applied sciences

MDPI

Review

Deep Multimodal Emotion Recognition on Human Speech: A Review

Panagiotis Koromilas * and Theodoros Giannakopoulos

Institute of Informatics and Telecommunications, National Center for Scientific Research—Demokritos, 15310 Athens, Greece; tyianak@iit.demokritos.gr

* Correspondence: pakoromilas@iit.demokritos.gr; Tel.: +30-210-650-3175

Abstract: This work reviews the state of the art in multimodal speech emotion recognition methodologies, focusing on audio, text and visual information. We provide a new, descriptive categorization of methods, based on the way they handle the inter-modality and intra-modality dynamics in the temporal dimension: (i) non-temporal architectures (NTA), which do not significantly model the temporal dimension in both unimodal and multimodal interaction; (ii) pseudo-temporal architectures (PTA), which also assume an oversimplification of the temporal dimension, although in one of the unimodal or multimodal interactions; and (iii) temporal architectures (TA), which try to capture both unimodal and cross-modal temporal dependencies. In addition, we review the basic feature representation methods for each modality, and we present aggregated evaluation results on the reported methodologies. Finally, we conclude this work with an in-depth analysis of the future challenges related to validation procedures, representation learning and method robustness.

Keywords: multimodal emotion recognition; multimodal temporal learning; multimodal signal processing; affective computing; speech emotion recognition

Citation: Koromilas, P.; Giannakopoulos, T. Deep Multimodal Emotion Recognition on Human Speech: A Review. *Appl. Sci.* **2021**, *11*, 7962. https://doi.org/10.3390/app11177962

Academic Editor: Len Gelman

Received: 26 July 2021
Accepted: 26 August 2021
Published: 28 August 2021

1. Introduction

The task of recognizing emotions in multimodal speech signals is vital and very challenging in the context of human–computer or human–human interaction applications. In such applications, interaction is not only characterized by the content of the dialogues but by how the involved parts feel when expressing their thoughts. In addition, as humans, we employ three modalities in a coordinated manner, in order to convey our intentions: language modality (words, phrases and sentences), vision modality (gestures and expressions) and acoustic modality (paralinguistics and changes in vocal tones) [1]. Theories of language origin identify the combination of language and nonverbal behaviors (vision and acoustic modality) as the prime form of communication utilized by humans throughout evolution [2].

Yet, why is it important for speech emotion recognition applications to adopt multimodality in the core of their architecture? To answer this question, let us see how the brain's mechanism functions in order to perceive and understand communication: different regions independently process and understand different modalities [3,4], while, at the same time, these individual regions are interconnected using neural links to achieve multimodal information integration [5]. The coordination between the different modalities in human communication introduces view-specific and cross-view dynamics. View-specific dynamics refer to dynamics within each modality independent of other modalities [6]. In order for a multimodal model to function similar to the way our brain perceives multimodality, it should therefore satisfy both of the aforementioned requirements.

Based on the above, the two key challenges to address when learning from multimodal data are that: (i) models must learn the complex intra-modal and cross-modal interactions to predict emotional content [6], and (ii) trained models must be robust to unexpected missing or noisy modalities during testing [7].

The two traditional and straightforward approaches that have usually been adopted in early multimodal fusion works related to speech emotion recognition are based on the strategies of early and late fusion. Early fusion approaches focus on concatenating multimodal features mostly at the input level [1,8,9], i.e., by simply concatenating feature vectors from different modalities. This fusion approach does not allow the intra-modality dynamics to be efficiently modeled. This is due to the fact that inter-modality dynamics can be more complex at the input level and can dominate the learning process or result in overfitting. Late fusion, instead, trains unimodal classifiers independently and performs decision voting [10,11] or other classifier combination strategies. This also prevents the model from learning inter-modality dynamics in an efficient way, by making the assumption that a simple aggregation of individual decisions is capable of capturing the multimodal relationships.

This work explores a wide range of more complex and sophisticated techniques for combining multimodal and temporal information in the emotion recognition application domain. In Section 2, we present the most common feature representation techniques per modality. In Section 3, we propose a hierarchical categorization of the related works, with regards to the way they face the modeling of the temporal dynamics and across modalities. In particular, we propose three general categories, namely: (i) non-temporal architectures, i.e., approaches that simplify the modeling of the temporal dimension in both unimodal and multimodal interactions, by assuming simple statistical representations; (ii) pseudo-temporal architectures, which also assume an oversimplification of the temporal dimension, in either unimodal or multimodal interactions (while the first category simplifies in both ways); and (iii) temporal architectures, i.e., methods that try to capture both unimodal and cross-modal temporal dependencies. In Section 4, we present detailed evaluation results in the most widely adopted datasets for emotion recognition. Finally, in Section 5, we conclude our work and introduce the key future challenges, as they arise from analysis of this work.

2. Modalities and Feature Extraction

2.1. Audio

The audio modality of a speech signal consists of a discrete temporal sequence sampled at frequencies varying from 8 KHz to 44.1 KHz. As with all machine learning tasks, feature extraction of audio signals is also crucial in multimodal speech emotion recognition (SER), as the feature representations must be informative with respect to the emotional class labels of the classification task. A small utterance of, say, a 5 s duration consists of 40 K samples at an 8 KHz sampling rate. Since the audio signal is obviously voluminous to be used as a feature vector (not to mention the duration dependency), it is necessary to transform the initial data representation to a more compact one. Of course, in the deep learning era, some methodologies function directly on the raw audio signal. However, in most cases, audio feature extraction in multimodal SER is achieved through hand-crafted features extracted from the time and frequency domains.

2.1.1. Feature Calculation

The features that can be calculated lie in four categories: (i) prosodic features, (ii) spectral features, (iii) voice quality features and (iv) Teager energy operator (TEO)-based features.

1. Prosodic features:
   These types of features are long term and are easily preserved and explained by humans. Typical such features are mostly based on energy, duration, and fundamental frequency characteristics, with some examples being rhythm and intonation.
2. Time-domain features:
   Such features are directly extracted from the samples of the audio signal. Typical examples include: short-term energy, zero crossing rate and entropy of energy. Such attributes cannot be strongly correlated to affective content in speech signals; however, they can carry information related to the strength of the underlying emotions (arousal).

3. Spectral features:
   Such features are calculated in the frequency domain and are widely used since characteristics of the vocal tract are well represented in that space [12]. They are usually extracted from overlapping speech short-term frames of length 20 to 50 ms that are applied through a windowing operation on the original audio signal. For each short-term frame, the discrete Fourier transform (DFT) is computed to provide a representation of the distribution of the frequency content. Features such as the spectral centroid, spectral flux, spectral rolloff and spectral entropy have been widely used in traditional audio and music classification tasks and in some speech emotion recognition methods that use hand-crafted audio features [13]. However, the most effective local features are extracted using the cepstrum, which is the result of computing the inverse Fourier transform of the logarithm of the estimated signal spectrum. Taking the cepstral coefficients of powerful signal transformations results in efficient and robust features. Some of these transformations are listed below.
   - Mel-Frequency Cepstral Coefficients (MFCC):
     These coefficients represent the short term power spectrum of the speech signal and consist of the most widely used spectral features for emotion recognition [14]. Before calculating the cepstral coefficients, the signal is transformed using a Mel-filter bank on a number of sub-band energies [15].
   - Linear Prediction Cepstral Coefficients (LPCC):
     These are task-specific coefficients that, in some cases, capture the emotional information expressed through vocal tract characteristics. The cepstral coefficients are derived from linear prediction (LP) analysis, which uses the energy values of linearly arranged filter banks to capture the contribution of all frequency components of a speech signal. One major drawback of LPC-produced features is that they are exposed to noise and thus need a processing technique to avoid additive noise error [16].
   - Gammatone Frequency Cepstral Coefficients (GFCC): This is a method that uses a feature extraction approach close to the MFCC procedure. More specifically, instead of the Mel-filter bank, it uses the Gammatone filter-bank on a range of sub-band energies [17].
4. Voice quality features:
   The voice quality features define the qualities of the glottal source by producing features such as noise ratio (HNR), shimmer and jitter. These are extremely useful since according to [18], the emotional content and voice quality of speech are correlated.
5. Teager Energy Operator-based features:
   TEO was introduced in [19,20] and was based on the observation that under stressful conditions, there is a change in fundamental frequency and critical bands of the hearing process due to the distribution of harmonics. The operator created by [19] was adopted in [20], in order to quantify the energy from speech by using a nonlinear process. TEO-decomposed frequency modulation variation, normalized TEO auto-correlation and critical band-based TEO auto-correlation are three basic features produced by [21].

There is a variety of open-source robust and efficient libraries, such as [22–25], that can be used in order to extract the aforementioned audio features.

2.1.2. Audio Representation Learning

The development of deep learning has resulted in a variety of representation learning techniques for different modalities. Convolutional neural networks (CNNs) have been widely used to "learn" image-related features in a supervised manner. Therefore, given a way to map audio signals into 2D representations (images), deep audio features can be learnt through CNNs. The most straightforward and obvious approach to achieve this is through spectrograms. For example, in [26–28], spectrograms were used as image inputs to CNNs in order to train classifiers for emotion recognition, and since then, this approach

of combining spectrograms (or melgrams) with CNNs has been widely adopted for speech emotion recognition tasks.

Lately, a variety of representation learning techniques and architectures has been proposed in order to learn audio representations from the initial waveform. The SincNet [29], a network that replaces the one-dimensional convolution kernels with sinc filter functions on the audio signal, is a supervised learning technique in such a direction. On the other hand, unsupervised methods for speech representation learning from audio are rapidly evolving and produce efficient architectures such as wav2vec [30], wav2vec 2.0 [31] and the newly produced HuBERT (Hidden-Unit BERT) [32], which is the currently state-of-the-art architecture for such representations.

2.2. Text

Since it was first shown that word-vectors produced by neural networks capture syntactic regularities [33], there has been a variety of unsupervised architectures that were designed to learn vector space representations of words. GloVe (Global Vectors) [34] is one of the first models to produce robust word embeddings, followed by the FastText model [35], which also included sub-word information, and ELMo (Embeddings from Language Models) [36], where the notion of deep context-dependent word representations has been introduced. With the advent of transformer networks [37], the architectures that produce word embeddings were further enriched with GPT (generative pre-trained) 1, 2 and 3 [38–40]; BERT (bidirectional encoder representations from transformers) [41]; ALBERT (a lite BERT) [42], etc. These models are efficiently designed, trained on vast amounts of data and have great representation abilities which capture both word meaning and context.

At this point, it has to be noted that, in order to be easily compared to the baseline models, most architectures of multimodal emotion recognition are adopting the GloVe embeddings.

2.3. Visual

Though theoretically speaking body language plays a significant part in expressing one's emotions, most of the related datasets available in multimodal emotion recognition capture faces. Therefore, the only visual information that is usually employed for the task of emotion recognition is facial expressions. At least six characteristics that are common to all humans, i.e., morphology, symmetry, duration, speed of onset and the coordination of apexes and ballistic trajectory, are correlated to emotional state [43]. According to [44], there are three main dimensions of facial variation—morphology, complexion and dynamics—with the third one being the most important for emotion recognition [45].

In [46], the authors developed the Facial Action Coding System (FACS), a system for objectively measuring facial movement, which, based on an implementation introduced by [47], became the standard for face movement recognition. The basic values of FACS are action units (AUs), which are the fundamental actions of either specific muscles or groups of muscles and are combined in several sets and identified by a sequence of numbers containing codes about head movement, eye movement, visibility, gross behavior and overall codes.

Existing tools for automatic facial recognition, such as IntraFace [48] and FACET [49], implement FACS and give automated facial feature tracking, head pose estimation, facial attribute recognition and facial expression analysis from video.

3. Multimodal Temporal Approaches

3.1. Methods Requirements and Categorization

The rapid development of deep learning architectures over the last decade resulted in a variety of different ideas for a range of multimodal sequential problems. One major distinguishing factor between the proposed architectures is the way they handle the temporal dimension. More specifically, such problems include both unimodal and cross-modal temporal dynamics. The unimodal representations should keep track and be adjusted

according to both short and long term temporal dependencies. The extracted multimodal representations can also be computed either by combining the final unimodal representations or by learning cross-modal interactions across time.

The goal of this paper is to provide a detailed review of deep learning methods for multimodal temporal speech emotion recognition. Thus, let us first focus on what *"deep"*, *"multimodal"* and *"temporal"* mean, in terms of method requirements. In order for a method to fall into the category of deep multimodal temporal emotion recognition, it should satisfy the following criteria:

1. *Deep*: adopt a deep network architecture at some part of its recognition and/or representation pipeline;
2. *Temporal*: model both short and long term unimodal dynamics. This means that in each modality, the temporal variation of the involved features should be modeled, either through explicit usage of temporal deep architectures or even through simple and "hand-crafted" statistical calculations on the feature sequences;
3. *Multimodal*: at least two different modalities are used in the representation and/or recognition pipeline of the method. The way these modalities are combined can be as simple as adopting a late fusion approach or more sophisticated, e.g., learning common multimodal representations. Note that in this work, we mostly focus on the audio, text and visual modalities, but emotion recognition in general has adopted other types of information (e.g., wearables).

The aforementioned general requirements are considered the minimum for a deep temporal multimodal method. However, as we will see in the rest of this section, some nice-to-have characteristics of more robust and efficient methods are listed below:

- *Modeling of cross-modal interactions through time*: The individual modalities include distinct unimodal dynamics that need to be learnt across time. However, apart from the unimodal representations, the temporal interactions between modalities form a dynamic phenomenon that needs to be learned through time.
- *Learning a joint representation space*: The representation of each modality has different geometric and statistical properties, and thus a simple combination of the unimodal representations cannot always be effective. Instead, a mapping from all individual feature vectors to a joint representation space would lead to more efficient solutions.
- *Learning the temporal alignment between modalities*: Different sample rates across modalities frequently result in sequences of different lengths that represent different parts of the multimodal file. For this reason, the alignment cannot be limited to a temporal window of fixed length, since a cross-modal interaction may happen asynchronously. These two reasons make the methods that assume cross-modal aligned data (we will call them forced alignment methods in this paper) have several limitations. Instead, the temporal alignment across modalities needs to be learned from pure data sequences during the training time.

Based on the aforementioned requirements (either minimum or nice-to-have), we propose the following categorization for multimodal temporal architectures for emotion recognition, which is also illustrated in Figure 1:

- Non-Temporal Architectures (NTA): In these approaches, both unimodal and cross-modal interactions do not take into account any temporal interaction. A typical case of such representations would be the statistical representation of unimodal feature sequences followed by an early or late fusion technique. In these cases, temporal modeling is not performed, and temporal information only "participates" in the task through particular statistics computed over the short-term feature sequences. In other words, these approaches do not model temporal evolution of emotions through the architecture itself but through simple feature engineering.
- Pseudo Temporal Architectures (PTA): Architectures in this category model a subset of temporal dynamics and thus lie under one of the following categories:
  - Unimodal Temporal Architectures (UTA): Model unimodal temporal dependencies;

  - Cross-Modal Temporal Architectures (CTA): Model cross-modal temporal dependencies.
- Temporal Architectures (TA): Approaches of this type lie in both UTA and CTA, since both unimodal and cross-modal temporal dependencies can be captured. TA can be further separated according to the degree of freedom of the temporal multimodal interactions. More specifically, a very typical constraint of these methods is that most approaches model multimodal interactions in a finite temporal window. However, in natural multimodal language the temporal progress differs across modalities, since one change in the visual medium cannot be directly aligned to the word uttered at that exact timestamp. This led us to define a clear distinction between the following subsets of methods:
  - Forced-Aligned Temporal Architectures (FTA): Architectures that force multimodal alignment with the use of a finite temporal window.
  - Aligned Temporal Architectures (ATA): Architectures that are free to use long term multimodal temporal interactions and learn multimodal alignment.

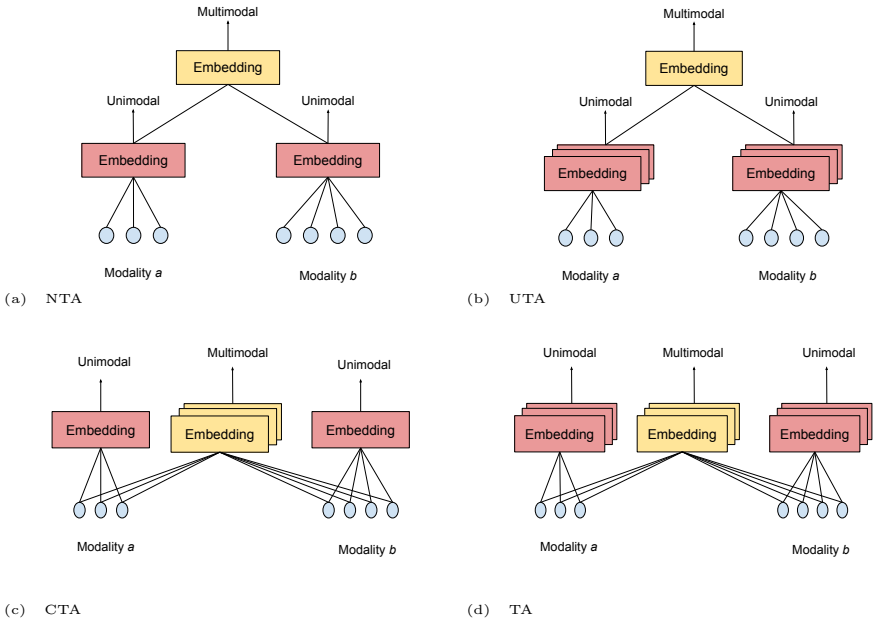

Figure 1. Deep temporal multimodal architectures categorization based on temporal learning properties. Parallel rectangles indicate that the learning process is performed on a temporal sequence of feature vectors, while standalone rectangles are used when learning is performed on a static—probably statistical—feature vector of predefined dimensions that represents the whole temporal sequence. The four listed categories are (**a**) NTA: Non-Temporal Architectures, (**b**) UTA: Unimodal Temporal Architectures, (**c**) CTA: Cross-Modal Temporal Architectures and (**d**) TA: Temporal Architectures. In (**a**,**b**) the multimodal representation is based on static vector representations, while in (**c**,**d**) it is calculated using the initial temporal sequence of each modality.

3.2. Non-Temporal Interactions

The methods that fall under the simplest among the above-defined categories adopt neither modality-specific temporal interactions nor cross-modality temporal interactions. These models simplify the problem by using feature-summarizing temporal observations for both cross-modal and intra-modal interactions.

An obvious starting point for combining multimodal representations is that of early fusion. One of the basic fusion ideas is applied on the task of persuasiveness prediction by [50] but evaluated on multimodal SER in a variety of latter works. In this architecture, the authors propose to represent the unimodal spaces with static vectors and train a feed-forward neural network for each modality. Both the resulting confidences c and

the complementary confidences $1 - c$ of each modality are passed through a deep neural network to perform the final inference. Due to this specificity, this procedure is called deep fusion (DF).

A work that tries to introduce the temporal dimension is proposed in [51]. Although the authors combine information from neighboring utterances, in order to capture the context of the session, they do not consider the temporal interactions between segments of the same utterance. More specifically, static feature vectors are extracted for every utterance and modality and are later passed through a contextual LSTM (long short term memory) which is applied on the sequence of utterance-based features and produces context-dependent representations. These representations are later concatenated and used as an input to a final contextual LSTM. In this way, the proposed architecture includes both unimodal and cross-modal utterance-level temporal interactions.

However, neither simple early nor late fusion techniques are capable of achieving robust multi-modal representations. A different approach is that of tensor fusion, which was introduced by [6]. The proposed tensor fusion network (TFN) concatenates each unimodal representation with the unit vector and then computes the outer product among each modality. The tensor fusion layer (i.e., outer product) explicitly captures the unimodal, bimodal and trimodal interactions using a three-fold Cartesian product from modality embeddings. The final representation is passed through an inference subnetwork which performs inference of the target value. Despite the representation power of [6], the problem of exponential dimensional growth is a drawback for this architecture. The work of [52] tries to tackle it by introducing the low rank fusion technique, which scales linearly in the number of modalities and has less parameters. The authors actually use matrix rank decomposition in order to parameterize the product of the final neural network with the fused tensor. Here, we have to note that both [6,52] do not include temporal representations for the acoustic and visual modalities, but instead consider the expected (average) vector for each utterance. That is the reason why, despite the completely temporal text representation, they cannot be counted as temporal architectures.

3.3. Pseudo-Temporal Interactions

These models simplify the problem by using feature-summarizing temporal observations for either cross-modal or intra-modal interactions. One of the first works that tries to effectively combine different feature spaces with the use of an attention mechanism is presented in [53]. The authors propose a deep architecture for the problem of speech emotion recognition, and thus they consider the two modalities of audio and text. For each medium, they implement a recurrent neural network (RNN) to capture the unimodal dynamics. In order to choose the most emotional words, the final audio embedding vector is used as a query for an attention mechanism against the hidden states of all text sequence elements. The representations of the selected "emotional" words are concatenated with the original audio vector to form the multimodal representation. However, the performance is not improved by the use of the attention layer.

The same authors extend their work and try to perform a more effective fusion technique. In [54], the multi-hop attention (MHA) mechanism is proposed and follows three steps: (a) MHA-1: predict the emotion based on the audio bidirectional-LSTM (BiLSTM) and query the output state (i.e., the representation of the audio) to the text BiLSTM in order to obtain a text representation H_1 that is associated with this very sound; (b) MHA-2: query H_1 to the audio BiLSTM in order to obtain the audio segments that are related to the selected words and represent them with H_2; and (c) MHA-3: query H_2 to the text BiLSTM for cross-check and obtain the overall representation H_3. However, the evaluation of this work states that the third attention step results in a worse performance compared to that achieved by MHA-2.

A mix of traditional fusion techniques seems to be effective when combined with strong segment feature representations, as shown in [55]. More specifically, this work tries to apply state-of-the art techniques for unimodal representations with the use of

ALBERT [42] embeddings for the text modality. As for the audio modality a spectrogram is calculated for every audio segment and passed through a CNN in order to learn the unimodal representation. The sequence of segmented features is the input to a Bi-LSTM and a multi-head self-attention layer in order to filter periods of aural silence. The resulting multimodal representations are firstly early-fused and passed through a prediction (bimodal) network and subsequently late-fused with the application of an ensemble classifier on the speech prediction, text prediction and bimodal prediction networks.

Designing a network that learns modality representations, instead of using existing modality embedding libraries, may result in more powerful representation spaces. In [9], the authors, in order to learn representations from visual sequences, propose the convolutional recurrent multiple kernel learning (CRMKL) network which uses a convolutional RNN, i.e., a CNN applied on the concatenation of pairs of consecutive images, with three different convolutional filter sizes. The text representations are learned with the application of a CNN on the text embeddings, while for the aural modality, the openSMILE [25] embeddings are used. The network is trained using a multiple kernel learning algorithm, and the final emotion prediction is inferred based on algorithmically selected features from the concatenation of the unimodal representations, and thus this specific architecture is categorized as UTA.

Attention mechanisms not only serve to produce better cross-modal representations, but they can also efficiently combine different feature extraction methods of the same modality. This is the case in [56], where the authors use two different, and widely used in text analysis, feature extraction techniques: (a) applying three Conv1D networks, with filter size 1, 2 and 3, respectively, on the text embedding sequence in order to extract a variety of n-word representations, and (b) using a Bi-RNN on the input embeddings and then apply method (a) to the output. In the aforementioned work, attention is applied along the two methods so as to get the best possible dynamics. Moreover, for the audio modality, a SincNet, a neural architecture that tends to replace CNNs in audio feature extraction [29], is used. Before concatenating the feature vectors, self-attention is calculated on the unimodal sequences in order to identify informative segments.

Despite the better feature extraction techniques and the use of attention to filter cross-modal information, there is a range of different ideas that deliver promising results. The first one is found in [57], where the authors break down the emotion recognition problem to three sub-tasks in order to extract direct and relative predictions. More specifically, they concatenate the different feature representations and separate emotion recognition to the following tasks: (a) In the multimodal local ranking task, the model is presented with two short segments randomly selected within a video and is tasked with determining whether there was an increase or decrease in emotional intensity. In this way, a binary classification task is created. (b) The second task is the *global ranking task*, which uses the previous results of local rankings to infer global emotion ranks using a Bayesian skill rating algorithm. (c) The third task involves *direct-relative fusion*. The global emotion ranks are incorporated with the raw multimodal inputs to estimate final emotion intensities.

The second is one of the crucial works on multimodal representation learning. The proposed architecture in [58] does not concentrate on finding an efficient way to fuse information. The multimodal cyclic translation network (MCTN) learns a joint representation space for different modalities. For two modalities it uses an encoder–decoder architecture to translate one modality to another by adopting a cyclic translation loss. The produced hidden embeddings serve as the cross-modal representation. For three modalities, a cyclic translation is formed: the resulting embeddings are the input to an encoder with the third modality being the target for the corresponding decoder. One more novelty in this approach is that once trained with multimodal data, only data from one source modality are needed during the inference time to construct both the joint representation and the label. This specificity makes this model suitable for production applications since it is more robust to noisy unimodal inputs.

All aforementioned works achieve efficiently modeling unimodal temporal dynamics. However, they lack cross-modal learning or fusion on the temporal level, since the cross-modal interaction is performed on the final representations instead of the sequence level. The MCTN [58] achieves modeling both temporal cross-modal learning and unimodal temporal dynamics but only for two modalities. When needing the third modality, there is no cross-modal temporal interactions for this specific input sequence.

In the pseudo-temporal architectures (PTA), modeling unimodal temporal representations (UTA) is more common. However, there are architectures that prefer to model cross-modal temporal interactions only (CTA). An example of such models is presented in [59], where the authors introduce the multimodal factorization model (MFM) that factorizes multimodal representations into multimodal discriminative factors and modality-specific generative factors. MFM tries to learn multimodal discriminative factors in order to perform inference, while it jointly learns unimodal generative factors that can be used to reconstruct missing or noisy information during testing time. To achieve this, there is an encoder for each modality (audio, text, video) and their fusion (audio + text + video), which creates the generative and discriminative factors. The corresponding decoders of the generative factors serve to approximate the input modalities, while the decoder of the discriminative factor is used for inference. For the encoder networks, the MFN architecture is used while LSTMs serve as decoders. Despite achieving competitive performance on six multimodal time series datasets, a big advantage of MFM is that reconstruction of missing modalities from observed modalities does not significantly impact discriminative performance.

3.4. Completely Temporal Interactions

This section presents an overview of the more temporal-specific methodologies that either capture modality-specific temporal interactions or cross-modality temporal interactions.

3.4.1. Temporally Unaligned across Modalities

One of the first ideas for capturing both unimodal and cross-modal temporal dynamics focused on using multimodal information to update the memory component of recurrent neural networks. For example, is has been proposed to adopt multimodal representation inside an LSTM's architecture in [60], where the authors have constructed the multi-view LSTM, which partitions the memory cell and the gates into regions corresponding to multiple modalities or views. This approach is based on two core ideas:

- The existence of one memory partition for each input view ensures that a view has its own internal dynamic.
- The memory partition of a specific view is flexible with regard to the way it integrates information from other views. This is achieved because the MV-LSTM allows four types of memory cells: (i) view-specific cells, which are affected by a hidden state from the same view; (ii) *coupled cells*, which are affected by a hidden state from other views; (iii) *fully-connected cells*, which are affected by both same-view and other-view hidden states; and (iv) *input-oriented cells*, which are not affected by either the same-view or other view hidden states.

This idea shaped a new line of works that tried to efficiently create multimodal memory components. The multi-attention recurrent network (MARN) by [61] is a good step towards capturing cross-modal temporal interactions. It has two basic components: the LSTHM, i.e., an LSTM with hybrid memory, and the multi-attention block (MAB). For each modality, a different LSTHM is used. In this network, the hybrid memory of each modality takes the modalities' hidden states along with the corresponding output of MAB in order to include cross-modal interactions related to the LSTHM's target modality. MAB takes as an input the concatenated hidden states from all modalities. It tries to find K (arbitrary number) cross-modal interactions by passing this vector in a pipeline of a neural network, a softmax function and dot-product with the input, which seems similar to an intra-vector attention. The output is a K vector matrix which is passed to three neural

networks (one for each modality) for dimensionality reduction. Obviously, this technique not only produces cross-modal interactions but intra-modal ones, too.

Instead of reflecting multimodal information to the memory cells, the memory fusion network (MFN), presented in [62], designs a multimodal memory component that keeps track of the multimodal interactions. In contrast to [61], it takes the two last memory components (instead of just the last) and passes them through a novel multimodal gated memory. It consists of three elements:

- *System of LSTMs*: One LSTM for each view.
- *Delta-Memory Attention Network (DMAN)*: For each modality, this takes the last two internal states (c of LSTM) and applies cross-view attention in order to combine the individual memories.
- *Multi-View Gated Memory*: This takes the output of DMAN and, based on the multimodal information, applies different neural networks to: (i) produce the new output, (ii) choose the amount of past information to keep and (iii) choose the amount of new information to keep.

MFN outputs the output of the multi-view gated memory together with the concatenation of each LSTM's hidden states.

An improvement of the memory component of this architecture is proposed in [63]. More specifically, a new hierarchical fusion method called dynamic fusion graph (DFG) is introduced in this work. This approach tries to model unimodal (t, v, a), bimodal (t, v, v, a, t, a) and trimodal (t, v, a) dynamics using a graph structure. It hierarchically combines all these multimodal subsets by using an efficacy scoring, represented by a sigmoid activated probability neuron for each connection, while it combines sets of these connections by using a neural network. The memory component of the memory fusion network (MFN) is replaced with DFG to form the graph memory fusion network (Graph-MFN). By visualizing the connections of DFG, it is shown that this network satisfactorily models priors on cross-modal interactions, since certain efficacies remain unchanged across cases and across time, while for the rest of the connections, it achieves extracting independent posteriors by changing its efficacies case by case and for each case over time.

Moving past the memory fusion idea, inspirations from neuroscience observations also seem promising. Based on works by [3–5], the authors of [64] argue that the brain's multimodal functionality is performed in a multi-stage manner, and, based on that, they introduce multi-stage fusion across modalities. For each modality, an LSTH(ybrid)M is used to extract intra-modal temporal representations, while for each timestamp, a multistage (K stages) fusion process (MFP) is followed: (i) the unimodal representations are concatenated and passed through a recurrent highlight network (LSTM), which outputs attention probabilities in order to select a features subset; (ii) this subset is passed through a recurrent fuse network (LSTM) in order to create a new joint representation of the selected features; and (iii) after K stages, the summarize network takes the last fuse output and creates a cross-modal representation which is passed in the LSTHM's memory.

However, the idea of tensor fusion remains promising in the manner of multimodal fusion. So far, conventional tensor fusion techniques are all restricted to the linear combination of multimodal features. More specifically, Refs. [6,52] are focused on modeling bilinear/trilinear cross-modal interactions, and thus their approach simply fuses multimodal features all at once, totally ignoring the local dynamics of interactions that are crucial to the final prediction. In [65], an architecture that employs polynomial cross-modal interactions is introduced. Its central building block is the polynomial tensor pooling (PTP), where a set of M feature vectors and the unit vector are first concatenated together into a long feature vector f. Then, a degree of P polynomial feature tensor F is formulated using a P-order tensor product of the concatenated feature vector f. F is capable of representing all possible polynomial expansions up to order P due to the incorporation of the unit vector. The effect of P polynomial interaction between features is transformed by a pooling weight tensor W which is applied using low-rank techniques.

By using windowing in the input vectors and applying PTP on the corresponding windows, they form the hierarchical polynomial fusion network (HPFN). Using a hierarchical architecture, the local temporal modality patterns of correlations can be recursively integrated via stacking PTPs in multiple layers. Here, we have to note that HPFN does not "force alignment" since the input modalities are time-averaged in sort blocks and thus avoid complete multimodal sequence interactions.

Learning joint representations has shown to result in high-performance models such as [58,59]. For this reason, there is a variety of works which try to apply representation learning in a more temporal approach. One good example is the adversarial representation graph fusion framework (ARGF), which was introduced in [66] and includes two basic units: the representation learning network and the hierarchical fusion network. The former firstly encodes each modality by using an encoder. The unimodal representations are used in two ways: (i) as an input to a decoder, which is trained with a reconstruction loss in order to retain the modality-specific dynamics, and (ii) as a grouped input to a discriminator, which is trained using adversarial loss in order to learn a joint representation, which are the input to the final classifier (trained with classification loss). As for the latter, it is a graph fusion network, which hierarchically combines multimodal representations.

Another representation learning technique that is completely temporal (TA) is proposed in [67] and aims to learn text-based representations since the authors defend that the principal modality for multimodal language analysis is that of text. The interaction canonical correlation network (ICCN) extracts temporal audio and video features by applying a Conv1D layer to the feature sequence which is later passed through an LSTM. The resulting feature vectors are used to individually perform the outer product with text embeddings. Both audio-text and video-text representations are later passed through a CNN to form two distinct representations. Using canonical correlation loss (CCL), these representations are concatenated with the initial text embeddings to form multimodal representations. In this way, the correlation between audio-based and video-based features is maximized.

Integrating multimodal information to memory components is a good overall approach for temporal multimodal learning. However, the goal of modeling cross-modal unaligned interactions across time is difficult to be served by the hidden states of an LSTM.

3.4.2. Temporal Alignment across Modalities

The approaches presented up to this point require forced phoneme alignment across modalities, while most of them demand same sequence size for all modalities. However, multimodal language dynamics are not temporally aligned by nature, since, for example, a gesture can happen before or after the corresponding phrase. For this reason, the architectures that learn cross-modal alignment can map the unimodal vectors in a more effective representation space.

One of the first works in this category is proposed for the speech emotion recognition problem in [68] and utilizes a basic multimodal sequence combination using an attention layer. More specifically, the architecture includes one LSTM network for each modality and applies cross-attention across the hidden states of both LSTMs. The resulting multimodal vector is concatenated with the text feature vector and passed through an LSTM which produces the final predictions. The application of the attention layer at the feature sequence produced by the LSTMs allows this architecture to be used in unaligned data sequences and learn an asynchronous alignment.

The recurrent attended variation embedding network (RAVEN) proposed in [69] models the nonverbal temporal interactions between subword units by adjusting (shifting) word representations based on the corresponding nonverbal behavior (i.e., acoustic and visual). RAVEN has three components: (i) nonverbal sub-networks, (ii) a gated modality-mixing network and (iii) multimodal shifting. For an input word in the utterance, the nonverbal sub-networks first compute the visual and acoustic embedding through modeling the sequence of visual and acoustic features lying in a word-long segment with a separate LSTM network. The gated modality-mixing network module then infers the nonverbal

shift vector as the weighted average over the visual and acoustic embedding based on the original word embedding. The multimodal shifting finally generates the adjusted multimodal word representation by integrating the nonverbal shift vector to the original word embedding. The multimodal-shifted word representation can then be used in the high-level hierarchy to predict target labels in the sentence. The produced embeddings of RAVEN are so powerful that they achieve close to state-of-the-art results by just using an LSTM and an FNN on the enriched word representations.

The architectures mentioned thus far mostly use the hidden states of LSTM networks in order to keep track of the temporal dynamics. However, it is questionable whether the presented architectures achieve keeping multimodal long-term dependencies, since the sequences of different modalities do not directly interact with each other. The use of transformer networks instead of LSTMs would be beneficial in this manner. The multimodal transformer (MulT) presented in [70] manages to efficiently learn and capture cross-modal relations with the use of transformers. For every target modality, there are $N-1$ transformers (corresponding to the other $N-1$ modalities), which align the feature sequences from their input modality to the targeted one. The output of these transformers is concatenated and passes from a self-attention transformer to produce the predictions. They use a Conv1D layer in the input sequences of embeddings in order to catch local temporal dependencies and to produce the same output size for all modalities.

The recent advances in the approaches of temporal aligned architectures (TAA) lead to the modal-temporal attention graph (MTAG) proposed in [71]. MTAG is a graph-based neural model which converts unaligned multimodal temporal data into a graph that captures the powerful interactions across modalities and time. Then, graph fusion (MTAG fusion) is performed along with a dynamic pruning and read-out technique, in order to process the modal-temporal graph and capture various interactions. MTAG learns to focus only on the important interactions within the graph and achieves state-of-the-art results in the unaligned task by just using 6% of the number of parameters used in [70] (MulT).

Write something about the unaligned task that [70,71] try to solve.

4. Evaluation

4.1. Datasets

Emotion recognition is a well-established subject in the greater area of machine learning, and as a result, there is a range of different datasets that address the growing need for data. The existing datasets can be classified into four categories according to the recording procedure that was followed during the data collection process. More specifically, these methods include one of the following: (i) *spontaneous speech:* the participants are unaware of the recording while their speech and reactions are recorded with hidden mechanisms in a real environment [72]; (ii) *acted speech:* the emotional condition of the speakers is acted; (iii) *elicited speech:* where the speaker is placed in a situation which evokes a specific emotional state [73]; and (iv) *annotated public speech*: data from public sources, such as YouTube, are annotated in order to associate them with a range of emotional states.

The models trained for the task of multimodal emotion recognition must have a great generalization ability that will help to both infer useful social information and be efficiently applied in industrial applications. Towards this end, the datasets for that very task must be realistic by having some desired properties, such as a variety in modalities, speakers, genders, subjects discussed, spoken languages, words used, emotional intensity and amount of data. Towards that end, the datasets presented in Table 1 try to address this need.

Table 1. Widely used multimodal emotion recognition datasets.

Dataset	Reference	Total Duration (h:m:s)	Speakers	Modalities	Language(s)
IEMOCAP	[74]	11:28:12	10	{A, T, V}	English
CMU-MOSEI	[63]	65:53:36	1000	{A, T, V}	English
RECOLA	[75]	03:50:00	46	{A, V}	French
VAM	[76]	12:00:00	20	{A, V}	German
SEWA	[77]	04:39:00	408	{A, V}	Chinese English German Greek Hungarian Serbian
HUMAINE	[78]	04:11:00	4	{A, V}	English
SEMAINE	[79]	06:30:00	20	{A, V}	English Greek Hebrew
AFEW	[80]	02:28:03	330	{A, V}	English
AFEW-VA	[81]	00:40:00	240	{A, V}	English
Mimicry	[82]	11:00:00	48	{A, V}	English

CMU-MOSEI [63] is the collection that achieves including most of the desired properties. It is the biggest dataset available that includes the most speakers and the greatest subject variety. It contains YouTube videos which makes it one of the most useful data sources, since many industrial products are used on similar data. CMU-MOSEI includes labels not only for the emotion recognition task but also for the (text-based) sentiment analysis problem. For this reason, it is mostly used for sentiment analysis architectures (e.g., [63,66,67,70,83–85]). However, its benefits must be adopted by researchers on emotion recognition in order to produce more robust models and gain a better insight into their actual performance.

One the contrary, IEMOCAP [74], the most widely used dataset, is collected in a detailed way in a lab environment. The use of ten actors which attach different realistic emotions in dyadic sessions makes this dataset the collection with the strongest ground truth. In addition, it contains visual features of high quality, since markers on the face, head and hands were used in order to capture facial expressions and hand movements. IEMOCAP is also one of the oldest, well recorded, annotated and maintained datasets, and thus the overwhelming majority of multimodal architectures are evaluated using this data. However, it has a variety of unwanted features since it only includes a small number of different speakers and subjects discussed.

The rest of the datasets are not extensively used in the literature of multimodal emotion recognition, and thus our analysis will focus completely on IEMOCAP, since this is the dataset which can be used to compare the effectiveness of different ideas for the problem of interest. However, the variety of different datasets is encouraging since they can be used in a cross-dataset evaluation scheme in order to produce highly generalizable results and upgrade the performance estimation to a more realistic manner.

4.2. Performance Evaluation Metrics

The performance of multimodal emotion recognition architectures is either measured on a class-based level or as an overall performance. For the former, binary accuracy and f1-score ($f_1 = 2 * (precision * recall)/(precision + recall)$) are used for each emotion class, while for the latter, the accuracy over all emotions is monitored together with the weighted accuracy (WA), where, according to [86],

$$WA = \frac{\mathrm{TP} * \mathrm{N}}{2\mathrm{N}(\mathrm{P} + \mathrm{TN})} \quad (1)$$

where N is the total number of negatives, P is the total number of positives, TP is the true positives and TN is the true negatives.

Here, we have to note that it is established to train and test the emotion recognition models on the four emotions of anger, happiness, sadness and neutral instead of the given nine (i.e., angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral), while beginning with [87], some works merge the happiness and excitement classes.

4.3. Evaluation Procedures

Every machine learning task that is associated with human speech must have speaker-independent training in order to avoid over-fitted testing results. However, this is not the case for all works in the field since there is a variety of works that use a simple cross-validation scheme (e.g., [53,54,56]). On the other hand, a variety of works use a speaker independent training procedure but measure performance on a single testing set. This type of evaluation is seen by either forming a test set with a percentage of speakers (e.g., train 80% speakers—test 20% speakers, as in [51]) or by separating the IEMOCAP sessions since each sessions includes two unique speakers (e.g., three sessions train—one session validation—one session testing, as in [6,52,59,61,62,69,70]). Here, we have to note that most of the one-session testing schemes have used the data and code given by CMU-MultimodalSDK (https://github.com/A2Zadeh/CMU-MultimodalSDK, accessed on 27 August 2021) for reasons of continuation.

The rest of the works use some statistically more stable approaches. More specifically, in such works, either leave-one-speaker-out (LO-Speaker-O) or leave-one-session-out (LO-Session-O) evaluation is performed, where either a speaker or a session is repeatedly left out during training in order to render the test set in a cross-validation-like manner. These methods are ideal for datasets of small or medium size such as IEMOCAP, since by using them, it is ensured that the resulting performance does not depend on a possible favorable choosing of speakers for the testing set.

4.4. Aggregated Reported Results

In Table 2, we only group some metrics associated with the IEMOCAP dataset, since most of the authors have tested their work on that collection. Some works present results for other datasets such as CMU-MOSEI (e.g., [63,66,84]) or REVOLA (e.g., [57]), but these works can be only grouped in small sets, and thus we think that the best overview of the field architectures can only be seen in an IEMOCAP comparison.

The results shown in Table 2 cannot be directly compared. This is because they do not all have the same validation procedure, and for the ones that do, it is not certain whether they used the same speakers or sessions for testing. For this reason, we juxtapose the authors' results in order to acquire a sense of the actual differences between each architecture's ideas. For an empirical comparison of such architectures, which trains from scratch a wide range of the presented architectures, we refer to [88].

We also report the results on the task of unaligned multimodal emotion recognition, which differs from the previous one, since the input data sequences are not forced-aligned by the dataset owners. This task is closer to reality and can result in approaches that extract more efficient multimodal temporal representations and are more robust in real-world applications.

In Table 3, the connectionist temporal classification (CTC) [89] method is adopted in order to use architectures that require forced-aligned data in an unaligned manner. Specifically, these models train to optimize the CTC alignment objective and the emotion multimodal objective simultaneously [70]. The results can be easily compared since all methods perform the exact same evaluation on the same sets of data. Thus, it can easily be inferred that the MTAG [71] is the architecture that holds the state-of-the-art performance in that task, while using just 0.14 million parameters.

Table 2. Multimodal emotion recognition evaluation results with rounding to one decimal place. The works in which a specific multimodal architecture was validated are indicated under the "Validated by" column, while the type of validation is marked in the "Validation" column.

Model	Reference	Validated by	Category	Modalities	Validation	Overall		Happy		Sad		Angry		Neutral	
						WA	UA	A	F1	A	F1	A	F1	A	F1
MV-LSTM	[60]	[69]	FTA	{A, T, V}	-	-	-	85.9	81.3	80.4	74.0	85.1	84.3	67.0	66.7
DF	[50]	[69]	NTA	{A, T, V}	-	-	-	86.0	81.0	81.8	81.2	75.8	65.4	59.1	44.0
BC-LSTM	[51]	[51,69]	NTA	{A, T, V}	Train-Test 80-20 Speakers	-	75.6	84.9	81.7	83.2	81.7	83.5	84.2	67.5	64.1
MARN	[61]	[69]	FTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	86.7	83.6	82.0	81.2	84.6	84.2	66.8	65.9
MFN	[62]	[69]	FTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2
TFN	[6]	[6]	NTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	-	83.6	-	82.8	-	84.2	-	65.4
RMFN	[64]	[69]	FTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	87.5	85.8	82.9	85.1	84.6	84.2	69.5	69.1
LMF	[52]	[52]	NTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	87.3	85.8	86.2	85.9	89.0	89.0	72.4	71.7
MDRE	[53]	[53]	UTA	{A, T,}	5-Fold Cross-Val	71.8	-	-	-	-	-	-	-	-	-
MHA-2	[54]	[54]	UTA	{A, T,}	10-Fold Cross-Val	75.6	76.5	-	-	-	-	-	-	-	-
CRMKL	[9]	[9]	UTA	{A, T, V}	10-Fold Cross-Val	-	-	72.2	-	75.6	-	79.2	-	80.4	-
MulT	[70]	[70]	ATA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	90.7	88.6	86.7	86.0	87.4	87.0	72.4	70.7
MFM	[59]	[59]	UTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	90.2	85.8	88.4	86.4	87.5	86.7	72.1	68.1
LAMER	[68]	[68]	ATA	{A, T,}	LO-Session-O	72.5	70.9	-	-	-	-	-	-	-	-
STSER	[55]	[55]	UTA	{A, T,}	LO-Session-O	71.1	72.0	-	-	-	-	-	-	-	-
ADF-III	[56]	[56]	UTA	{A, T,}	10-Fold Cross-Val	79.2	80.5	-	-	-	-	-	-	-	-
RAVEN	[69]	[69]	ATA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
MCTN	[58]	[70]	UTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	84.9	83.1	80.5	79.6	79.7	80.4	62.3	57.0
HPFN	[65]	[65]	FTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	-	86.2	-	86.6	-	88.8	-	72.5
ICCN	[67]	[67]	UTA	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	-	87.4	84.7	88.6	88.0	86.3	85.9	69.7	68.5

Table 3. Multimodal emotion recognition evaluation results for unaligned sequential input data. The works in which a specific multimodal architecture was validated are indicated under the "Validated by" column, while the type of validation is marked in the "Validation" column. For architectures that were not created to work in an unaligned manner, the CTC method was used.

Model	Reference	Validated by	Modalities	Validation	Happy		Sad		Angry		Neutral	
					A	**F1**	**A**	**F1**	**A**	**F1**	**A**	**F1**
CTC + RAVEN	[69]	[70]	{A, T, V}	Train-Val-Test 3-1-1 Sessions	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
CTC + MCTN	[58]	[70]	{A, T, V}	Train-Val-Test 3-1-1 Sessions	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT	[70]	[70]	{A, T, V}	Train-Val-Test 3-1-1 Sessions	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
MTAG	[71]	[71]	{A, T, V}	Train-Val-Test 3-1-1 Sessions	-	**86.0**	-	**79.9**	-	**76.7**	-	**64.1**

5. Conclusions and Future Challenges

In this work we presented a review on multimodal temporal deep learning architectures for speech emotion recognition. The key role of the temporal dimension has been emphasized, and a temporal-learning-based categorization has been defined. Based on this rationale, we categorized a wide range of methodologies in the literature into one of three clusters: (i) Non-Temporal Architectures (NTA), (ii) Pseudo-Temporal Architectures (PTA) and (iii) Temporal Architectures (TA). The feature extraction methods used in the literature have been extensively discussed as well, while aggregated evaluation results have been presented. However, based on the aforementioned analysis, there are certain future challenges that arise and can be narrowed down to three key factors: (i) evaluation, (ii) robustness and (iii) representation learning.

First, despite the fact that emotion recognition is already 20 years old, one of the major challenges that is still active is associated with the existence of a common evaluation procedure. More specifically, the proposed architectures are evaluated using not only different types of evaluation pipelines (e.g., random subsampling train-val split vs leave-one-speaker out) but also different classification metrics (e.g., binary metrics for each label vs multiclass aggregated metrics). From the evaluation procedures presented in Section 4, it can be inferred that a speaker-dependent experimentation may result in overfitted results (in terms of speaker identity). We strongly believe that leave-one-speaker-out evaluation should be used as a standard not only to avoid overfitting to speaker identities but also to have a common evaluation procedure across different datasets. In addition, in order to construct a general evaluation framework, the experiments have to produce generic results and list both binary and multiclass metrics and also extend the produced results in order to capture both the dimensional (i.e., valence, arousal and dominance) and the categorical (discrete emotions) dimensions of emotion recognition.

One more major problem with the proposed architectures is associated with their real-world adaptation. These models are difficult to be adapted in an industrial application, since they are usually engineered on specific datasets (thus lacking generalization power), relying on unrealistic data such as forced-aligned multimodal sequences, error-less text transcriptions, non realistic recording conditions, artificial dialog context and perfect information on all modalities. Thus, in practice, more robust models need to be designed, and thus future works should produce multimodal temporal architectures for the problem of speech emotion recognition that are: (i) trained in a cross-dataset evaluation manner, while utilizing the power of unsupervised [90] or supervised [91] domain adaptation methodologies, in order to better evaluate their generalization power; (ii) capable of performing inference on unaligned temporal multimodal data; (iii) capable of performing inference in cases of noisy or absent modalities; and (iv) capable of performing under unexpected ASR errors.

Apart from the aforementioned challenges that will result in more robust and better statistically evaluated systems, more efficient methods need to be adopted for signal representation. More specifically, in the task of multimodal speech emotion recognition, a few specific low-level modality representation techniques have been established. In particular, for the aural modality, signal-analysis-based features that are associated with both the spectral and temporal domains are used. As for the textual and visual modalities, basic word embeddings are used for the former, while traditional facial features are used for the latter. However, the development of deep learning techniques has resulted in better representation capabilities for all three of these modalities. Thus, multimodal architectures need to be in line with the state-of-the-art techniques for unimodal representations and find efficient ways to combine them by taking into account the temporal dimension . Finally, each modality can be enriched with new representations that also take into account parts of the information that have not been so widely used until now. For example, despite the fact that body language is strongly correlated with expressed emotions, the only information that is captured from the visual modality is that of facial expressions.

Last but not least, one more concept that is absent in the literature of multimodal emotion recognition is unsupervised representation. In other machine learning application domains, there are many powerful unsupervised representation methods, such as BERT [41] for text, wav2vec 2.0 [31] for audio and MoCo [92] for vision, that are adopted in an application-independent manner. Several well-defined datasets have been created for the task of emotion recognition, and research that combines them in order to generate generic multimodal emotion-oriented representations may result in a exponential growth of the use of emotional analysis in other application domains. As an example, an "emotional" multimodal embedding methodology could be used in a speaker verification system to improve the robustness of the system to the emotional state of the speaker.

Author Contributions: Conceptualization, P.K. and T.G.; methodology, P.K.; validation, T.G.; investigation, P.K. and T.G.; visualization, P.K.; supervision, T.G.; writing—original draft preparation, P.K. and T.G.; writing—review and editing, P.K. and T.G. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
2. Müller, F.M. *Lectures on the Science of Language: Delivered at the Royal Institution of Great Britain in February, March, April and May, 1863*; Charles Scribner: New York, NY, USA, 1865; Volume 2.
3. Kuzmanovic, B.; Bente, G.; von Cramon, D.Y.; Schilbach, L.; Tittgemeyer, M.; Vogeley, K. Imaging first impressions: Distinct neural processing of verbal and nonverbal social information. *Neuroimage* **2012**, *60*, 179–188. [CrossRef] [PubMed]
4. Sergent, J.; Signoret, J.L. Functional and anatomical decomposition of face processing: Evidence from prosopagnosia and PET study of normal subjects. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **1992**, *335*, 55–62.
5. Jiang, J.; Dai, B.; Peng, D.; Zhu, C.; Liu, L.; Lu, C. Neural synchronization during face-to-face communication. *J. Neurosci.* **2012**, *32*, 16064–16069. [CrossRef]
6. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 1103–1114.
7. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal Deep Learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696
8. Pérez-Rosas, V.; Mihalcea, R.; Morency, L.P. Utterance-level multimodal sentiment analysis. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Volume 1, pp. 973–982.
9. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.

10. Wang, H.; Meghawat, A.; Morency, L.P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 949–954.
11. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
12. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [CrossRef]
13. Lalitha, S.; Mudupu, A.; Nandyala, B.V.; Munagala, R. Speech emotion recognition using DWT. In Proceedings of the 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, 10–12 December 2015; pp. 1–4.
14. Kuchibhotla, S.; Vankayalapati, H.D.; Vaddi, R.; Anne, K.R. A comparative analysis of classifiers in emotion recognition through acoustic features. *Int. J. Speech Technol.* **2014**, *17*, 401–408. [CrossRef]
15. Gupta, D.; Bansal, P.; Choudhary, K. The state of the art of feature extraction techniques in speech recognition. In *Speech and Language Processing for Human-Machine Communications: Proceedings of CSI 2015*; Springer: Singapore, 2018; pp. 195–207.
16. Gupta, K.; Gupta, D. An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. In Proceedings of the 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, 14–15 January 2016; pp. 493–497.
17. Chatterjee, M.; Zion, D.J.; Deroche, M.L.; Burianek, B.A.; Limb, C.J.; Goren, A.P.; Kulkarni, A.M.; Christensen, J.A. Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hear. Res.* **2015**, *322*, 151–162. [CrossRef]
18. Guidi, A.; Gentili, C.; Scilingo, E.P.; Vanello, N. Analysis of speech features and personality traits. *Biomed. Signal Process. Control* **2019**, *51*, 1–7. [CrossRef]
19. Teager, H.; Teager, S. Evidence for nonlinear sound production mechanisms in the vocal tract. In *Speech Production and Speech Modelling*; Springer: Dordrecht, The Netherlands, 1990; pp. 241–261.
20. Kaiser, J.F. Some useful properties of Teager's energy operators. In Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 27–30 April 1993; Volume 3, pp. 149–152.
21. Zhou, G.; Hansen, J.H.; Kaiser, J.F. Nonlinear feature based classification of speech under stress. *IEEE Trans. Speech Audio Process.* **2001**, *9*, 201–216. [CrossRef]
22. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
23. Giannakopoulos, T. pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS ONE* **2015**, *10*, e0144610. [CrossRef]
24. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (Icassp), Florence, Italy, 4–9 May 2014; pp. 960–964.
25. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 1459–1462.
26. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
27. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
28. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [CrossRef]
29. Ravanelli, M.; Bengio, Y. Speaker recognition from raw waveform with sincnet. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 1021–1028.
30. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1904.05862.
31. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**, arXiv:2006.11477.
32. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**, arXiv:2106.07447.
33. Mikolov, T.; Yih, W.T.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
34. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
35. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
36. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.

37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
38. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 27 August 2021).
39. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
40. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
43. Ekman, P. Facial expression and emotion. *Am. Psychol.* **1993**, *48*, 384. [CrossRef]
44. Marechal, C.; Mikolajewski, D.; Tyburek, K.; Prokopowicz, P.; Bougueroua, L.; Ancourt, C.; Wegrzyn-Wolska, K. Survey on AI-Based Multimodal Methods for Emotion. In *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*; Springer International Publishing: Cham, Switzerland, 2019; pp. 307–324.
45. Jack, R.E.; Schyns, P.G. The human face as a dynamic tool for social communication. *Curr. Biol.* **2015**, *25*, R621–R634. [CrossRef]
46. Ekman, P.; Friesen, W.; Hager, J. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press, Inc.: Palo Alto, CA, USA, 1978.
47. Hjortsjö, C. *Man's Face and Mimic Language*; Studentlitteratur: Lund, Sweden, 1969.
48. De la Torre, F.; Chu, W.S.; Xiong, X.; Vicente, F.; Ding, X.; Cohn, J. IntraFace. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015.
49. iMotions Global. Facial Expression Analysis Solutions. Available online: https://imotions.com/biosensor/fea-facial-expression-analysis/ (accessed on 27 August 2021).
50. Nojavanasghari, B.; Gopinath, D.; Koushik, J.; Baltrušaitis, T.; Morency, L.P. Deep multimodal fusion for persuasiveness prediction. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 284–288.
51. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in user-generated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 873–883.
52. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.P. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2247–2256.
53. Yoon, S.; Byun, S.; Jung, K. Multimodal speech emotion recognition using audio and text. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 112–118.
54. Yoon, S.; Byun, S.; Dey, S.; Jung, K. Speech emotion recognition using multi-hop attention mechanism. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2822–2826.
55. Chen, M.; Zhao, X. A multi-scale fusion framework for bimodal speech emotion recognition. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 374–378.
56. Priyasad, D.; Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. Attention driven fusion for multi-modal emotion recognition. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3227–3231.
57. Liang, P.P.; Zadeh, A.; Morency, L.P. Multimodal local-global ranking fusion for emotion recognition. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 472–476.
58. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6892–6899.
59. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning Factorized Multimodal Representations. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada , 30 April–3 May 2018.
60. Rajagopalan, S.S.; Morency, L.P.; Baltrusaitis, T.; Goecke, R. Extending long short-term memory for multi-view structured learning. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 338–353.
61. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
62. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.

63. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2236–2246.
64. Liang, P.P.; Liu, Z.; Zadeh, A.B.; Morency, L.P. Multimodal Language Analysis with Recurrent Multistage Fusion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 150–161.
65. Hou, M.; Tang, J.; Zhang, J.; Kong, W.; Zhao, Q. Deep multimodal multilinear fusion with high-order polynomial pooling. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 12136–12145.
66. Mai, S.; Hu, H.; Xing, S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 164–172.
67. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8992–8999.
68. Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; Li, X. Learning Alignment for Multimodal Emotion Recognition from Speech. *arXiv* **2019**, arXiv:1909.05645.
69. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7216–7223.
70. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; p. 6558.
71. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; pp. 1009–1021.
72. Cao, H.; Verma, R.; Nenkova, A. Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Comput. Speech Lang.* **2015**, *29*, 186–202. [CrossRef] [PubMed]
73. Basu, S.; Chakraborty, J.; Bag, A.; Aftabuddin, M. A review on emotion recognition using speech. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 10–11 March 2017; pp. 109–114.
74. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
75. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8.
76. Grimm, M.; Kroschel, K.; Narayanan, S. The Vera am Mittag German audio-visual emotional speech database. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23 June–26 April 2008; pp. 865–868.
77. Kossaifi, J.; Walecki, R.; Panagakis, Y.; Shen, J.; Schmitt, M.; Ringeval, F.; Han, J.; Pandit, V.; Toisoul, A.; Schuller, B.W.; et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1022–1040 [CrossRef] [PubMed]
78. Douglas-Cowie, E.; Cowie, R.; Sneddon, I.; Cox, C.; Lowry, O.; McRorie, M.; Martin, J.C.; Devillers, L.; Abrilian, S.; Batliner, A.; et al. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 12–14 September 2007; pp. 488–500.
79. 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]
80. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Ann. Hist. Comput.* **2012**, *19*, 34–41. [CrossRef]
81. Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* **2017**, *65*, 23–36. [CrossRef]
82. Bilakhia, S.; Petridis, S.; Nijholt, A.; Pantic, M. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern Recognit. Lett.* **2015**, *66*, 52–61. [CrossRef]
83. Siriwardhana, S.; Reis, A.; Weerasekera, R.; Nanayakkara, S. Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *arXiv* **2020**, arXiv:2008.06682.
84. Shenoy, A.; Sardana, A. Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In Proceedings of the Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML), Online, 10 July 2020; pp. 19–28.
85. Mai, S.; Xing, S.; Hu, H. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1424–1437. [CrossRef]

86. Tong, E.; Zadeh, A.; Jones, C.; Morency, L.P. Combating human trafficking with multimodal deep models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1547–1556.
87. Metallinou, A.; Lee, S.; Narayanan, S. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 2462–2465.
88. Gkoumas, D.; Li, Q.; Lioma, C.; Yu, Y.; Song, D. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. *Inf. Fusion* **2021**, *66*, 184–197. [CrossRef]
89. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
90. Deng, J.; Zhang, Z.; Eyben, F.; Schuller, B. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process. Lett.* **2014**, *21*, 1068–1072. [CrossRef]
91. Abdelwahab, M.; Busso, C. Supervised domain adaptation for emotion recognition from speech. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5058–5062.
92. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.