into: (1) 6 emotions, (2) 3 valences, and (3) a binary (emotional/neutral data) category. In this section, we introduce our visual speech emotion recognition part based on an EMO-3DCNN-GRU network. We then describe our lip-reading model based on 3DCNN-BiLSTM architecture. Both models are trained specifically for six, three, and two emotion classes.

![Figure 4. Distribution of recordings per phrase in the Train/Development/Test subsets of the CREMA-D corpus.](image)

![Figure 5. EMOLIPS general processing flow in terms of three emotional strategies: (a) 6 emotions, (b) 3-level valence, (c) binary.](image)

All three strategies of the EMOLIPS approach shown in Figure 5 work on the same principle. The input data of the approach are a visual signal $V = \{F^0, F^2, \cdots, F^{(K-1)}\}$ consisting of $K$ frames. A face landmark detection algorithm [15] is used to detect the face and lip region images. The face and lip images are normalized, scaled to the correct size, and written to the corresponding lists $I_F = \{I_F^0, I_F^{step}, \cdots, I_F^{(M-1)-step}\}$ (for face images) and $I_L = \{I_L^0, I_L^{step}, \cdots, I_L^{(K-1)}\}$ (for lip images). For emotion recognition, only five Frames per Second (FPS) are analyzed. The step value range depends on the video FPS; $M$ is the number of frames left after reduction. Depending on the emotional strategy, the facial images are fed into one of three models: (1) an emotional model $EMO_6^{classes}$ for 6-emotion recognition, (2) an emotional model $EMO_3^{classes}$ for for 3-valence recognition, and (3) an emotional model $EMO_2^{classes}$ for binary classification. The emotional model predicts an emotional class $P_{emo}$. According to the obtained label, the lip images are fed into a specific lip-reading model $LIP_{emo}$ trained on phrases spoken with the predicted emotion.
Finally, the lip-reading model predicts a phrase class $P_{\text{phrase}}$ presented on the visual signal. The algorithm of the proposed EMOLIPS approach is presented in Algorithm 1.

**Algorithm 1 EMOLIPS Approach**

1: procedure PHRASEPREDICTION($V$, $S$)  
2: $V \leftarrow \{F^0, F^2, ..., F^{(K-1)}\}$ \hfill $\triangleright$ Input visual signal  
3: Forming the list of required frame numbers $F_{\text{red}} = \{0, \text{step}, ..., \text{step} \cdot (i - 1)\}$, where \text{step} is a frame decimation step; \text{i} ranges from 0 to $M$, the number of frames left after reduction.  
4: for $j \leftarrow 0$ to $K$ do  
5: \hspace{1em} Detect face $I^j_F$ and lip $I^j_L$ images in $F^j$ using [15]  
6: \hspace{1em} Reshape lip images $I^j_L = \text{Reshape}(I^j_L, \text{size}_L)$ \hfill $\triangleright$ $\text{size}_L$ is the correct size of a lip  
7: \hspace{1em} Normalization of lip images $I^j_L = \text{Norm}_{\text{Lips}}(I^j_L)$  
8: Add lip image $I^j_L$ to $I_L$  
9: \hspace{1em} if $j$ in $F_{\text{red}}$ then  
10: \hspace{2em} Reshape lip images $I^j_F = \text{Reshape}(I^j_F, \text{size}_F)$ \hfill $\triangleright$ $\text{size}_F$ is the correct size of a face  
11: \hspace{2em} Normalization of lip images $I^j_F = \text{Norm}_{\text{Faces}}(I^j_F)$  
12: \hspace{2em} Add lip image $I^j_F$ to $I_F$  
13: \hspace{1em} end if  
14: end for  
15: if $S$ is 6-Emotions strategy then \hfill $\triangleright$ $S$ refers to strategy  
16: Predict emotional class $P_{\text{emo}} = \text{argmax}(\text{EMO}_6\text{classes}(I_F))$  
17: else if $S$ is 3-level Valence strategy then  
18: Predict emotional class $P_{\text{emo}} = \text{argmax}(\text{EMO}_3\text{classes}(I_F))$  
19: else \hfill $\triangleright$ Binary strategy  
20: Predict emotional class $P_{\text{emo}} = \text{argmax}(\text{EMO}_2\text{classes}(I_F))$  
21: end if  
22: Predict phrase class $P_{\text{phrase}} = \text{argmax}(\text{LIP}_{\text{emo}}(I_L))$  
23: return $P_{\text{phrase}}$  
24: end procedure

4.1. Data Pre-Processing

Both VER and lip-reading models require a certain pre-processing of the raw video speech data. Identifying the face ROI on each video sequence is the first step. We use the MediaPipe open-source library to detect a face ROI in each frame [15]. For both classification tasks, we divide the videos into 2-second segments; the final prediction is the maximum sum of the probability vectors of all segments. Then we consider further pre-processing that is different for each model.

We apply an emotion recognition pre-processing pipeline similar to [107] that includes: (1) video downsampling to five FPS (ten frames per 2-second segment) to ensure the same processing conditions for recurrent NNs in terms of temporality; (2) channel normalization of the face image; (3) resizing the image to $224 \times 224$ pixels. Lip-reading pre-processing includes detection of a mouth ROI with the same approach (MediaPipe) on each frame and cropping it. We resized the mouth image to $88 \times 88 \times 3$. In order to maintain the mouth proportions, we pad the image to the desired size with the average pixel value along the vertical of the image itself. We then use a linear normalization technique. We do not apply any FPS downsampling at this stage and use all 60 frames for a 2-second segment. This data pre-processing showed the highest performance results in our related research [20]; we keep it unchanged in the current research.

4.2. Visual Emotion Recognition

For efficient emotion recognition, we use the open-source static 2DCNN model proposed in Ref. [107]. This model has confirmed its performance in the emotion recognition
task in comparison to other open-source models [108]. Similar to article [109], we convert the 2DCNN model based on ResNet50 architecture to a 3DCNN model with the addition of recurrent layers to capture changes in facial expression over time. However, unlike previous SOTA approaches [107,109], we extract features from different residual blocks of the CNN. We freeze the weights up to the 2nd residual block, and we fine-tune the weights of the 3rd and 4th blocks in the new research corpora. This approach allows us to combine the features of different scales in order to capture both larger and smaller facial-feature changes, as actively employed in object detection [110]. We use GRU as RNN, since it requires less computational effort and is not inferior in performance to LSTM [111]. Each GRU layer extracts a feature vector of size 128. To fuse the contribution of each feature vector within the emotion classification, we use the gated fusion mechanism [109,112], which is calculated as follows:

\[
g_a = \tanh(W_a a),
\]

\[
g_b = \tanh(W_b b),
\]

\[
z = \sigma(W_z [a, b]),
\]

\[
g = z \odot g_a + (1 - z) \odot g_b,
\]

where \(\tanh(\cdot)\) is the tanh activation function and \(W_a, W_b\) and \(W_z\) are the weight matrices for \(a, b\) and \(z\) feature vectors, respectively. The weight matrices have a size of \(N \times U\), where \(N\) refers to the number of features in a feature vector and \(U\) to the number of output neurons, which is equal to 64. \(g_a\) and \(g_b\) are gated representations of the \(a\) and \(b\) feature vectors, respectively. \(\sigma(\cdot)\) is the sigmoid activation function, \(\odot\) is the element-wise product, and \(g\) is the final feature vector for the following classification.

4.3. Lip-Reading

We use the 3DCNN [42] and BiLSTM to tackle emotional speech lip-reading (see the right part of Figure 6).

Figure 6. EMO-3DCNN-GRU model (left) and LIP-3DCNN-BiLSTM model (right).
Unlike other emotional models, this model is based on the ResNet18 [42] architecture. We use a simpler model architecture because lip recognition requires analyzing each frame of a video, which increases the computational effort. Additionally, the region covering the lips within each frame is smaller compared to the whole face, so there is no need for a more complex, deeper model. However, since the sequence of analyzed frames in the lip-reading model is six times greater than in the emotional model, we use a BiLSTM. The BiLSTM network consists of two BiLSTM layers with 512 consecutive neurons in each layer. Our model was previously trained and achieved SOTA results for lip-reading on the LRW corpus [8].

Three-dimensional CNN inputs an image that passes through the first 3D convolution layer and the pooling layer, then four residual blocks with 3D convolution layers follow, each of which is repeated twice. The 3D average pooling and reshape layers are next. For each image in a sequence of length \( T \) equal to 60 frames, the 3DCNN returns a feature vector \( r_i = [f_i^1, \ldots, f_i^N] \), where \( N \) is the number of features and \( i \in \{1, \ldots, T\} \). This is followed by the first BiLSTM layer of the sequence-to-sequence type, which consists of forward and backward LSTMs (\( \rightarrow LSTM_1 \) and \( \leftarrow LSTM_1 \), respectively). It returns a new feature matrix \( h \) by concatenating the output of the two LSTMs. The last BiLSTM layer is of the sequence-to-one type; it takes the feature matrix \( h \) as input and returns the final feature vector \( s \) for the last time-step of the sequence to perform the next classification. In general, the operation of the model can be represented by the following equations:

\[
\begin{align*}
\overrightarrow{h_i} &= LSTM^1_i(r_i), \\
\overleftarrow{h_i} &= LSTM^1_i(r_i), \\
\overrightarrow{h} &= \left[ \left[ \overrightarrow{h'}, \overleftarrow{h'} \right], \ldots, \left[ \overrightarrow{h'}, \overleftarrow{h'} \right] \right], \\
\overrightarrow{s_i} &= LSTM^2_i(h), \\
\overleftarrow{s_i} &= LSTM^2_i(h), \\
s &= \left[ \overrightarrow{s}, \overleftarrow{s} \right].
\end{align*}
\]

5. Experimental Results

To develop and research the EMOLIPS approach, we conducted a series of experiments. Initially, we trained three emotion recognition models. Then, we trained several lip phrase recognition models. Finally, we assessed the performance of the proposed EMOLIPS approaches by applying the trained models.

We assessed the experimental results using three metrics: the Mean Accuracy (mAcc) demonstrates the assessment of phrase recognition in the context of emotion/valence class; Acc demonstrates phrase recognition performance without taking into account emotions/valency; Unweighted Average Recall (UAR), Acc, and F1-measure (F1) are used to evaluate emotion recognition by the EMO-3DCNN-GRU model.

We trained the emotional models using the Stochastic Gradient Descent (SGD) optimizer with the Cosine Annealing Warm Restarts [113] scheduler for 100 epochs and five restart cycles. The learning rates varied from 0.001 to 0.0001. The lip models were trained using the Adam optimizer with a learning rate of 0.0001. The parameters of the models and of the training were selected experimentally using grid search.

5.1. Visual Emotion Recognition

Emotion recognition models were intended for specific tasks: (1) six-emotion recognition, (2) three-level valence recognition, and (3) binary emotion recognition. We excluded
recordings of phrases with the CA and SU emotions from the RAVDESS research corpus, as they are not present in the CREMA-D corpus. Additionally, we conducted a tenfold cross-validation on both research corpora including all recordings of phrases to allow for a comparison of our proposed emotion recognition approach with SOTA approaches.

The performance measures for each emotional strategy across the different corpora are presented in Table 4. The results obtained from training on two research corpora simultaneously are also included. As expected, a reduction in the number of classes resulted in higher average performance measures. However, combining the two corpora did not enhance the average performance measures for each emotional strategy. This phenomenon is likely due to variations in emotional expressions within the research corpora, which could complicate and confuse the model.

Figure 7 shows the confusion matrices reflecting the UAR values for the emotions from Table 4. The multi-corpus confusion matrices replicate patterns that are comparable to the CREMA-D corpus. Primarily, this is because the CREMA-D corpus offers a larger volume of training data in comparison to RAVDESS. Overall, in the case of balanced classes, the RAVDESS corpus displays accurate recognition of all six emotions with a minimum Acc (recall) of 78%, while the CREMA-D corpus varies between 44% and 95%. The other two strategies demonstrate class imbalance increases; however, the average Acc across all classes remains no lower than 69%. This result was attained by implementing logarithmic class weighting, as suggested in the research conducted by the authors of Ref. [109].

<table>
<thead>
<tr>
<th>Strategy</th>
<th>RAVDESS</th>
<th>CREMA-D</th>
<th>Multi-Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UAR</td>
<td>F1</td>
<td>Acc</td>
</tr>
<tr>
<td>6 Emotions</td>
<td>84.2</td>
<td>83.9</td>
<td>83.4</td>
</tr>
<tr>
<td>3-level Valence</td>
<td>93.8</td>
<td>91.2</td>
<td>95.0</td>
</tr>
<tr>
<td>Binary</td>
<td>92.7</td>
<td>91.6</td>
<td>96.9</td>
</tr>
</tbody>
</table>

Table 4. Comparison of the emotion recognition performance measures for three strategies.

![Figure 7](image-url)

Figure 7. Confusion matrices for 6 emotions, 3 valences, and binary classification obtained for Test subsets: (a) RAVDESS corpus, (b) CREMA-D corpus, (c) multi-corpus.

Additionally, we compared the Acc measures of our approach with SOTA approaches presented for the same corpora using a tenfold cross-validation scheme, as presented in
Table 5. The comparison highlights that our approach demonstrates at least 1% superiority over the SOTA for both corpora.

Table 5. Comparison of the emotion recognition Acc with SOTA in 10-fold cross-validation.

<table>
<thead>
<tr>
<th>Approach</th>
<th>MAX</th>
<th>MEAN</th>
<th>STD</th>
<th>Approach</th>
<th>MAX</th>
<th>MEAN</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human [96]</td>
<td>–</td>
<td>75.0</td>
<td>–</td>
<td>Human [94]</td>
<td>–</td>
<td>58.2</td>
<td>–</td>
</tr>
<tr>
<td>MATER [114]</td>
<td>–</td>
<td>52.0</td>
<td>–</td>
<td>MATER [114]</td>
<td>–</td>
<td>60.0</td>
<td>–</td>
</tr>
<tr>
<td>VO-LSTM [115]</td>
<td>–</td>
<td>60.5</td>
<td>–</td>
<td>VO-LSTM [115]</td>
<td>–</td>
<td>66.8</td>
<td>–</td>
</tr>
<tr>
<td>Our</td>
<td>66.8</td>
<td>61.7</td>
<td>3.9</td>
<td>Our</td>
<td>76.6</td>
<td>67.9</td>
<td>5.1</td>
</tr>
</tbody>
</table>

The best result for each corpus is highlighted in bold.

5.2. Lip-Reading

We conducted four experiments with the lip-recognition model. In the first experiment, we trained six models based on phrases spoken with one of six emotions. We provide the results of phrase recognition Acc obtained by all models in the context of each emotion class separately. We also trained the models based on phrases spoken with each emotion. But since in this case the Train subset was increased by six times, we divided it into six evenly distributed folds by gender, age, phrase labels, and speakers. The first experiment results are presented in Table 6.

Table 6. Comparison of the phrase recognition Acc for 6-emotion strategy. TM is the training model.

<table>
<thead>
<tr>
<th>TM</th>
<th>NE</th>
<th>HA</th>
<th>AN</th>
<th>SA</th>
<th>DI</th>
<th>FE</th>
<th>mAcc</th>
<th>NE</th>
<th>HA</th>
<th>AN</th>
<th>SA</th>
<th>DI</th>
<th>FE</th>
<th>mAcc</th>
<th>NE</th>
<th>HA</th>
<th>AN</th>
<th>SA</th>
<th>DI</th>
<th>FE</th>
<th>mAcc</th>
<th>NE</th>
<th>HA</th>
<th>AN</th>
<th>SA</th>
<th>DI</th>
<th>FE</th>
<th>mAcc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NE</td>
<td>93.8</td>
<td>79.7</td>
<td>85.2</td>
<td>89.1</td>
<td>92.2</td>
<td>70.3</td>
<td>85.1</td>
<td>92.1</td>
<td>75.0</td>
<td>95.3</td>
<td>82.2</td>
<td>72.2</td>
<td>88.2</td>
<td>84.2</td>
<td>72.3</td>
<td>55.5</td>
<td>73.1</td>
<td>69.1</td>
<td>62.5</td>
<td>68.8</td>
<td>66.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA</td>
<td>82.8</td>
<td>93.0</td>
<td>87.5</td>
<td>85.9</td>
<td>91.4</td>
<td>93.8</td>
<td>89.1</td>
<td>78.9</td>
<td>95.5</td>
<td>90.5</td>
<td>82.9</td>
<td>83.0</td>
<td>89.9</td>
<td>86.8</td>
<td>59.1</td>
<td>69.0</td>
<td>62.4</td>
<td>63.4</td>
<td>64.2</td>
<td>66.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AN</td>
<td>84.4</td>
<td>75.8</td>
<td>89.1</td>
<td>82.8</td>
<td>79.7</td>
<td>68.8</td>
<td>80.1</td>
<td>85.1</td>
<td>76.5</td>
<td>93.1</td>
<td>70.2</td>
<td>67.5</td>
<td>82.7</td>
<td>79.2</td>
<td>67.3</td>
<td>65.6</td>
<td>72.1</td>
<td>60.7</td>
<td>61.4</td>
<td>65.4</td>
<td>65.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>89.1</td>
<td>92.2</td>
<td>68.0</td>
<td>89.8</td>
<td>85.9</td>
<td>85.9</td>
<td>85.2</td>
<td>82.0</td>
<td>86.7</td>
<td>89.8</td>
<td>86.5</td>
<td>85.7</td>
<td>91.4</td>
<td>87.0</td>
<td>71.0</td>
<td>61.5</td>
<td>70.9</td>
<td>70.9</td>
<td>65.2</td>
<td>71.2</td>
<td>68.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DI</td>
<td>73.4</td>
<td>69.5</td>
<td>68.0</td>
<td>70.3</td>
<td>73.4</td>
<td>70.3</td>
<td>70.8</td>
<td>81.1</td>
<td>85.8</td>
<td>88.2</td>
<td>84.5</td>
<td>84.4</td>
<td>89.0</td>
<td>85.5</td>
<td>62.7</td>
<td>64.9</td>
<td>70.7</td>
<td>62.1</td>
<td>66.3</td>
<td>66.2</td>
<td>65.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE</td>
<td>71.9</td>
<td>68.0</td>
<td>63.3</td>
<td>87.5</td>
<td>79.7</td>
<td>78.1</td>
<td>74.8</td>
<td>85.1</td>
<td>86.0</td>
<td>92.3</td>
<td>85.1</td>
<td>82.6</td>
<td>92.3</td>
<td>87.2</td>
<td>67.9</td>
<td>70.2</td>
<td>71.1</td>
<td>66.6</td>
<td>63.5</td>
<td>72.1</td>
<td>68.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mAcc</td>
<td>82.6</td>
<td>79.7</td>
<td>76.9</td>
<td>84.2</td>
<td>83.7</td>
<td>77.9</td>
<td>86.2</td>
<td>84.1</td>
<td>84.3</td>
<td>91.5</td>
<td>81.9</td>
<td>79.2</td>
<td>88.9</td>
<td>90.7</td>
<td>67.9</td>
<td>65.3</td>
<td>65.2</td>
<td>63.7</td>
<td>68.0</td>
<td>71.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>81.3</td>
<td>91.4</td>
<td>74.2</td>
<td>92.2</td>
<td>93.0</td>
<td>75.0</td>
<td>84.5</td>
<td>93.0</td>
<td>89.9</td>
<td>92.5</td>
<td>86.5</td>
<td>83.5</td>
<td>86.8</td>
<td>87.7</td>
<td>66.5</td>
<td>70.1</td>
<td>71.8</td>
<td>64.0</td>
<td>66.2</td>
<td>69.3</td>
<td>68.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The performance measure values of 6-emotion strategy for each corpus are highlighted in bold.

Table 6 demonstrates the results of the six-emotions strategy, as well as the classical (trained only on emotionless NE phrases) and emotional (trained on all emotional phrases with six emotions) approaches. In the case where the VSR approach was trained only on NE phrases (NE Train model), the mAcc is equal to 85.1%, 84.2%, and 66.9% for RAVDESS, CREMA-D, and multi-corpus data, respectively. In the case where the VSR approach was trained on all emotional phrases with six emotions (All Train model), the mAcc is equal to 84.5%, 88.7%, and 68.0% for RAVDESS, CREMA-D, and multi-corpus data, respectively. For our approach with the six-emotions strategy, we trained six different models for every emotion (see values on the diagonal) and the mAcc is equal to 86.2%, 90.7%, and 71.3% for RAVDESS, CREMA-D, and multi-corpus data, respectively. Using our approach, we achieved an average increase of 4% compared to the classical approach and 2% compared to the emotional approach across all corpora.

It is worth noting that, depending on the emotional training model, the values of mAcc (horizontally) vary. For the RAVDESS corpus, the mAcc varies from 70.8% to 89.1%, displaying an 18.3% contrast between the DI and HA training models. Conversely, for the CREMA-D corpus, mAcc shows a range of 79.2% to 87.2%, with an 8% deviation between the AN and FE training models. For the multi-corpus data, the mAcc increases from 65.4% to 68.5%, or by 3%, depending on whether the AN/DI or SA/FE training models are used. The mAcc (vertically) for emotional phrase recognition also varies regardless of the emotional training model used. For AN and SA spoken phrases in the RAVDESS corpus, the mAcc ranges from 76.9% to 84.2%. Meanwhile, the CREMA-D corpus shows an increase in mAcc from 79.2% to 91.5% for phrases uttered with DI and AN emotions. A
comparable trend is observed in the multi-corpus data, where mAcc changes from 63.7% to 71.2%. This phenomenon arises because all emotions elicit specific lip movements that enable differentiation between emotions. This variation in articulation impacts phrase recognition performance.

In the second experiment, we grouped all categorical emotions into classes according to their valence: Negative (NG) valence (includes AN, DI, FE, and SA emotions), NE valence, Positive (PO) valence (includes HA emotion). Since we obtained the models for the NE and PO valence classes in the first experiment, we only trained the models for the NG valence class. Since the size of the Train subset for NE valence exceeded the size of the Train subset for other valences by four times, we divided the Train subset into four folds, similarly to the experiment in the first experimental study. We report the experimental results in Table 7.

Table 7. Comparison of the phrase recognition Acc for 3-level valence strategy. TM is the training model.

<table>
<thead>
<tr>
<th>TM</th>
<th>RAVDESS</th>
<th>CREMA-D</th>
<th>Multi-Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NE</td>
<td>PO</td>
<td>NG</td>
</tr>
<tr>
<td>NE</td>
<td>93.8</td>
<td>79.7</td>
<td>86.2</td>
</tr>
<tr>
<td>PO</td>
<td>82.8</td>
<td>93.0</td>
<td>89.1</td>
</tr>
<tr>
<td>NG</td>
<td>95.3</td>
<td>88.3</td>
<td>90.6</td>
</tr>
<tr>
<td>mAcc</td>
<td>90.6</td>
<td>87.0</td>
<td>88.6</td>
</tr>
</tbody>
</table>

The results presented in Table 7 show that when emotions are grouped according to their valence, the mAcc values for recognition of twelve phrases are 92.5%, 92.6%, and 71.8% for RAVDESS, CREMA-D, and multi-corpus data, respectively. These values exceed mAcc using our six-emotion strategy by 6.3%, 1.9%, and 0.5% for the same respective corpora (where mAcc = 86.2%, 90.7%, and 71.3%, respectively; see Table 6). This result shows that all emotions of NG valence have similar facial features that differ from other valence classes.

In addition, in Table 8 we examine how the Acc of phrase recognition varies with respect to vocal channel (for the RAVDESS corpus) and emotional intensity (for the RAVDESS and CREMA-D corpora). In the case of NE valence in the RAVDESS corpus, strong emotional intensity is absent. Similarly, for NE valence in the CREMA-D corpus, low, medium and high intensities are missing. Table 8 shows that the Acc of phrase recognition is higher for the speech vocal channel than for the song one. Furthermore, for the RAVDESS corpus, the Acc of SR is higher for normal emotional intensity than for strong emotional intensity, with a difference of no more than 5.7%. On the other hand, for CREMA-D, emotional intensity plays a crucial role in phrase recognition, as the difference between different emotional valences and intensities is 62.9%. It can therefore be concluded that variations in articulation significantly distort the Acc of phrase recognition across different vocal channels and emotional intensities.

Table 8. Comparison of the phrase recognition Acc for 3-level valence strategy in terms of vocal channels and emotional intensities. TM is the training model. SP and SO refer to speech and song vocal channels, respectively. NR and ST refer to normal and strong emotional intensities, respectively. LO, MD, and HI refer to low, medium, and high emotional intensities, respectively.
In the third experiment, we grouped the emotions into two classes: NE and Emotional (EM) (all emotions excluding NE), i.e., we implement binary classification. The model for the NE class remains unchanged. For the EM class, we divided the Train subset into five folds. Similarly to the previous experimental studies, we present the maximal results obtained in only one of the folds in Table 9.

Tables 7 and 9 show that for the RAVDESS corpus and multi-corpus data, the Acc of phrase recognition is approximately the same for the three-level valence and binary strategies. For the CREMA-D corpus, the binary strategy shows a 1.4% reduction in mAcc (92.6% vs. 91.2%). Since the NG and PO valences have completely different facial features, the union of these two opposite valences results in an mAcc value drop.

In Table 10, we examine how the Acc of phrase recognition varies with respect to vocal channel and emotional intensity, similar to Table 8. The Acc of the vocal channel phrase recognition is higher than that of the speech one for emotionally spoken phrases. However, the recognition Acc is higher for normal-intensity phrases than for strong-intensity phrases, as in the three-level valence strategy on the RAVDESS corpus. Furthermore, for the CREMA-D corpus, the Acc of phrase recognition is approximately the same for different levels of intensity, but higher for medium intensity.

Table 9. Comparison of the phrase recognition Acc for binary strategy. TM is the training model.

<table>
<thead>
<tr>
<th>TM</th>
<th>RAVDESS</th>
<th>CREMA-D</th>
<th>Multi-Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NE</td>
<td>EM</td>
<td>mAcc</td>
</tr>
<tr>
<td>NE</td>
<td>93.8</td>
<td>84.7</td>
<td>89.3</td>
</tr>
<tr>
<td>EM</td>
<td>93.8</td>
<td>91.8</td>
<td>92.8</td>
</tr>
<tr>
<td>mAcc</td>
<td>93.8</td>
<td>88.3</td>
<td>92.8</td>
</tr>
</tbody>
</table>

The performance measure values of the binary strategy for each corpus are highlighted in bold.

Table 10. Comparison of the phrase recognition Acc for binary strategy in terms of vocal channels and emotional intensities. TM is the training model. SP and SO refer to speech and song vocal channels, respectively. NR and ST refer to normal and strong emotional intensities, respectively. LO, MD, and HI refer to low, medium, and high emotional intensities, respectively.

<table>
<thead>
<tr>
<th>TM</th>
<th>RAVDESS</th>
<th>CREMA-D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NE</td>
<td>EM</td>
</tr>
<tr>
<td>NE</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td>EM</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td>mAcc</td>
<td>93.8</td>
<td>87.7</td>
</tr>
</tbody>
</table>

5.3. EMOLIPS

The values of mAcc = 86.2%, 90.7%, and 71.3% (six-emotion strategy); 92.5%, 92.6%, and 71.8% (three-level valence strategy); and 93.8%, 91.2%, and 72.0% (binary strategy) have been achieved with the assumption that an emotional model makes predictions with a high Acc. However, at the moment there are no such models that make such predictions with a high Acc on different corpora. We therefore applied the three emotional models proposed in this article (see Section 5.1) as the first level of the EMOLIPS approach.

The result of recognition of twelve phrases (see Table 11) taking into account the distribution of emotions using the EMO-3DCNN-GRU model and two-level recognition was achieved relative to two baselines (a model trained only on neutral phrases and all emotional phrases). The experimental results demonstrate that the EMOLIPS approach with a six-emotion strategy does not yield any improvement over the baselines for the RAVDESS and CREMA-D corpora. For the CREMA-D corpus, the maximum Acc of 90.9% was achieved with a three-level valence strategy, which is 6.7% higher than the NE baseline and 2.2% higher than the All baseline. For the RAVDESS and multi-corpus data, the binary
strategy outperformed the three-level valence strategy by an average of 1%, resulting in a maximum phrase recognition Acc of 91.9% and 71.7%, respectively.

Table 11. Comparison of mAcc of our two-level approaches with baseline (model trained only on NE phrases and all emotional phrases).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Baseline (NE)</th>
<th>Baseline (All)</th>
<th>6 Emotions</th>
<th>3-Level Valence</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAVDESS</td>
<td>85.1</td>
<td>84.5</td>
<td>83.8</td>
<td>90.6</td>
<td>91.9</td>
</tr>
<tr>
<td>CREMA-D</td>
<td>84.2</td>
<td>88.7</td>
<td>86.4</td>
<td>90.9</td>
<td>90.4</td>
</tr>
<tr>
<td>Multi</td>
<td>66.9</td>
<td>68.0</td>
<td>68.8</td>
<td>69.9</td>
<td>71.7</td>
</tr>
</tbody>
</table>

The best result for each corpus is highlighted in bold.

It is worth noting that our EMOLIPS approach, based on the CREMA-D corpus, outperformed our previous result by 23.1% [20]. Compared to our previous approach, we have not only introduced a two-stage phrase recognition based on lip-reading, but also improved the phrase recognition model itself.

6. Conclusions

In this article, we proposed the EMOLIPS approach for automatic emotional speech lip-reading. This two-level approach based on visual data processing was motivated by human perception and the recent developments in multimodal DL. The proposed approach uses visual speech data to determine the type of speech emotion. The speech data are then processed using one of the emotional lip-reading models trained from scratch. This essentially resolves the multi-emotional lip-reading issue associated with most real-life scenarios. We implemented these models as a combination of EMO-3DCNN-GRU architecture for emotion recognition and 3DCNN-BiLSTM architecture for automatic lip-reading.

In addition, this article provided a detailed review of the recent advances in automated lip-reading and emotion recognition that have been developed over the last 5 years (2018–2023). In comparison to existing research, we mainly focused on the valuable progress brought with the introduction of DL to the field and skip the description of traditional approaches. Moreover, we presented an extensive experimental investigation that demonstrates how different emotions (happiness, anger, disgust, fear, sadness, and neutral), valence (positive, neutral, and negative) and binary (emotional and neutral) affect automatic lip-reading.

The evaluation on the RAVDESS and CREMA-D corpora demonstrated that the proposed approach achieved SOTA performance with up to 91.9% and 90.9% emotional phrase recognition Acc, respectively. We also trained three emotional models for specific tasks: (1) six-emotion recognition, (2) three-level valence recognition, and (3) binary emotion recognition. Our proposed emotional model outperformed SOTA results for emotion recognition by more than 1% in terms of the Acc using a tenfold cross-validation setup.

There are various potential areas for future research. These include improving the performance of emotional lip-reading models by using more complex model architectures and applying them to real-time applications in affective computing and Human-Computer Interaction (HCI). Additionally, it is vital to explore multimodal approaches that integrate lip-reading with other modalities such as audio and text. The results of this research are crucial for pushing the boundaries of emotion recognition and automatic lip-reading.

All source code and trained models are available on our GitHub repository at https://github.com/SMIL-SPCRAS/EMOLIPS (accessed on 25 October 2023).

Author Contributions: Conceptualization, E.R. and D.I.; methodology, E.R.; validation, E.R.; formal analysis, D.R.; investigation, E.R.; resources, D.I.; writing—original draft preparation, E.R. and D.I.; writing—review and editing, D.R. and D.I.; visualization, D.R., E.R. and D.I.; supervision, D.R.; project administration, D.R.; funding acquisition, D.I. All authors have read and agreed to the published version of the manuscript.
Funding: Sections 1, 3, 4, 5.2, 5.3 and 6 were financially supported by the Russian Science Foundation (No. 23-71-01056, https://rscf.ru/en/project/23-71-01056/, accessed on 25 October 2023), Section 5.1 was financially supported by a grant (No. MK-42.2022.4), and Section 2 was financially supported by the State research (Topic No. FFZF-2022-0005).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this study, we used two large-scale publicly available corpora: RAVDESS—https://zenodo.org/records/1188976 (accessed on 25 October 2023) and CREMA-D—https://github.com/CheyneyComputerScience/CREMA-D (accessed on 25 October 2023).

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

Abbreviations
The following abbreviations are used in this manuscript:

1D 1-Dimensional
2D 2-Dimensional
3D 3-Dimensional
4D 4-Dimensional
Acc Accuracy
AN Angry
ASR Audio Speech Recognition
AV Audio-Visual
AVSR Audio-Visual Speech Recognition
Bi Bidirectional
CA Calm
CNN Convolutional Neural Network
CREMA-D Crowd-Sourced Emotional Multimodal Actors Dataset
CV Computer Vision
DBN Deep Belief Network
DI Disgust
DL Deep Learning
DNN Deep Neutral Network
EM Emotional
F1 F1-measure
FC Fully Connected
FE Fearful
FPS Frames per Second
GELU Gaussian Error Linear Unit
GRU Gated Recurrent Unit
HA Happy
HCI Human-Computer Interaction
KL Kullback-Leibler
LN Layer Normalization
LOCO Leave-One-Corpus-Out
LRS Lip-Reading Sentences in-the-Wild dataset
LRW Lip Reading in the Wild dataset
LSTM Long-Short Term Memory
mAcc Mean Accuracy
ML Machine Learning
MLP Multilayer Perceptron
MSA Multi-head Self-Attention
MSTCN Multi-Scale Temporal Convolution Network
NE Neutral
NG Negative
NN Neutral Network
PO Positive
RAVDESS  Ryerson Audio-Visual Database of Emotional Speech and Song
RNN  Recurrent Neural Network
ROI  Region-of-Interest
RUME  Recurrence based Uni-Modality Encoder
STT  Speech-to-Text
SA  Sad
SER  Speech Emotion Recognition
SGD  Stochastic Gradient Descent
SOTA  State-of-the-Art
SR  Speech Recognition
STCNN  Spatio-Temporal Convolutional Neural Network
STFT  Short Time Fourier Transform
SU  Surprised
SVM  Support Vector Machine
TSN  Two-Stream Network
UAR  Unweighted Average Recall
VER  Visual Emotion Recognition
VGG  Visual Geometry Group
ViT  Vision Transformer
VSR  Visual Speech Recognition

References
24. Kansizoglu, I.; Misirlis, E.; Tsintotas, K.; Gasteratos, A. Continuous Emotion Recognition for Long-Term Behavior Modeling through Recurrent Neural Networks. *Technologies* 2022, 10, 59. [CrossRef]
34. Kumar, P.; Raman, B. A BERT Based Dual-Channel Explainable Text Emotion Recognition System. *Neural Netw.* 2022, 150, 392–407. [CrossRef]


49. Middya, A.I.; Nag, B.; Roy, S. Deep Learning Based Multimodal Emotion Recognition Using Model-Level Fusion of Audio-Visual Modalities. Knowl.-Based Syst. 2022, 244, 108880. [CrossRef]


78. Chern, J.S.; Zisserman, A. Lip Reading in the Wild. In Asian Conference on Computer Vision; Springer: Cham, Switzerland, 2016. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.