frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 09 February 2023
DOI 10.3389/fcomp.2023.1072479

Check for updates

OPEN ACCESS

EDITED BY
Theodora Chaspari,
Texas A&M University, United States

REVIEWED BY
Xiaoming Zhao,
Taizhou University, China
Alessio Brutti,
Bruno Kessler Foundation (FBK), Italy

*CORRESPONDENCE
Andreas Triantafyllopoulos
✉ andreas.triantafyllopoulos@uni-a.de

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 17 October 2022
ACCEPTED 24 January 2023
PUBLISHED 09 February 2023

CITATION
Triantafyllopoulos A, Reichel U, Liu S, Huber S, Eyben F and Schuller BW (2023) Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Front. Comput. Sci.* 5:1072479. doi: 10.3389/fcomp.2023.1072479

COPYRIGHT
© 2023 Triantafyllopoulos, Reichel, Liu, Huber, Eyben and Schuller. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Multistage linguistic conditioning of convolutional layers for speech emotion recognition

Andreas Triantafyllopoulos[1,2]*, Uwe Reichel[1], Shuo Liu[2], Stephan Huber[1], Florian Eyben[1] and Björn W. Schuller[1,2,3]

[1]audEERING GmbH, Gilching, Germany, [2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, [3]Group on Language, Audio, & Music, Imperial College London, London, United Kingdom

Introduction: The effective fusion of text and audio information for categorical and dimensional speech emotion recognition (SER) remains an open issue, especially given the vast potential of deep neural networks (DNNs) to provide a tighter integration of the two.

Methods: In this contribution, we investigate the effectiveness of deep fusion of text and audio features for categorical and dimensional SER. We propose a novel, multistage fusion method where the two information streams are integrated in several layers of a DNN, and contrast it with a single-stage one where the streams are merged in a single point. Both methods depend on extracting summary linguistic embeddings from a pre-trained BERT model, and conditioning one or more intermediate representations of a convolutional model operating on log-Mel spectrograms.

Results: Experiments on the MSP-Podcast and IEMOCAP datasets demonstrate that the two fusion methods clearly outperform a shallow (late) fusion baseline and their unimodal constituents, both in terms of quantitative performance and qualitative behavior.

Discussion: Overall, our multistage fusion shows better quantitative performance, surpassing alternatives on most of our evaluations. This illustrates the potential of multistage fusion in better assimilating text and audio information.

KEYWORDS

speech emotion recognition, multimodal fusion, speech processing, natural language processing, machine learning

1. Introduction

Automatic emotion recognition (AER) is an important component of human-computer interfaces, with applications in health and wellbeing, multimedia information retrieval, and dialogue systems. Human emotions are expressed in, and can accordingly be identified from, different modalities, such as speech, gestures, and facial expressions (Zeng et al., 2008; Calvo and D'Mello, 2010). A plethora of previous works have thus investigated different ways to improve AER by combining several information streams: e.g., audio, video, text, gestures, and physiological signals. Underlying these computational approaches are distinct emotion theories, with the two most commonly-used ones being *discrete (or basic) emotion theories* (Ekman, 1992) and *dimensional* ones (Russell and Mehrabian, 1977). The former specifies a list of discrete categories, e.g., "happy" or "sad," with one or more of them being used to characterize the emotional state of an individual at any given moment (Ekman, 1992). The latter defines a list of dimensions over which the emotional state varies; usually, these dimensions are arousal (emotional intensity), valence (the pleasantness of a stimulus), and dominance (degree of control) (Russell and Mehrabian, 1977).

Over the years, different modalities have proven more conducive to the recognition of different emotional states. For example, video and text have shown better performance at valence recognition, whereas acoustic cues are better indicators of arousal (Calvo and D'Mello, 2010; Schuller, 2018). This has led several prior works to pursue a fusion of one or several of those to improve performance. The dominant computational paradigm for fusing different modalities has been that of *shallow fusion* (Zeng et al., 2008; Atrey et al., 2010), where two or more modalities are first processed independently before being merged at a single point in the processing chain. This paradigm can be further broken down to *late, decision-level fusion* and *early, feature-level fusion* approaches (Zeng et al., 2008; Atrey et al., 2010). Late fusion corresponds to training separate models for each modality and subsequently combining their independent decisions, whereas early fusion approaches typically train a single model using input features derived from different modalities.

Recently, the success of deep learning (DL) has given rise to *deep fusion* approaches (Tzirakis et al., 2017), which utilize the power of deep neural networks (DNNs) to learn better multimodal representations. Deep fusion is conceptually similar to shallow fusion, with both jointly processing features from two or more modalities. However, the term "deep" is used to distinguish from simple, monolithic architectures which simply accept input features from two modalities (Atrey et al., 2010) as these fall under the category of shallow fusion. Instead, deep fusion architectures consists of several differential modules, some of them unimodal and others multimodal, which are jointly trained (Tzirakis et al., 2017). Such architectures have been successfully utilized for several tasks, such as multimodal AER (Tzirakis et al., 2017; Siriwardhana et al., 2020b).

In the present contribution, we are primarily interested in speech emotion recognition (SER). Speech, as the primary mode of human communication, is also a major conduit of emotional expression (Calvo and D'Mello, 2010). Accordingly, SER has been a prominent research area in affective computing for several years (Schuller, 2018). Although it constitutes a single modality, it is often analyzed as two separate information streams, a *linguistic* (*what* has been said) and a *paralinguistic* one (*how* it has been said) (Calvo and D'Mello, 2010; Schuller, 2018; Atmaja et al., 2022). However, the two streams are not independent. Previous studies have established that the interaction between acoustic descriptors and emotional states depends on the linguistic content of an utterance (Scherer et al., 1984). Moreover, text information is better suited for valence and audio for arousal recognition (Calvo and D'Mello, 2010). As a result, several works have attempted to more tightly integrate the two streams in an attempt to model their complex interrelationship and obtain more reliable recognition performance (Zeng et al., 2008; Calvo and D'Mello, 2010; Atmaja et al., 2022).

Recently, deep fusion architectures have proven very successful at utilizing linguistic and acoustic cues for SER (Lee et al., 2018; Chen et al., 2019; Georgiou et al., 2019; Siriwardhana et al., 2020b). However, most existing such methods are primarily based on sequential models [e.g., long short-term memorys (LSTMs)] operating on expert-based acoustic descriptors (Lee et al., 2018; Chen et al., 2019). Whereas, such methods have a long history in the field of SER, in recent years convolutional neural networks (CNNs) operating on raw audio or low-level features have been shown capable of learning good representations that lead to better performance (Trigeorgis et al., 2016; Fayek et al., 2017; Neumann and Vu, 2017). Thus, the combination of multistage fusion with the representation power of CNNs for auditory processing is a natural next step in the attempt to closely integrate the acoustic and linguistic information streams.

Inspired by such works, we propose a novel, CNN-based multistage fusion architecture where summary linguistic embeddings extracted from a pre-trained language model are used to condition multiple intermediate layers of a CNN operating on log-Mel spectrograms. We contrast it with a single-stage DNN architecture, where the two information streams are fused at a single point, and a late fusion method, where unimodal models are trained independently and their decisions are aggregated. Furthermore, we complement our experimental results with an in-depth analysis on two previously underresearched factors that influence model behavior: the role of the underlying dialogue act, and the difference between scripted vs. improvised text in an acted emotional dataset. This analysis helps to elucidate why bimodal fusion is beneficial for SER, and, in particular, why it benefits some dimensions and classes more than others.

The rest of this contribution is organized as follows. We begin by presenting an overview of related works in Section 2, with an emphasis on single-stage and multistage fusion approaches. We then describe our architecture and experimental settings in Section 3. Results and analyzes are reported in Section 4. We end with a discussion in Section 5 and our conclusion in Section 6.

2. Related work

In this section, we present an overview of related works. We begin by a brief overview of the state-of-the-art in unimodal, either text- or acoustic-based, AER methods. We then present an overview of fusion methods. For a more in-depth, recent survey on bimodal SER, we refer the reader to Atmaja et al. (2022).

2.1. Audio-based emotion recognition

The goal of audio-based SER systems is to estimate the target speaker's emotion by analysing the non-verbal content of their voice (Schuller, 2018). This is traditionally handled by extracting a set of low-level descriptors (LLDs), such as Mel frequency ceptral coefficients (MFCCs) or pitch, which capture relevant paralinguistic cues (Schuller et al., 2013). From these, a set of higher-level descriptors (HLDs) can be derived, such as statistical functionals of the LLDs like mean or standard deviation, resulting in a vector of fixed dimensionality which contains aggregated information over the entire utterance. This fixed-length vector then forms the input of a machine learning model such as a support vector machine (SVM) (Schuller et al., 2013) or a fully connected neural network (FCNN) (Parthasarathy and Busso, 2017).

In recent years, DNN-based methods that exploit the temporal information inherent in LLDs have come to dominate the modeling paradigm (Fayek et al., 2017). These fall broadly under two categories: *sequential* and *convolutional* ones. Sequential models, such as LSTMs, jointly process all available frame-level LLDs. CNNs, on the other hand, operate on a subset of LLDs which lend themselves well to the inductive biases of convolution, typically (log-Mel) spectrograms or MFCCs

(Fayek et al., 2017; Neumann and Vu, 2017; Zhao et al., 2021). The recent advances in attention-based models (Vaswani et al., 2017) have also been utilized for SER, for example by augmenting both sequential and convolutional models with attention (Zhao et al., 2021), or using transformer-based architectures (Wagner et al., 2022). Finally, end-to-end methods that operate on raw audio input have also shown competitive performance (Trigeorgis et al., 2016).

2.2. Text-based emotion recognition

We begin by noting that there is a rich body of work on text-based *sentiment* detection methods. While sentiment serves an important role in affective computing, it is nevertheless different from emotion (Munezero et al., 2014). Providing a full definition of both terms is beyond the scope of this work and we refer the reader to Scherer (2000) and Munezero et al. (2014) as entry points on this topic. For our purposes, we settle on a working definition of sentiment as a specific disposition held toward an object, topic, or situation (Munezero et al., 2014) and of emotion as an (episodic) change in the affective state of an individual as a result of external or internal stimuli (Scherer, 2000). We have thus intentionally limited our overview to text-based methods addressing the emotion recognition problem, with an emphasis on methods operating on spoken, rather than written, corpora. The reader is referred to Zhang et al. (2018) for a recent survey on text-based sentiment detection.

Early works on text-based emotion recognition utilized affective lexica, such as the WordNet-Affect dictionary (Strapparava et al., 2004), to generate word-level scores, which could then be combined using expert rules to derive a sentence-level prediction (Perikos and Hatzilygeroudis, 2013). In recent years, the use of learnt textual representations, like word2vec (Mikolov et al., 2013), has substituted these methods. These representations are used as input features for machine learning (ML) models, such as LSTMs (Tseng et al., 2021), and the systems are trained on available emotional data.

Moreover, attention-based models (transformers) (Vaswani et al., 2017) have shown exceptional performance on several natural language processing (NLP) tasks. These models are usually pre-trained on large, unlabelled corpora using some proxy task, e.g., masked language prediction (Devlin et al., 2019), which enables them to learn generic text representations. They are then fine-tuned on available emotional datasets to learn the emotion recognition task (Siriwardhana et al., 2020b; Acheampong et al., 2021).

2.3. Shallow fusion

Early work in multimodal fusion has primarily followed the shallow fusion paradigm (Atrey et al., 2010; Poria et al., 2017). Several early systems depended on hand-crafted features, usually HLDs, extracted independently for each modality, which were subsequently processed by a fusion architecture adhering to the early or late fusion paradigm. Early fusion corresponds to feeding the HLDs from both information streams as input to a single classifier. For example, Schuller et al. (2005) used acoustic HLDs and bag-of-words (BoW) linguistic features as input to an SVM. Late fusion, on the other hand, is achieved by training unimodal classifiers independently, and subsequently aggregating their predictions. The aggregation can consist of simple rules (e.g., averaging the predictions) or be delegated to a cascade classifier (Steidl et al., 2009).

2.4. Deep fusion

With the advent of DL, traditional, shallow fusion methods have been substituted by end-to-end multimodal systems (Tzirakis et al., 2017) where the different modalities are processed by jointly-trained modules. We differentiate between *single-stage* and *multistage* fusion.

Single-stage fusion is the natural extension of shallow fusion methods, where the different modalities are first processed separately by independent differentiable modules. These modules produce intermediate unimodal representations which are then merged in a downstream fusion module. Finally, this fusion stage is followed by one or more output layers which process the now multimodal representations to generate a final prediction.

Several works follow this fusion paradigm. For example, Lee et al. (2018) first process text and audio independently using unimodal CNNs, before combining both with cross-modal attention and using another CNN to do the final classification. Chen et al. (2019) use textual embeddings to attend to acoustic embeddings. Priyasad et al. (2020) utilize attention to fuse linguistic and acoustic representations extracted by a bidirectional long short-term memory (bLSTM) network processing GloVe embeddings, and a SincNet network processing raw audio, respectively. Finally, Yang et al. (2020) leverage the power of attention-based architectures and perform multimodal attention on the learnt representations of a BERT model (Devlin et al., 2019).

Although these methods have consistently outperformed both the unimodal baselines and shallow fusion alternatives, they nevertheless build on independently learnt unimodal representations constructed by modules agnostic to the presence of other modalities. In an attempt to utilize the power of DL to learn useful representations after several layers of processing, the community has also pursued multistage fusion paradigms, where the processing of different modalities is intertwined in multiple layers of a DNN. Prior multistage fusion works are primarily based on sequential models. Tseng et al. (2021) combine text tokens and acoustic embeddings using a sigmoid gating function, with each gated token being the input of a multimodal language modeling bLSTM. Georgiou et al. (2019) extend the previous to incorporate a hierarchical fusion model and explicitly model word and sentence level interactions. Zadeh et al. (2018) propose a hierarchical, dynamic fusion graph for combining the intermediate representations of unimodal modules.

The above works all investigate a single fusion paradigm: either single- or multistage. It is thus not clear how those two mechanisms fare in comparison to one another. Recently, Siriwardhana et al. (2020b) explicitly contrast a single-stage fusion of audio and text representations coming from vq-wav2vec (Baevski et al., 2019) and RoBERTa (Liu et al., 2019), respectively, with a multistage, attention-based fusion of these representations, and find that the former, simpler mechanism gives better performance.

Our proposed multistage fusion mechanism draws inspiration from several recent approaches in style transfer (Karras et al., 2019), speech synthesis (van den Oord et al., 2018), denoising

(Keren et al., 2018; Gfeller et al., 2020), and speaker adaptation (Triantafyllopoulos et al., 2021). In general, all those approaches utilize two DNNs: one devoted to the primary task, and a second providing additional information through some fusion mechanism. The two networks can be trained independently (van den Oord et al., 2018) or jointly (Keren et al., 2018; Triantafyllopoulos et al., 2021). The fusion mechanisms utilized in these works all operate on the same principle: the embeddings produced by the secondary network modulate the output of several (usually all) convolution layers of the primary network either by shifting or a combination of scaling and shifting. We use shifting as it is simpler and more stable. This fusion mechanism has the advantage of injecting the additional information in multiple layers of a model; these layers can thus specialize on learning the necessary features conditioned on this additional information, rather than having to learn generic ones.

3. Methods

In this section, we present our proposed multi-stage fusion architecture and baselines in Section 3.1, as well as our datasets in Section 3.2 and experimental setup in Section 3.3.

3.1. Architectures

The focus of the current contribution is on tightly integrating acoustic and linguistic information. This is achieved by proposing a multistage fusion approach where linguistic embeddings condition several intermediate layers of a CNN processing audio information. The overall architecture comprises of two constituent networks: a (pre-trained) text-based model that provides linguistic embeddings and an auditory CNN whose intermediate representations are conditioned on those embeddings.

The two unimodal architectures can obviously be trained independently for emotion recognition and their predictions can be aggregated in a straightforward way (e.g., by averaging) to produce a combined output. This simple setup is used in some of our experiments as a shallow (late) fusion baseline. Additionally, the two information streams can be combined in single-stage fashion, with the linguistic embeddings fused at a single point with the embeddings generated by the CNN; a setup which forms a deep fusion baseline with which to compare our method.

As our unimodal text model, we use BERT (Devlin et al., 2019), a pre-trained model with a strong track-record on several NLP tasks. The BERT model has an identical architecture with the transformer encoder originally proposed by Vaswani et al. (2017). BERT, embeddings of which are employed in our fusion approaches, consists of twelve transformer blocks. Each of them contains one self-attention layer with twelves heads and a hidden size (H) of 768, followed by a 2-layer feed-forward network with hidden size of $4 \cdot H$. Each of these two sub-networks is in turn followed by layer normalization and has a residual connection around them. The input embeddings to the BERT model are defined as the sum of token embeddings, sentence embeddings, and position embeddings. BERT's deep contextualized word representations are pre-trained with self-supervised learning on two pre-training tasks and a large amount of unlabelled text data. In the first pre-training task, masked language modeling, masked input tokens are predicted via the corresponding final output vector which rests upon contextual information from the other input tokens. The other pre-training task—next sentence prediction—considers the binarised problem of whether one sentences comes after another.

As our baseline acoustic model, we use the CNN14 architecture introduced by Kong et al. (2020), which was found to give good performance for emotional dimensions (Triantafyllopoulos and Schuller, 2021). While pre-trained speech transformer models have shown better performance than CNNs in recent years, they are unsuitable for our method, as they are only better when pre-trained on an upstream (e.g., self-supervised) task, and injecting linguistic information in their earlier layers would break this pre-training (Wagner et al., 2022). In contrast, CNN models can be trained from scratch more effectively on (generally limited) emotional datasets.

CNN14 consists of 12 convolution layers and 2 linear ones, with mean and max pooling following the last convolution layer to aggregate its information over the time and frequency axes. CNN14 follows the VGG architecture design. It consists of 6 convolutional blocks of two convolutional layers each, with each block followed by max pooling. All convolutional layers are using a 3×3 kernel and a stride of 1×1, whereas max pooling layers use a stride of 2×2. After the last convolution layer, features are pooled using both mean and max pooling, and subsequently fed into two linear layers. Dropout with a probability of 0.2 is applied after every each convolution block. As features, we use log-Mel spectrograms computed with 64 Mel bins, a window size of 32 ms, and a hop size of 10 ms, similar to the original authors. Our sampling rate for all experiments is 16 kHz. The network can process an audio feature sequence of arbitrary length due to its mean and max pooling mechanism, which is what we do during test time. During training, we randomly crop all sequences to the same length so we can process them in batches, as is the common practice in audio processing (Neumann and Vu, 2017). For our experiments, we choose 5 s as our constant utterance length during training time. In principle, our approach can also be combined with recent advances in handling variable length sequences during training (Lin and Busso, 2021).

These unimodal networks form the building blocks of our fusion methods, which are illustrated in Figure 1. Both architectures first pass the text input through a pre-trained BERT model, and then use it to condition one or more layers of the CNN14 base architecture. For the purposes of this work, we chose not to fine-tune the layers of the pre-trained BERT when using it in the fusion architectures. This was done because, as also shown by our experiments, BERT is a very powerful model capable of achieving very high performance on its own (when fine-tuned on the target task). However, our emphasis is primarily on investigating the behavior of deep fusion architectures. Fine-tuning BERT alongside CNN14 might confound our findings on how these fusion architectures behave. Nevertheless, we expect that jointly fine-tuning both models would result in even better performance, as shown by recent works (Siriwardhana et al., 2020b).

Our proposed multistage fusion method, shown in Figure 1A, relies on fusing linguistic representations in the form of embeddings extracted from a pre-trained BERT model with the intermediate layer outputs of an acoustics-based CNN model. Linguistic embeddings are computed by averaging the token-embeddings returned by BERT for each utterance. Similar to prior works (Keren et al., 2018; Triantafyllopoulos et al., 2021), we use the averaged embeddings to shift the intermediate representations of each block. Given an input $\boldsymbol{X} \in \mathbb{R}^{T_{\text{in}} \times F_{\text{in}}}$ to each convolution block, with T and F being the

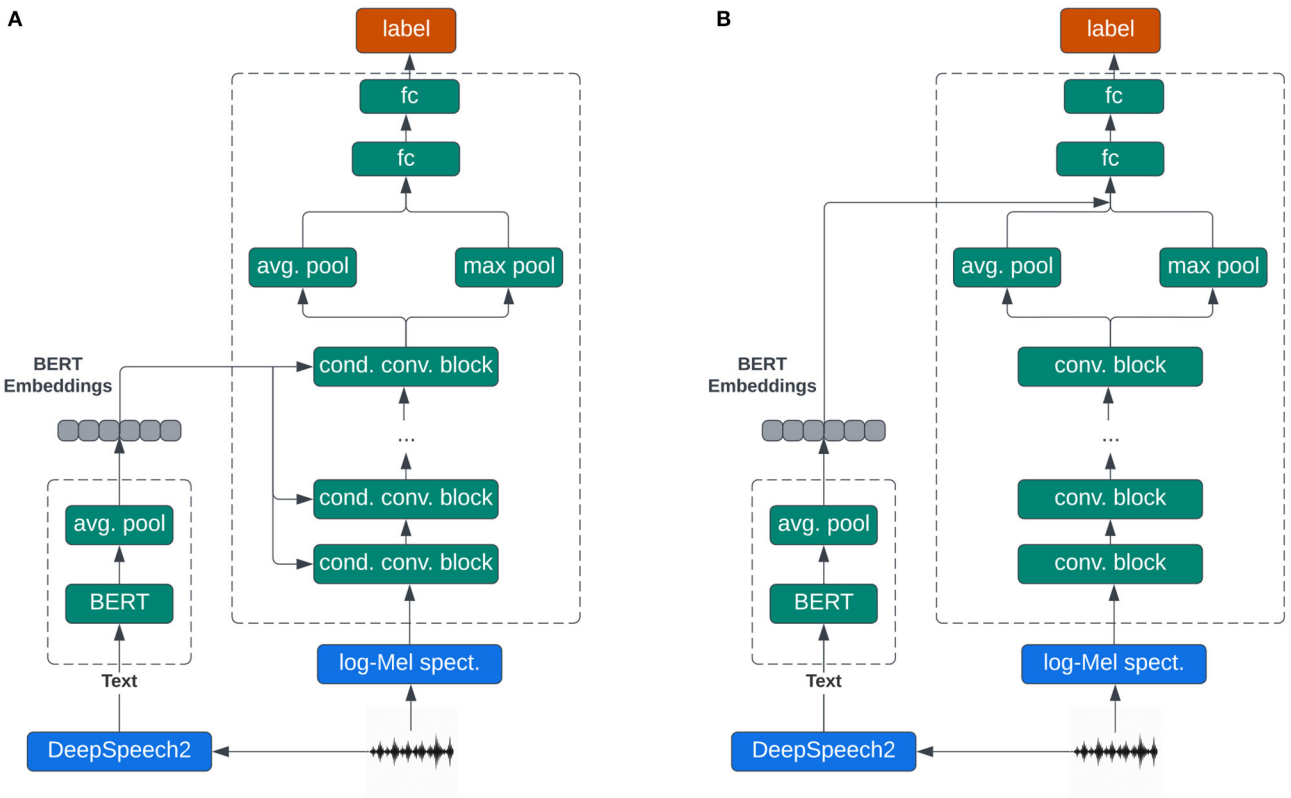

FIGURE 1
Diagrams of the architectures used in this work illustrating the processing of a single utterance. **(A)** MFCNN14: Multistage fusion architecture. **(B)** SFCNN14: Single-stage fusion architecture. The same architecture was used for the baseline CNN14, but without the fusion.

number of time windows and frequency bins, respectively, and the average BERT embeddings $\boldsymbol{E}_{\mathrm{L}} \in \mathbb{R}^{\mathrm{L_{dim}}}$, the output, $\boldsymbol{Y} \in \mathbb{R}^{T_{\mathrm{out}} \times F_{\mathrm{out}}}$, is computed as follows:

$$\boldsymbol{H}_1 = \mathrm{ReLU}(\mathrm{BN}(\mathrm{CONV}(\boldsymbol{X}))), \quad (1)$$

$$\boldsymbol{H}_2 = \mathrm{ReLU}(\mathrm{BN}(\mathrm{CONV}(\boldsymbol{H}_1))), \quad (2)$$

$$\boldsymbol{H_3} = \mathrm{MaxPool}(\boldsymbol{H}_2), \text{ and} \quad (3)$$

$$\boldsymbol{Y} = \boldsymbol{H}_3 + \mathrm{PROJ}(\boldsymbol{E}_{\mathrm{L}}), \quad (4)$$

Where ReLU stands for the rectified linear unit activation function (Nair and Hinton, 2010), BN for batch normalization (Ioffe and Szegedy, 2015), and CONV for 2D convolutions. The projection (PROJ) is implemented as a trainable linear layer which projects the input embeddings $\boldsymbol{E}_{\mathrm{L}}$ to a vector, $\boldsymbol{E}_{\mathrm{P}} \in \mathbb{R}^{F_{\mathrm{out}}}$, with the same dimensionality as the output feature maps:

$$\boldsymbol{E}_{\mathrm{P}} = \boldsymbol{W} \times \boldsymbol{E}_{\mathrm{L}} + \boldsymbol{b}, \quad (5)$$

Where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the trainable weight and bias terms, respectively. Thus, this conditioning mechanism is tantamount to adding a unique bias term to each output feature map of each convolution block.

The single-stage fusion architecture integrates acoustics and linguistics at a single point: immediately after the output of the last CNN14 convolution layer. The linguistic embeddings are first projected to the appropriate dimension, and then added to the acoustic representations produced by the convolution network. The shallow fusion architecture is shown in Figure 1B.

3.2. Datasets

3.2.1. IEMOCAP

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008), a multimodal emotion recognition dataset collected from 5 pairs of actors, each acting a set of scripted and improvised conversations, resulting in a total of 10,039 utterances. It has been annotated for the emotional dimensions of *arousal*, *valence*, and *dominance* on a 5-point Likert scale, with individual ratings averaged over all annotators to produce the gold standard. It has also been annotated for the emotion categories of *neutral*, *excited*, *surprised*, *happy*, *frustrated*, *sad*, *angry*, *afraid*, and *disgusted*. The dataset additionally contains gold standard transcriptions, which we use in our experiments. As IEMOCAP does not contain official train/dev/test splits, we follow the established convention of evaluating using leave-one-speaker-out (LOSO) cross-validation (CV) (Poria et al., 2018; Latif et al., 2020; Priyasad et al., 2020; Tseng et al., 2021), where we use all utterances of each speaker once as the test set, each time using the utterances of their pair as the validation set, resulting in a total of 10 folds.

3.2.2. MSP-Podcast

MSP-Podcast (Lotfian and Busso, 2019) is a recently-introduced data set for SER. The dataset is constantly growing and new releases are made every year; we used version v1.7, which was the latest one available to us for our experiments. It is split into

speaker independent partitions, with a training set consisting of 38,179 segments, a development set of 7,538 segments, collected from 44 speakers (22 male–22 female), and a 12,902 segment test set, consisting of 60 speakers (30 male–30 female). The dataset has been annotated for the emotional dimensions of *arousal*, *valence*, *dominance*, as well as for the emotion categories of *angry*, *contemptful*, *disgusted*, *afraid*, *happy*, *neutral*, *sad*, and *surprised*. The emotional dimensions have been annotated on a 7-point Likert scale on the utterance level, and scores by individual annotators have been averaged to obtain a consensus vote. All experiments on MSP-Podcast are performed on the official train/dev/test splits.

As we have no ground truth transcriptions for MSP-Podcast, we generated them automatically using an open-source implementation of DeepSpeech2 (Amodei et al., 2016).[1] Whereas, other works have used more advanced, proprietary models (Pepino et al., 2020), we opted for a widely-used open-source alternative for reproducibility. However, this model achieves a worse automatic speech recognition (ASR) performance than that of proprietary models. As it has been shown by several previous works that the performance of text-based and fusion approaches improves with better ASR models (Yoon et al., 2018; Sahu et al., 2019), we also expect our method to yield correspondingly better results, and do not consider this a critical limitation of our work.

3.3. Experimental procedure

As discussed in Section 1, we considered both a categorical and a dimensional model of emotion. For discrete emotion recognition, most works on the two datasets considered here pursue a 4-class classification problem, utilizing the emotion classes of *{angry, happy, neutral, sad}*, while further fusing the emotion class of *excited* with that of *happy* for IEMOCAP (Atmaja et al., 2022). To make our results comparable, we followed this formulation as well. For these experiments, we report unweighted average recall (UAR) (%), the standard evaluation metric for this task which also accounts for class imbalance, and additionally show confusion matrices. To mitigate the effect of class imbalance, which, as seen in Figure 2, is particularly pronounced for MSP-Podcast, we used a weighted variant of cross-entropy, where the loss for each sample is weighted by the inverse frequency of the class it belongs to. For dimensional SER, where we have continuous values for the dimensions of arousal, valence and dominance, we formulated our problem as a standard regression task and evaluated based on concordance correlation coefficient (CCC)—the standard evaluation metric for dimensional emotion (Parthasarathy and Busso, 2017; Li et al., 2021)—and also trained with the CCC loss, which is averaged over the 3 dimensions (Parthasarathy and Busso, 2017; Li et al., 2021).

However, multi-tasking potentially entangles the three dimensions and therefore complicates our analysis. Moreover, whereas the CCC loss is widely used for emotional dimension modeling, it is not the standard loss for regression tasks. Thus, we begin by considering single-task models trained with the standard mean squared error (MSE) loss.

We performed our experiments by separately training on IEMOCAP and MSP-Podcast. As mentioned, we performed 10-fold LOSO CV for the first and use the official train/dev/test partitions for the latter. We also report cross-domain results. Cross-domain results were obtained by evaluating models trained with one dataset on the other. For MSP-Podcast, where a single model was trained, we evaluated it on the entire IEMOCAP dataset. For IEMOCAP, where 10 models were trained for each experiment, we evaluated all 10 of them on the test set of MSP-Podcast, and computed the average performance metric for each task. As the emotional dimensions of the two datasets are annotated with different scales (5-point scale for IEMOCAP and 7-point scale for MSP-Podcast), we evaluated cross-corpus performance using Pearson correlation coefficient (PCC) instead of CCC, as the former is unaffected by differences in the scale.

For each dataset and task, we thus always perform the following experiments:

- **CNN14**: the unimodal, acoustics-only baseline,
- **SFCNN14**: our single-stage fusion architecture,
- **MFCNN14**: our multistage fusion architecture.

All models were trained for 60 epochs with a learning rate of 0.01 and a batch size of 64 using stochastic gradient descent (SGD) with a Nesterov momentum of 0.9. We selected the model that performed best on each respective validation set. In order to avoid statistical fluctuations due to random seeds, we ran each experiment 5 times and report mean and standard deviation.

Additionally, for some configurations, we fine-tuned the pre-trained BERT model, a practice which has recently emerged as the standard linguistic baseline showing strong performance on several NLP tasks. As pre-trained model, we selected *bert-base-uncased* distributed by Huggingface[2] with a final linear layer. As in the other experiments, we use the weighted cross-entropy loss for classification, the MSE loss for single task regression, the MSE loss for single-task experiments, and the mean CCC loss averaged over all targets for multitask regression. For all conditions, the maximum token length was set to 128, and the batch size to 32. For fine-tuning, we chose the Adam optimiser with fixed weight decay (learning rate: $2e$–5, betas: 0.9 and 0.999, epsilon: $1e$–06, weight decay: 0.0, no bias correction), and a linear schedule with 1,000 total and 100 warmup steps. We trained for 4 epochs with early stopping based on a UAR or CCC decrease on the development set for classification and regression, respectively.

4. Results

In this section, we begin by presenting our results on emotional categories in Section 4.1 and dimensions in Section 4.2, followed by our analysis on dialogue acts in Section 4.3 and on scripted vs. improvised conversations in IEMOCAP in Section 4.4.

4.1. Emotional categories

We begin by considering the 4-class emotion classification problem discussed in Section 3. Table 1 presents in- and cross-domain results on MSP-Podcast and IEMOCAP for both unimodal baselines and both fusion methods. Interestingly, CNN14 performs

1 https://github.com/mozilla/DeepSpeech

2 https://huggingface.co/bert-base-uncased

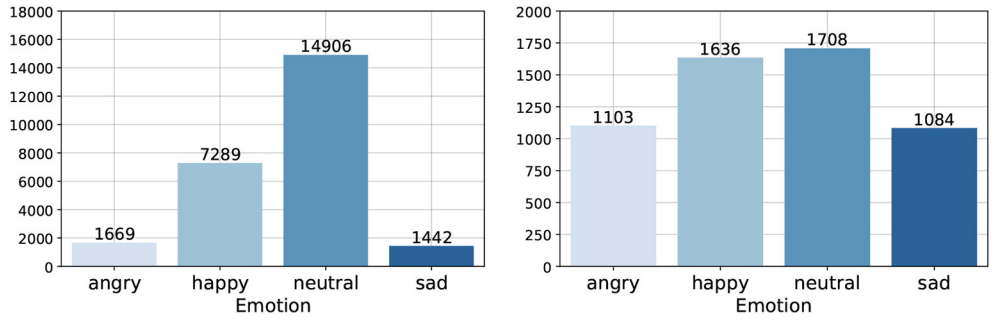

FIGURE 2
Distribution of emotion categories for MSP-Podcast **(left)** and IEMOCAP **(right)** for the herein considered 4-class classification problem using the {angry, happy, neutral, sad} categories. For IEMOCAP, "excited" samples are fused with "happy" ones.

TABLE 1 UAR(%) in- and cross-domain results for 4-class emotion recognition (chance level: 25 %).

Train	MSP-Podcast		IEMOCAP	
Test	MSP-Podcast	IEMOCAP	IEMOCAP	MSP-Podcast
	UAR	UAR	UAR	UAR
BERT	48.9 (0.4)	45.0 (0.7)	71.1 (0.2)	**40.5 (0.1)**
CNN14	48.3 (0.6)	41.3 (2.9)	55.8 (1.0)	31.5 (1.9)
SFCNN14	**56.0 (1.3)***†	**46.4 (3.9)**	64.1 (0.7)*†	33.3 (1.8)*†
MFCNN14	54.2 (0.8)*†	**46.4 (2.0)***	**72.6 (0.7)***†	35.6 (2.1)*†

MSP-Podcast in-domain results computed on the official test set, whereas IEMOCAP in-domain results correspond to leave-one-speaker-out cross-validation, where data from each speaker is used once as the test set and the other speaker in their session is used as the development set. Cross-domain results are reported on the official test set for MSP-Podcast and the entire dataset for IEMOCAP. Average (and standard deviation) results computed over 5 runs. Fusion results (SFCNN14 and MFCNN14) that are significantly different than the unimodal baselines (CNN14 and BERT) as determined by two-sided independent t-tests ($p < 0.05$) are marked by * and †, respectively. Bold values specify the best-performing model in each (sub-)table and column.

worse than BERT on both MSP-Podcast, with a UAR of 48.3 vs. 48.9 %, and on IEMOCAP (55.8 vs. 71.1 %), while also achieving the best cross-corpus performance when training on IEMOCAP and evaluating on MSP-Podcast. This shows that linguistics carry more emotional information on both datasets, but more so for IEMOCAP. On the one hand, this could be due to the noisy transcriptions used for MSP-Podcast. On the other hand, this dataset is more naturalistic than IEMOCAP, where actors could have relied more on text for conveying their emotions, especially in the case of scripted conversations.

Bimodal fusion leads to consistently higher performance compared to CNN14. Both architectures perform significantly better than the unimodal audio baseline for both datasets, with SFCNN14 performing slightly better on MSP-Podcast, and MFCNN14 considerably outperforming it on IEMOCAP. Moreover, in the case of IEMOCAP, only MFCNN14 is better than BERT, whereas SFCNN14 is significantly worse than it. In terms of cross-corpus results, the two fusion models yield the same performance when trained on MSP-Podcast and tested on IEMOCAP, with MFCNN14 being better on the reverse setup. In both cases, performance is severely degraded, which illustrates once more the challenges associated with cross-corpus AER.

With respect to the state-of-the-art, Pepino et al. (2020) outperforms our best result (59.1 vs. 56.0 %). We attribute this performance difference to the fact that they obtained their transcriptions with Google ASR, which has better performance than the open-source DeepSpeech2, and thus led to better performance on the linguistics. This is further corroborated by differences in the case of IEMOCAP as well, where our BERT model obtains better performance than their unimodal text model(71.1 vs. 55.2%), albeit with the caveats on scripted conversations raised by Pepino et al. (2020) and discussed further in Section 4.4. Thus, we expect both our BERT and fusion models to yield even better performance as the quality of transcriptions improves, in line with previous works (Yoon et al., 2018; Sahu et al., 2019; Amiriparian et al., 2021).

Priyasad et al. (2020), on the other hand, is showing overall stronger performance for IEMOCAP. However, their unimodal audio model is already substantially better than ours (69.9 vs. 55.8 %), indicating that our fusion approach could further improve when combined with a stronger unimodal baseline. Therefore, we conclude that while our performance does not surpass what is reported by other works, it still fares competitively well, and would benefit from improvements (e.g., better unimodal baselines or ASR) introduced in those works.

Figure 3 additionally shows the confusion matrices for the best performing models (based on the validation set) on the MSP-Podcast test set. For both CNN14 and BERT, we observe poor performance, with large off-diagonal entries. Notably, BERT is better at recognizing *happy* and *neutral* while CNN14 is better for *angry* and *sad*. For CNN14, the most frequent misclassifications occur when *happy* is misclassified into *angry* and *neutral* into *sad*. The latter is particularly problematic as more *neutral* samples are classified as *sad* (1, 596) than *neutral* (1, 216).

These problems are largely mitigated through the use of multimodal architectures. The improvements introduced by both SFCNN14 and MFCNN14 are mostly concentrated on the *happy* and *neutral* classes, where the true positive rate (TPR) improves by +33/39 and +47/27 %, respectively. However, the two architectures exhibit a different behavior on their off-diagonal entries. MFCNN14 substantially worsens the false positives on the *happy* class, with a large increase on the amount of *angry* (+35 %), *neutral* (+84 %), and *sad* (+122 %) samples misclassified as *happy*. In contrast, SFCNN14 exhibits only a minor deterioration (+13 %) on *neutral* to *happy* misclassifations. This illustrates that, although the UAR of both models, as shown in Table 1, is comparable for MSP-Podcast, the

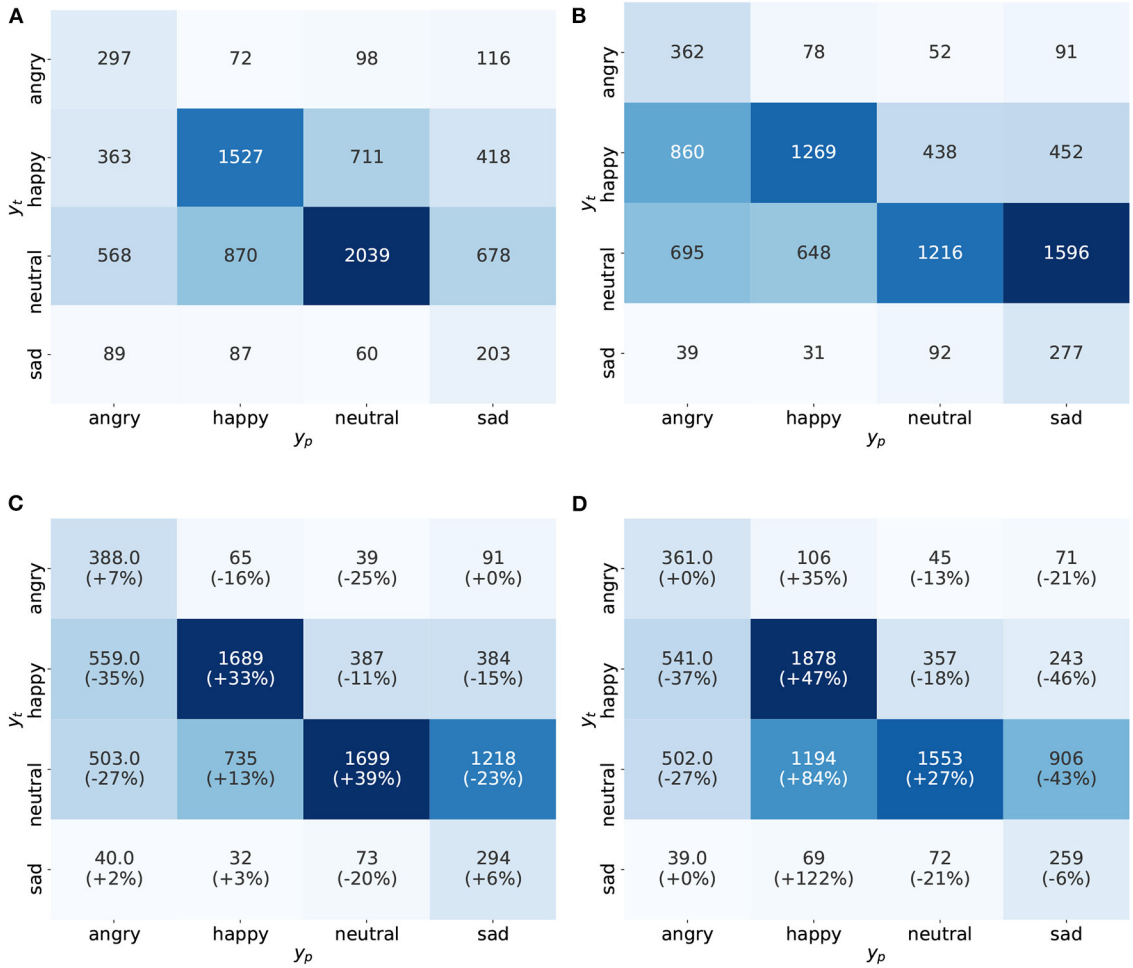

FIGURE 3
Confusion matrices for the 4-class emotion classification task on MSP-Podcast. For each approach, we show results for the best performing seed. For the two fusion models, we additionally show % change with respect to CNN14 for easier comparison. **(A)** BERT, **(B)** CNN14, **(C)** SFCNN14, and **(D)** MFCNN14.

more granular view provided by the confusion matrices clearly positions SFCNN14 as the winning architecture for this experiment.

Overall, our results clearly demonstrate that the proposed deep fusion methods can lead to substantial gains compared to their baseline, unimodal counterparts. MFCNN14 shows a more robust behavior with respect to the different datasets, whereas SFCNN14 shows more desirable properties in terms of confusion matrices.

4.2. Emotional dimensions

After evaluating our methods on categorical emotion recognition, we proceed with modeling emotional dimensions. As discussed in Section 3, we begin with single-task models trained with an MSE loss in Section 4.2.1, which allows us to study the effects of fusion independently for each dimension. Then, in Section 4.2.2, we investigate the combination of our methods with multi-task training and a CCC loss, which, in line with previous works, enables us to obtain better performance.

4.2.1. Single-task models

Our first experiments are performed on the emotional dimensions of MSP-Podcast and IEMOCAP. In Table 2, we report CCC and PCC results for in-domain and cross-domain performance, respectively. The performance of both fusion models is compared to that of the baseline CNN14 using two-sided independent sample t-tests.

The best performance on valence, both in- and cross-domain is achieved by MFCNN14, which reaches a mean CCC of 0.407 on MSP-Podcast and 0.664 on IEMOCAP. This is significantly higher than CNN14 and considerably outperforms SFCNN14, showing that multistage fusion can better utilize the textual information in this case. Cross-domain performance is severely degraded when training on IEMOCAP and testing on MSP-Podcast, while not so much when doing the opposite. This illustrates how training on large, naturalistic corpora leads to better generalization for SER systems, both unimodal and bimodal ones.

In the case of arousal, CNN14 performs better on MSP-Podcast with an average CCC of 0.664 (vs. 0.620 and 0.627 for SFCNN14 and MFCNN14), but this difference is not statistically significant. On IEMOCAP, MFCNN14 shows a marginally better performance than CNN14, while SFCNN14 is significantly worse than its unimodal baseline. This curious case shows how additional information can also hamper the training process. We hypothesize that this is because textual information is not conducive to arousal modeling, and leads it to perform worse on this task. This is corroborated by BERT models trained to jointly predict arousal/valence/dominance presented in

TABLE 2 CCC/PCC in-/cross-domain results for emotional dimension prediction using single-task models trained with an MSE loss.

Train	MSP-Podcast					
Test	MSP-Podcast			IEMOCAP		
	Arousal	Valence	Dominance	Arousal	Valence	Dominance
	CCC	CCC	CCC	PCC	PCC	PCC
CNN14	**0.664 (0.036)**	0.217 (0.010)	**0.583 (0.022)**	**0.593 (0.009)**	0.323 (0.018)	**0.471 (0.019)**
SFCNN14	0.620 (0.025)	0.367 (0.018)*	0.523 (0.019)*	0.543 (0.019)*	0.328 (0.026)	0.417 (0.014)*
MFCNN14	0.627 (0.010)	**0.407 (0.009)***	0.539 (0.011)*	0.558 (0.015)*	**0.381 (0.017)***	0.470 (0.017)
Train	IEMOCAP					
Test	IEMOCAP			MSP-Podcast		
	Arousal	Valence	Dominance	Arousal	Valence	Dominance
	CCC	CCC	CCC	PCC	PCC	PCC
CNN14	0.618 (0.010)	0.385 (0.015)	0.424 (0.020)	0.418 (0.008)	0.087 (0.010)	**0.438 (0.007)**
SFCNN14	0.551 (0.019)*	0.622 (0.007)*	0.438 (0.013)	0.271 (0.017)*	0.212 (0.004)*	0.210 (0.024)*
MFCNN14	**0.628 (0.005)**	**0.664 (0.005)***	**0.503 (0.006)***	**0.432 (0.006)***	**0.219 (0.005)***	0.365 (0.008)*

MSP-Podcast in- and cross-domain results computed on the test set, whereas IEMOCAP in-domain results correspond to leave-one-speaker-out cross-validation and cross-domain results reported on the entire dataset. Average (and standard deviation) results computed over 5 runs. Fusion results (SFCNN14 and MFCNN14) that are significantly different than the unimodal baseline (CNN14) as determined by two-sided independent sample t-tests ($p < 0.05$) are marked by *. Bold values specify the best-performing model in each (sub-)table and column.

Section 4.2.2. As mentioned in Section 3, we did not train BERT models for each dimension in isolation to reduce the computational load of our experiments; thus, we return to this point in Section 4.2.2.

Finally, we note that cross-domain performance for arousal, while also lower than in-domain performance, is not as low as valence, especially for CNN14. Interestingly, PCC on IEMOCAP for CNN14 models trained on MSP-Podcast is now significantly higher than the PCC obtained by either SFCNN14 or MFCNN14 (0.593 vs. 0.543 and 0.558, respectively). On the contrary, MFCNN14 shows significantly better performance on the opposite setup (0.432 vs. 0.418 PCC) than CNN14, while SFCNN14 remains significantly worse. This illustrates once more that the paralinguistic information stream carries more information on arousal than the linguistic one, and models trained on that can generalize better across different datasets.

Results on dominance follow the trends exhibited by arousal. CNN14 is significantly better than SFCNN14 and MFCNN14 on in-domain MSP-Podcast results (0.583 vs. 0.523 and 0.539), but, in the case of MFCNN14, this large in-domain difference does not translate to better cross-domain generalization, as both models are nearly equivalent on IEMOCAP PCC performance (0.471 vs. 0.470). This tendency is reversed on IEMOCAP; there MFCNN14 achieves significantly better in-domain results (0.503 vs. 0.424), but shows evidence of overfitting by performing significantly worse cross-domain (0.365 vs. 0.438).

Overall, our results show that bimodal fusion significantly improves performance on the valence dimension both in- and cross-domain for both datasets, with MFCNN14 achieving consistently superior performance to SFCNN14. For the case of MSP-Podcast, the other two dimensions fail to improve, while for IEMOCAP they improve only in-domain, and only for MFCNN14, while SFCNN14 performs consistently worse than CNN14. As we discuss in Section 4.2.2, this is because BERT is not good at modeling arousal and dominance, and this propagates to the fusion models. It thus appears that linguistic information, which by itself is not adequate to learn the tasks, hampers the training process and results in worse fusion models as well. This undesirable property seems to affect SFCNN14 more strongly than MFCNN14, which is able to circumvent it, and, in some cases, benefit from linguistic information. Thus, in the case of dimensional emotion recognition, MFCNN14 so far shows a better behavior than SFCNN14.

TABLE 3 CCC results for emotional dimension prediction using multi-task models on MSP-Podcast trained with a CCC loss.

Architecture	Arousal	Valence	Dominance
	CCC	CCC	CCC
BERT	0.232 (0.006)	0.503 (0.003)	0.214 (0.008)
CNN14	0.660 (0.012)	0.291 (0.029)	0.578 (0.011)
SFCNN14	0.665 (0.008)†	0.497 (0.013)*†	0.598 (0.025)†
MFCNN14	**0.678 (0.005)*†**	**0.521 (0.004)*†**	**0.604 (0.001)*†**

Models trained to jointly optimize the CCC for all dimensions. Average (and standard deviation) results reported over 5 runs. Fusion results (SFCNN14 and MFCNN14) that are significantly different than the unimodal baselines for each dimension as determined by two-sided independent sample *t*-tests ($p < 0.05$) are marked by * (for CNN14) and † (for BERT). Bold values specify the best-performing model in each (sub-)table and column.

4.2.2. Multi-task models

We end our section on emotional dimensions by considering a multi-task problem with a CCC loss. This is motivated by several recent works who have gotten better performance by switching to this formulation (Parthasarathy and Busso, 2017; Li et al., 2021). To reduce the footprint of our experiments, we only evaluate this approach on MSP-Podcast. CCC results for 5 runs are shown in Table 3. As previously discussed, Table 3 additionally includes results with a fine-tuned BERT model. As expected, we observe that BERT performs much better than CNN14 on valence prediction (0.503 vs. 0.291), but lacks far behind on arousal (0.232 vs. 0.660) and dominance (0.214 vs. 0.578). This clearly illustrates how the two

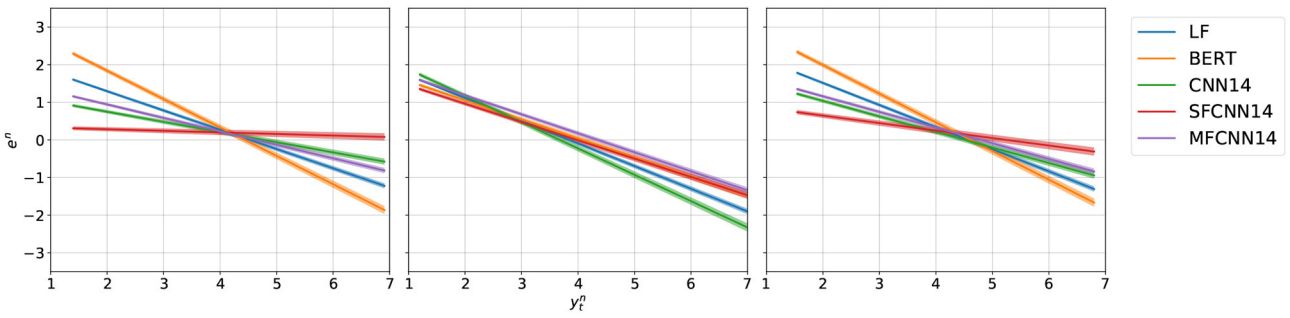

FIGURE 4
Error residual plots for arousal **(left)**, valence **(middle)**, and dominance **(right)**. Linear curves fitted on error residuals ($e^n = y_t^n - y_p^n$) for each architecture and dimension and plotted against the gold standard (y_t^n). With the exception of SFCNN14 on arousal, all models show higher errors toward the edges of their scale. Plots best viewed in color.

streams, acoustics and linguistics, carry complementary information for emotion recognition.

Both fusion methods improve on all dimensions compared to CNN14. In particular, MFCNN14 is significantly better on all three dimensions (0.678 vs. 0.660, 0.521 vs. 0.291, and 0.604 vs. 0.578), whereas SFCNN14 is significantly better only for valence (0.497 vs. 0.291) but not for arousal (0.665 vs. 0.660) and dominance (0.598 vs. 0.578). Moreover, of the two fusion methods, only MFCNN14 significantly improves on valence performance compared to BERT (0.521 vs. 0.503), while SFCNN14 performs marginally (but significantly) worse (0.497). This demonstrates once more that multistage fusion can better utilize the information coming from the two streams.

Finally, we are interested in whether the models show a heteroscedastic behavior by examining their error residuals. To this end, we pick the best models on the validation set and examine their residuals. Figure 4 shows fitted linear curves on the error residuals for each model and task. We use fitted curves for illustration purposes as superimposing the scatterplots for each model would make our plots uninterpretable. The curves are least-squares estimates over all error residuals. Our analysis reveals that most models show non-uniform errors, with their deviation from the ground truth increasing as we move away from the middle of the scale. This 'regression toward the mean' phenomenon is highly undesirable, especially for real-world applications where users would observe a higher deviation from their own perception of a target emotion for more intense manifestations of it.

SFCNN14 is the only model which escapes this undesirable fate, primarily for arousal and dominance. BERT, in contrast, shows the worst behavior for those two dimensions, but is comparable to SFCNN14 for valence. CNN14 and MFCNN14 show improved a much behavior compared to BERT, but some bias still appears, with the late fusion baseline naturally falling in the middle between BERT and CNN14. For valence, SFCNN14 and MFCNN14 both closely follow BERT in showing a low, but nevertheless existent bias, while CNN14 performs worse.

Interestingly, the residuals are also showing an asymmetric behavior. For valence, in particular, CNN14 is showing higher errors than the other models for the upper end of the scales, but is comparable to them for the lower end. Conversely, BERT shows higher errors for the lower end of arousal and dominance. This indicates that models struggle more with the scales. Naturally, this is partly explained by the sparseness of data for the more extreme values, as naturalistic datasets tend to be highly imbalanced toward neutral. Nevertheless, this continues to pose a serious operationalisation problem for AER systems.

4.3. Dialogue act analysis

We continue our analysis with an investigation of the interactions between dialogue acts and model performance. This gives us another lens with which to analyze model behavior w.r.t.the added linguistic knowledge.

Previous studies in psychology have established that the interaction between acoustic descriptors and emotional states depends on the linguistic content of an utterance. In addition to the case of questions discussed in Section 1 (Scherer et al., 1984), the *frequency code* hypothesis postulated by Ohala (1994), if applied to dialogue acts, predicts low or falling F0 in dialogue acts with inherently higher dominance (e.g., statements expressing a stance). For low-dominance dialogue acts in contrast (e.g., yes/no questions when expressing a request to the interlocutor), the hypothesis predicts high or rising F0. Such characteristic prosodic properties, which are utilized in dialogue act classification models (Shriberg et al., 1998), constrain the range of the unmarked dialogue act acoustics. To this end, linguistic information can establish reference points, with paralinguistics impacting speech acoustics relative to those. For example, in many languages statements are characterized by a falling F0. A rising F0 in such statements can thus work as a paralinguistic signal indicating increased insecurity and low dominance. Conversely, in many languages F0 in yes/no questions is usually rising. A falling F0 in these questions would then serve as a paralinguistic signal indicating impatience and high dominance.

This means that for certain emotional categorisations, acoustic information alone is not sufficient. This raises the question whether the benefits in performance obtained by adding linguistic information are exclusively attributable to the predictive power of BERT embeddings, which have been shown to contain information relevant for several natural language understanding (NLU) tasks

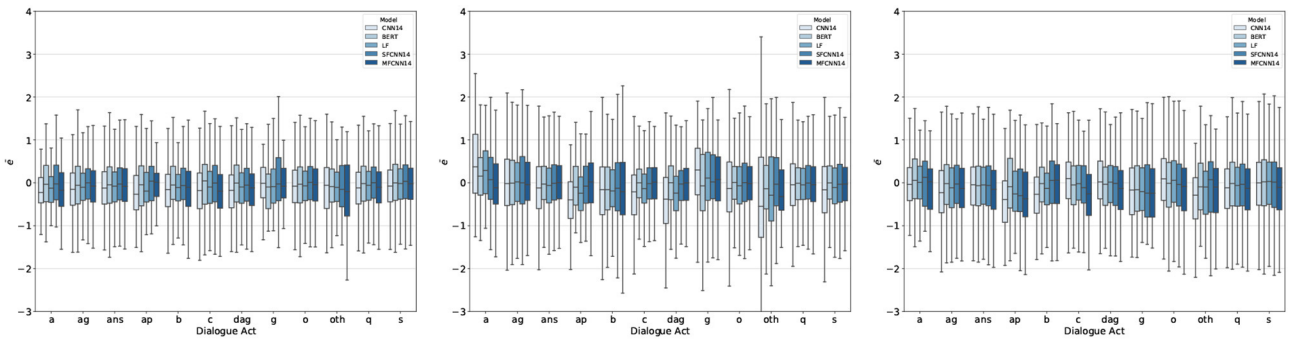

FIGURE 5
Box plots showing the distribution of error residuals ($\bar{e} = \sum_1^N e^n = \sum_1^N y_t^n - y_p^n$) for emotional dimension prediction on IEMOCAP for **(left** to **right)** CNN14, BERT, LF, SFCNN14, and MFCNN14. Error residuals show if models are consistently under- ($e^n > 0$) or over-predicting ($e^n < 0$) the respective dimension for specific dialogue acts. Dialogue act taxonomy comprises of: acknowledgment (a), agreement (ag), answer (ans), apology (ap), backchanneling (b), command (c), disagreement (d), greeting (g), opinion (o), question (q), statement (non-opinion) (s), and other (oth).

(Wang et al., 2018), or if information about the linguistic structure of the utterance is important as well.

One particular facet of the linguistic structure that is known to impact acoustics are dialogue acts. Dialogue acts can be predicted both by prosodic and lexical cues, with the latter being more effective (Shriberg et al., 1998; Stolcke et al., 2000). Moreover, (Saha et al., 2020) recently investigated the relationship between dialogue acts and text-based emotion recognition using two emotional datasets, one of them IEMOCAP, and found that jointly modeling both can improve emotion recognition performance. They released their dialogue act annotations,[3] thus opening a new avenue of exploration for the interaction between the underlying linguistic content and model performance. In total, the following dialogue acts have been annotated: *acknowledgment*, *agreement*, *answer*, *apology*, *back-channeling*, *command*, *disagreement*, *greeting*, *opinion*, *question*, *statement (non-opinion)*, and *other*.

We use this additional information to investigate hidden biases in our dimensional models, and investigate how the proposed fusion methods deal with them. Specifically, we are interested in whether model performance changes w.r.t.the dialogue act. This is measured by looking at the *error residual* of each sample in the dataset:

$$e^n = y_t^n - y_p^n, \tag{6}$$

Where y_t^n and y_p^n are the label and prediction for sample n, respectively. A negative error means that the model is over-predicting, whereas a positive one is an indication of under-prediction.

What we are interested in is whether a given model is consistently over-/under-predicting for specific dialogue acts. In particular, the presence of bias for the acoustics-only CNN14, and its subsequent alleviation by the introduction of linguistic information, would be indicative of more complex interactions between acoustics and linguistics, rather than the linguistic information being merely used for its predictive power. As an example, consider the case discussed by Scherer et al. (1984). An acoustic model that was confronted by two different ways to express 'positivity' (i.e., , valence), namely both rising and falling F0 contours for different types of questions, would struggle to learn this complex relationship, which would require some understanding of the underlying linguistic context of an utterance. Failing to do that, it might revert to learning just one of those contradictory patterns, e.g., the one for which more data is available. Linguistic information would resolve this issue by helping to differentiate between wh- and yes/no questions; this would inform the bimodal model to treat the two question types differently and learn the correct acoustic pattern for each one.

To evaluate this, we investigate the best-performing IEMOCAP models (based on overall CV performance) introduced in Section 4.2.1. In addition, we train a single BERT model (using LOSO CV) for each dimension of IEMOCAP, and furthermore compute late fusion (LF) results by averaging the predictions of BERT and CNN14; this forms our simple, shallow fusion baseline.

Figure 5 shows the distribution of residuals for each dialogue act. For a systematic evaluation of bias, we also performed two-sided one-sample t-tests for a sample mean of 0 of the error residuals. Rejecting the null hypothesis means the model shows a biased prediction for a particular dialogue act. We use a significance level of 0.05. P-values are shown in Table 4.

We begin our discussion by considering each dimension separately, starting with arousal. CNN14 and LF are generally over-predicting for most dialogue acts. In contrast, BERT is only moderately over-predicting for answers ($p = 0.005$) and opinions ($p = 0.002$). SFCNN14 improves on that by only (slightly) over-predicting on answers ($p = 0.043$). On the contrary, MFCNN14 is showing a highly biased behavior and follows an over-prediction trend.

For valence, CNN14 and LF again show a biased behavior for all acts except agreement ($p = 0.627$) and question ($p = 0.104$). Here, however, the biases are more inconsistent; for instance, acknowledgments and greetings are under-predicted, whereas apologies and disagreements are over-predicted. BERT is once more showing a balanced behavior; this time, the null hypothesis is only rejected for backchanneling ($p = 0.027$). This is now mirrored by both SFCNN14 and MFCNN14, for which the null hypothesis is rejected only for agreement ($p = 0.037$) and backchanneling ($p = 0.021$), respectively.

Dominance is the only dimension for which CNN14 and LF show a balanced behavior. BERT is once again stable, with the exception

3 https://github.com/sahatulika15/EMOTyDA

TABLE 4 *P*-values from a two-sided *t*-test for 0-mean error residuals for emotional dimension prediction on IEMOCAP stratified for the different dialogue acts.

	Arousal					Valence					Dominance				
	CNN14	BERT	LF	SF	MF	CNN14	BERT	LF	SF	MF	CNN14	BERT	LF	SF	MF
a	0.005	0.961	0.079	0.706	0.031	0.001	0.396	0.007	0.242	0.631	0.644	0.517	0.907	0.773	0.206
ag	0.000	0.402	0.000	0.928	0.002	0.627	0.585	0.545	0.037	0.422	0.000	0.330	0.000	0.320	0.000
ans	0.000	0.005	0.000	0.043	0.000	0.000	0.836	0.017	0.544	0.603	0.000	0.000	0.000	0.000	0.000
ap	0.000	0.971	0.060	0.897	0.520	0.000	0.909	0.003	0.540	0.896	0.000	0.464	0.012	0.082	0.002
b	0.000	0.290	0.000	0.081	0.000	0.001	0.027	0.001	0.185	0.021	0.000	0.363	0.000	0.556	0.143
c	0.000	0.789	0.001	0.604	0.000	0.000	0.477	0.000	0.915	0.869	0.197	0.217	0.965	0.016	0.000
dag	0.000	0.571	0.000	0.068	0.000	0.000	0.929	0.000	0.423	0.578	0.281	0.056	0.628	0.163	0.000
g	0.915	0.590	0.715	0.728	0.621	0.005	0.316	0.318	0.440	0.665	0.147	0.148	0.125	0.120	0.014
o	0.000	0.002	0.000	0.879	0.000	0.000	0.507	0.000	0.661	0.313	0.000	0.001	0.384	0.000	0.000
oth	0.148	0.407	0.203	0.378	0.011	0.018	0.927	0.094	0.932	0.246	0.000	0.469	0.016	0.597	0.082
q	0.000	0.059	0.000	0.826	0.000	0.104	0.696	0.208	0.271	0.274	0.000	0.034	0.000	0.002	0.000
s	0.000	0.996	0.003	0.172	0.011	0.000	0.777	0.000	0.352	0.132	0.085	0.279	0.129	0.002	0.000

Dialogue act taxonomy comprises of: acknowledgment (a), agreement (ag), answer (ans), apology (ap), back-channeling (b), command (c), disagreement (dag), greeting (g), opinion (o), question (q), statement (non-opinion) (s), and other (oth). Results are presented for CNN14, BERT, their late fusion (LF), SFCNN14 (SF), and MFCNN14 (MF).

of answers ($p < 0.001$), opinions ($p = 0.001$), and questions ($p = 0.034$). The same acts show up as biased for SFCNN14 with the addition of commands ($p = 0.016$) and statements ($p = 0.002$). In contrast, the null hypothesis is rejected for all acts except acknowledgment ($p = 0.206$), backchanneling ($p = 0.143$), and other ($p = 0.082$) for MFCNN14.

We now attempt to aggregate these disparate observations and form a more coherent picture. It is evident that CNN14 is highly inconsistent and shows a biased (over-/under-predicting) behavior for most dialogue acts. This also occurs for the tasks of arousal and dominance, where it is showing an overall high performance. Especially for valence, its behavior is more erratic; it is severely over-predicting for some dialogue acts while under-predicting for others. On the contrary, BERT is showing an overall balanced behavior across all dialogue acts for all dimensions. Interestingly, the late fusion model seems to closely follow the behavior of CNN14 for all dimensions. This makes sense as we are combining a biased (CNN14) with an unbiased (BERT) estimator. On the other hand, the deep fusion methods, and in particular SFCNN14, seem to strike a better balance between BERT and CNN14. Especially for valence, to which linguistic features are more suited, we observe an unbiased behavior for almost all acts.

Turning back to the original question of whether linguistic information primarily helps because of its raw predictive power, or whether it provides additional information w.r.t.the underlying linguistic content, the evidence is mixed. On the one hand, BERT is by itself performing strongly on all emotional dimensions and shows a low bias on all of them. Thus, any benefits could be attributable to its absolute predictive power, rather than information on the utterance type; the fusion models merely utilize this information to improve their predictions on some samples for which BERT does well, and thus appear to have a lower bias. However, even on the cases of arousal and dominance, where adding linguistic information does not improve performance compared to the audio-only baseline, deep fusion methods become more unbiased—SFCNN14 more so than MFCNN14. Moreover, the late fusion baseline does not benefit from the debiasing effects of BERT, even though it achieves the lowest MSE on all three dimensions. Therefore, although we cannot definitively conclude that deep fusion methods indirectly benefit from information on the underlying linguistic content, there are nevertheless intriguing findings pointing to this direction which warrant a closer investigation in future work.

4.4. IEMOCAP: Scripted vs. improvised

We conclude our investigation by discussing an important point raised in Pepino et al. (2020). As mentioned in Section 3.2, IEMOCAP consists of both improvised and scripted conversations. When doing LOSO CV, the same scripts can find themselves in both the training and the testing partition. This introduces an amount of information leakage that benefits the linguistic models, who could exploit this spurious correlation and obtain better performance.

Rather than modifying the standard 10 LOSO folds as Pepino et al. (2020) did, we decided to investigate the effect of scripted conversations *post-hoc*. We do this by performing a stratified evaluation for the best performing IEMOCAP models examined in Sections 4.1, 4.2.

The results of this evaluation are shown in Table 5. The first column shows UAR (%) on emotion categories. As Pepino et al. (2020) hypothesized, BERT is indeed showing substantially better performance for scripted conversations. However, we note that information leakage is not the only reason why this is expected. As discussed in Section 3, BERT was pretrained on written text data and not on spontaneous speech transcriptions; thus, we expect it to perform better on scripted conversations which resemble written text more. Additionally, the scripted elicitations in IEMOCAP were selected to facilitate the emergence of the target emotions in text (Busso et al., 2008); therefore it is expected that text-based models perform better there.

TABLE 5 UAR (%) and CCC results for the best performing dimensional and categorical emotion models on IEMOCAP.

Architecture	Emotion [S/I]	Arousal [S/I]	Valence [S/I]	Dominance [S/I]
	UAR [%]	CCC	CCC	CCC
BERT	**73.5**/59.8	**0.444**/0.404	**0.757**/0.593	**0.482**/0.308
CNN14	49.5/**60.9**	**0.661**/0.637	0.333/**0.433**	**0.504**/0.455
SFCNN14	58.3/**69.0**	**0.609**/0.571	**0.698**/0.557	**0.536**/0.386
MFCNN14	71.9/**73.1**	**0.639**/0.636	**0.714**/0.625	**0.575**/0.461

Results are computed separately for the scripted (S) and improvised (I) utterances of IEMOCAP (reported as S/I). For each model and task, we highlight the type of utterances that perform best. Bold values specify the best-performing model in each (sub-)table and column.

TABLE 6 F_1 scores for different categorical emotions for the scripted (S) vs. improvised (I) utterances of IEMOCAP (reported as S/I).

Architecture	Angry	Happy	Neutral	Sad
	F_1	F_1	F_1	F_1
BERT	**82.5**/47.7	**79.2**/67.0	56.5/**65.4**	**74.6**/61.2
CNN14	**69.2**/51.4	32.7/**55.2**	32.5/**53.6**	48.5/**63.9**
SFCNN14	**72.3**/60.7	54.8/**68.5**	39.9/**66.9**	56.7/**71.0**
MFCNN14	**83.0**/62.2	**74.1**/73.9	52.6/**70.5**	72.0/**79.1**

For each model and emotion, we highlight the type of utterances that perform best. Bold values specify the best-performing model in each (sub-)table and column.

In contrast, CNN14 shows the opposite trend, with much higher performance on improvised conversations. This is in line with previous research showing that acoustic models do better for improvised speech (Neumann and Vu, 2017). SFCNN14 and MFCNN14 follow CNN14 and show better performance for this type of speech as well. Notably, MFCNN14 shows a more balanced behavior and performs almost equally well for both types of speech.

However, a different story emerges for the three emotional dimensions, shown in the last three columns of Table 5. Surprisingly, we observe that all models perform better on the scripted conversations for all dimensions. The only exception is CNN14 for the valence task. This is unexpected because acoustic models should remain unaffected by scripted conversations, and shows that information leakage does not tell the whole story.

We further investigate the effects of scripted speech on categorical emotions by inspecting the performance of each model in terms of individual emotions. To this end, we compute the F_1 score, which is the harmonic mean between precision and recall. Results are shown in Table 6.

Several interesting patterns emerge. All models are better at identifying angry samples from scripted conversations, and all of them are also better at identifying neutral samples from improvised conversations. For happy and sad samples, BERT clearly fares better on scripted conversations. In contrast, all other models perform better on improvised conversations for sad, with CNN14 and SFCNN14 additionally performing better for happy samples, as well. MFCNN14 shows almost identical behavior for happy samples in both speech types.

For the acoustic and fusion models, these findings are in line with those reported by Neumann and Vu (2017), who found angry samples to be best identified from scripted conversations, sad ones to be better predicted from improvised ones, and neutral to have an overall low accuracy. For BERT, these findings illustrate that information leakage, although an important issue, does not equally affect performance on all emotions.

We end our discussion on scripted vs. improvised speech by inspecting the distributions of categories and dimensions for these types of speech. In Figure 6, we show the distribution of categorical samples across the two categories, followed by the distribution of the three emotional dimensions within each category. We observe that anger is more highly represented in scripted conversations; this could explain the higher performance for it in this stratum. As expected, improvised conversations show more extreme (perceived) values on the associated dimensions. Anger is expressed with a higher arousal and happiness with a higher valence. However, as discussed in Section 4.2.2, all models struggle with the more extreme values of the dimensional scales. This could explain the lower performance of dimensional models for improvised conversations. Thus, our main takeaway is that (textual) information leakage from scripted conversations is not the only factor leading to over-optimistic performance, but also a) an increased prominence of some categories in scripted data, and b) the fact that improvised conversations show more extreme emotions on the dimensional scales, with which models were found to be struggling.

5. Discussion

In this section, we summarize our findings. Overall, our experiments have shown that deep fusion can significantly improve performance for emotion recognition. In the case of emotional categories (Section 4.1), SFCNN14 and MFCNN14 were consistently better than the audio-based CNN14. MFCNN14 was also consistently better than BERT, whereas SFCNN14 was only better for MSP-Podcast, where the linguistic model seemed to struggle with the lower quality transcriptions.

Interestingly, inspecting model performance on scripted vs. improvised dialogues revealed that BERT performed better on scripted conversations, whereas CNN14 and the fusion models performed better for improvised ones. Rather than attributing this purely to information leakage from the presence of the same scripts in training and test partitions, the breakdown of performance per emotion showed that this behavior is partly influenced by data imbalance. For example, a lot more angry samples appear for scripted dialogues than improvised ones (see Figure 6), leading all models to perform better for that sub-population. Moreover, BERT was pre-trained on written text and is thus expected to perform better for scripted conversations, rather than improvised ones which mostly resemble naturalistic speech; a fact which also accounts for its higher performance on that data. Notably, MFCNN14 showed an overall more balanced behavior which, combined with its higher quantitative performance for that database, showed that this fusion method has great potential to combine the benefits from its two input information streams.

MFCNN14 was additionally better than SFCNN14 and the late fusion baseline in terms of quantitative performance on emotional dimension prediction. The largest gains over CNN14 appeared on the valence dimension, which is expected as this dimension is better predicted by linguistic information. Surprisingly, when training single-task models for each dimension separately, performance

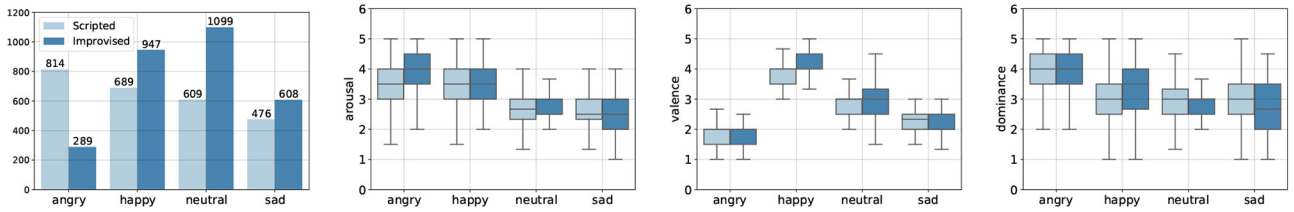

FIGURE 6
Comparison of scripted **(left)** vs. improvised **(right)** sessions of IEMOCAP. A countplot for the 4-class classification problem used here is shown on the left. The three plots on the right illustrate the distribution of arousal **(left)**, valence **(middle)**, and dominance **(right)** gold standard annotations for scripted vs. improvised sessions.

sometimes deteriorated for both SFCNN14 and MFCNN14, showing that adding linguistic information can also hamper the training process. This impacted the late fusion baseline as well. However, this was mitigated by multi-task training, where MFCNN14 positioned itself as the clear winner, leading to statistically significant benefits over both CNN14 and BERT for all emotional dimensions.

Our qualitative evaluations with respect to error residuals (Figure 4) and the influence of dialogue acts (Section 4.3) add more nuance to our analysis. MFCNN14 and SFCNN14 showed a less biased behavior than CNN14 w.r.t.the different dialogue acts, a trait not shared by the late fusion baseline. This demonstrates how deep fusion can lead to a tighter integration of the linguistic and paralinguistic streams and facilitate a better utilization of extra information. Additionally, SFCNN14 showed more uniform residuals, especially for arousal and dominance, which is a desirable property for real life application systems.

In comparison to the recent state-of-the-art, our proposed approach fares competitively, though not always better. Rather than a limitation of our proposed fusion method, we attribute this to experimental factors such as the choice of the ASR system (Pepino et al., 2020), the strength of the unimodal constituents (Priyasad et al., 2020; Siriwardhana et al., 2020b), and pre-training the acoustic model as well (Siriwardhana et al., 2020b; Li et al., 2021). Moreover, we chose not to jointly fine-tune the BERT model during fusion, so as to better isolate the benefits from adding linguistic information, but this has been shown to lead to additional benefits (Siriwardhana et al., 2020a,b). Thus, all these factors could be improved and lead to better performance for our method as well.

In conclusion, we find that single- and multistage fusion are both very competitive forms of fusion that outperform their constituent unimodal architectures and a shallow, late fusion baseline. In terms of absolute performance, multistage is consistently better than single-stage. These results are obtained for the specific domains of MSP-Podcast (naturalistic podcasts) and IEMOCAP (acted conversations). While MSP-Podcast is a large dataset by SER standards, our claims are nevertheless only valid for these particular domains. They are furthermore constrained to the English language and to overall high-quality recording conditions. Testing their generalization in a more diverse set is a necessary follow-up.

6. Conclusion

In the present contribution, we investigated the performance of deep fusion methods for emotion recognition. We introduced a novel method for combining linguistic and acoustic information for AER, relying on deep, multistage fusion of summary linguistic features with the intermediate layers of a CNN operating on log-Mel spectrograms, and contrasted it with a simpler, single-stage fusion one where information is only combined at a single point. We demonstrated both methods' superiority over unimodal and shallow, decision-level fusion baselines. In terms of quantitative evaluations, multistage fusion fares better than the single-stage baseline, thus illustrating how a tighter coupling of acoustics and linguistics inside CNNs can lead to a better integration of the two streams. This can be combined with recent advances on the use of attention (Zhao et al., 2021) and self-supervised pre-training (Li et al., 2021) to yield state-of-the-art performance on SER.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AT, FE, and BS conceptualized the fusion architecture. AT, UR, SL, and SH designed the experimental protocol and conducted model training and evaluation. AT and UR conducted the subsequent analysis. AT wrote the first draft. All authors revised and edited later draft versions and the final manuscript.

Funding

This work was partially funded from the DFG project No. 442218748 (AUDI0NOMOUS).

Conflict of interest

FE, BS, and AT have a patent pending for the multi-stage fusion architecture (US appl. no. 17542564).

AT, UR, SH, FE, and BS were employed by audEERING GmbH.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of bert-based approaches. *Artif. Intell. Rev.* 54, 5789–5829. doi: 10.1007/s10462-021-09958-2

Amiriparian, S., Sokolov, A., Aslan, I., Christ, L., Gerczuk, M., Hübner, T., et al. (2021). On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era. *arXiv preprint* arXiv:2104.10121. doi: 10.48550/arXiv.2104.10121

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2016). "Deep speech 2: end-to-end speech recognition in english and mandarin," in *Proceedings of ICML* (New York, NY), 173–182.

Atmaja, B. T., Sasou, A., and Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* 2022, 2. doi: 10.1016/j.specom.2022.03.002

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.* 16, 345–379. doi: 10.1007/s00530-010-0182-0

Baevski, A., Schneider, S., and Auli, M. (2019). "vq-wav2vec: self-supervised learning of discrete speech representations," in *Proceedings of ICLR* (New Orleans, LA).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluat.* 42, 335. doi: 10.1007/s10579-008-9076-6

Calvo, R. A., and D'Mello, S. (2010). Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* 1, 18–37. doi: 10.1109/T-AFFC.2010.1

Chen, F., Luo, Z., Xu, Y., and Ke, D. (2019). Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv preprint* arXiv:1904.08138. doi: 10.48550/arXiv.1904.08138

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of ACL* (Austin, TX), 4171–4186.

Ekman, P. (1992). An argument for basic emotions. *Cogn. Emot.* 6, 169–200. doi: 10.1080/02699939208411068

Fayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks* 92, 60–68. doi: 10.1016/j.neunet.2017.02.013

Georgiou, E., Papaioannou, C., and Potamianos, A. (2019). "Deep hierarchical fusion with application in sentiment analysis," in *Proceedings of INTERSPEECH* (Graz), 1646–1650.

Gfeller, B., Roblek, D., and Tagliasacchi, M. (2020). One-shot conditional audio filtering of arbitrary sounds. *arXiv preprint* arXiv:2011.02421. doi: 10.1109/ICASSP39728.2021.9414003

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML* (Lille), 448–456.

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of CVPR* (Long Beach, CA), 4401–4410.

Keren, G., Han, J., and Schuller, B. (2018). "Scaling speech enhancement in unseen environments with noise embeddings," in *Proceedings of CHiME Workshop on Speech Processing in Everyday Environments* (Montreal, QC), 25–29.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 28, 2880–2894. doi: 10.1109/TASLP.2020.3030497

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Epps, J., and Schuller, B. W. (2020). "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," in *IEEE Transactions on Affective Computing*. Available online at: https://ieeexplore.ieee.org/abstract/document/9052467

Lee, C. W., Song, K. Y., Jeong, J., and Choi, W. Y. (2018). "Convolutional attention networks for multimodal emotion recognition from speech and text data," in *Proceedings of ACL* (Melbourne, VIC), 28. Available online at: https://aclanthology.org/W18-3304/

Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., et al. (2021). "Contrastive unsupervised learning for speech emotion recognition," in *Proceedings of ICASSP* (Toronto, ON: IEEE), 6329–6333.

Lin, W.-C., and Busso, C. (2021). Chunk-level speech emotion recognition: a general framework of sequence-to-one dynamic temporal modeling. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2021.3083821

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692

Lotfian, R., and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.* 10, 471–483. doi: 10.1109/TAFFC.2017.2736999

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Proceedings of NeurIPS* (Tahoe, CA), 3111–3119.

Munezero, M., Montero, C. S., Sutinen, E., and Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.* 5, 101–111. doi: 10.1109/TAFFC.2014.2317187

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of ICML* (Haifa), 807–814.

Neumann, M., and Vu, N. T. (2017). "Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech," in *Proceedings of INTERSPEECH* (Stockholm), 1263–1267. Available online at: https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/0917.PDF

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. *Sound Symbolism* 2, 325–347. doi: 10.1017/CBO9780511751806.022

Parthasarathy, S., and Busso, C. (2017). "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proceedings of INTERSPEECH* (Stockholm), 1103–1107.

Pepino, L., Riera, P., Ferrer, L., and Gravano, A. (2020). "Fusion approaches for emotion recognition from speech using acoustic and text-based feature," in *Proceedings of ICASSP* (Barcelona: IEEE), 6484–6488.

Perikos, I., and Hatzilygeroudis, I. (2013). "Recognizing emotion presence in natural language sentences," in *International Conference on Engineering Applications of Neural Networks* (Halkidiki: Springer), 30–39. Available online at: https://link.springer.com/chapter/10.1007/978-3-642-41016-1_4

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. *Inf. Fusion* 37, 98–125. doi: 10.1016/j.inffus.2017.02.003

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., and Hussain, A. (2018). Multimodal sentiment analysis: addressing key issues and setting up the baselines. *IEEE Intell. Syst.* 33, 17–25. doi: 10.1109/MIS.2018.2882362

Priyasad, D., Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2020). "Attention driven fusion for multi-modal emotion recognition," in *ICASSP* (Barcelona: IEEE), 3227–3231.

Russell, J. A., and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *J. Res. Pers.* 11, 273–294. doi: 10.1016/0092-6566(77)90037-X

Saha, T., Patra, A., Saha, S., and Bhattacharyya, P. (2020). "Towards emotion-aided multi-modal dialogue act classification," in *Proceedings of ACL* (Seattle, WA: Association for Computational Linguistics), 4361–4372. Available online at: https://aclanthology.org/2020.acl-main.402/

Sahu, S., Mitra, V., Seneviratne, N., and Espy-Wilson, C. Y. (2019). "Multi-modal learning for speech emotion recognition: an analysis and comparison of asr outputs with ground truth transcription," in *Proceedings of INTERSPEECH* (Graz), 3302–3306.

Scherer, K. R. (2000). Psychological models of emotion. *Neuropsychol. Emot.* 137, 137–162.

Scherer, K. R., Ladd, D. R., and Silverman, K. E. (1984). Vocal cues to speaker affect: testing two models. *J. Acoust. Soc. Am.* 76, 1346–1356. doi: 10.1121/1.391450

Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005). "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles," in *Proceedings of European Conference on Speech Communication and Technology* (Lisbon). Available online at: https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2005/i05_0805.pdf

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., et al. (2013). "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH* (Lyon), 148–152.

Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* 61, 90–99. doi: 10.1145/3129340

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., et al. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Lang Speech*. 41, 443–492. doi: 10.1177/002383099804100410

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., and Nanayakkara, S. (2020a). Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* 8, 176274–176285. doi: 10.1109/ACCESS.2020.3026823

Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. (2020b). "Jointly fine-tuning "bert-like" self supervised models to improve multimodal speech emotion recognition," in *Proceedings of Interspeech* (Shanghai), 3755–3759. Available online at: https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/1212.pdf

Steidl, S., Batliner, A., Schuller, B., and Seppi, D. (2009). "The hinterland of emotions: facing the open-microphone challenge," in *Proceedings of ACII* (Amsterdam: IEEE), 1–8.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist*. 26, 339–373. doi: 10.1162/089120100561737

Strapparava, C., and Valitutti, A. (2004). "Wordnet affect: an affective extension of wordnet," in *Proceedings of ELREC, Vol. 4* (Lisbon), 1083–1086.

Triantafyllopoulos, A., Liu, S., and Schuller, B. W. (2021). "Deep speaker conditioning for speech emotion recognition," in *Proceedings of ICME* (Shenzhen: IEEE.), 1–6.

Triantafyllopoulos, A., and Schuller, B. W. (2021). "The role of task and acoustic similarity in audio transfer learning: insights from the speech emotion recognition case," in *Proceedings of ICASSP* (Toronto: IEEE), 7268–7272.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., et al. (2016). "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP 2016* (Shanghai: IEEE), 5200–5204.

Tseng, S.-Y., Narayanan, S., and Georgiou, P. (2021). Multimodal embeddings from language models for emotion recognition in the wild. *IEEE Signal Process. Lett*. 28, 608–612. doi: 10.1109/LSP.2021.3065598

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process*. 11, 1301–1309. doi: 10.1109/JSTSP.2017.2764438

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2018). "Wavenet: agenerative model for raw audio," in *Proceedings of ISCA Speech Synthesis Workshop* (Sunnyvale, CA), 125–125.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of NeurIPS* (Long Beach, CA), 5998–6008.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Eyben, F., and Schuller, B. W. (2022). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *arXiv preprint* arXiv:2203.07378. doi: 10.48550/arXiv.2203.07378

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). "Glue: a multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of EMNLP* (Brussels), 353–355.

Yang, K., Xu, H., and Gao, K. (2020). "CM-BERT: cross-modal bert for text-audio sentiment analysis," in *Proceedings of ACM Multimedia* (Seattle, WA), 521–528.

Yoon, S., Byun, S., and Jung, K. (2018). "Multimodal speech emotion recognition using audio and text," in *Proceedings of SLT* (Stuttgart: IEEE), 112–118.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). "Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of ACL* (Melbourne, VIC), 2236–2246.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2008). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell*. 31, 39–58. doi: 10.1109/TPAMI.2008.52

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: a survey. *Wiley Interdisc. Rev*. 8, e1253. doi: 10.1002/widm.1253

Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., et al. (2021). Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition. *Neural Networks* 141, 52–60. doi: 10.1016/j.neunet.2021.03.013