Versatile Audio-Visual Learning for Emotion Recognition

Lucas Goncalves, *Student Member, IEEE*, Seong-Gyun Leem, *Student Member, IEEE*, Wei-Cheng Lin, *Student Member, IEEE*, Berrak Sisman, *Member, IEEE*, and Carlos Busso, *Fellow, IEEE*

***Abstract*—Most current audio-visual emotion recognition models lack the flexibility needed for deployment in practical applications. We envision a multimodal system that works even when only one modality is available and can be implemented interchangeably for either predicting emotional attributes or recognizing categorical emotions. Achieving such flexibility in a multimodal emotion recognition system is difficult due to the inherent challenges in accurately interpreting and integrating varied data sources. It is also a challenge to robustly handle missing or partial information while allowing direct switch between regression or classification tasks. This study proposes a *versatile audio-visual learning* (VAVL) framework for handling unimodal and multimodal systems for emotion regression or emotion classification tasks. We implement an audio-visual framework that can be trained even when audio and visual paired data is not available for part of the training set (i.e., audio only or only video is present). We achieve this effective representation learning with audio-visual shared layers, residual connections over shared layers, and a unimodal reconstruction task. Our experimental results reveal that our architecture significantly outperforms strong baselines on the CREMA-D, MSP-IMPROV, and CMU-MOSEI corpora. Notably, VAVL attains a new state-of-the-art performance in the emotional attribute prediction task on the MSP-IMPROV corpus.**

***Index Terms*—Audio-visual modeling, handling missing modalities, multimodal emotion recognition, transformers, versatile learning.**

I. Introduction

EFFECTIVE human interactions often include the expression and perceptions of emotional cues conveyed across multiple modalities to accurately comprehend and convey a message. During human interactions, there are two modalities that stand out as particularly significant: speech and facial expressions. These modalities play a crucial role in facilitating communication, making it essential for emotion recognition systems to incorporate speech-based cues [1] and facial expression-based cues [2], [3]. These modalities are intrinsically connected [4], proving complementary information [5]. Therefore, emotion recognition systems can be more accurate if they incorporate audio-visual solutions [6], [7], [8], [9], mirroring the way humans interact in natural, real-world settings.

It is important to note that while humans primarily rely on these two modalities to recognize emotional states, there are scenarios in which only one modality may be utilized. In some cases, visual cues might be unavailable or occluded, leaving individuals to rely solely on acoustic information. Conversely, there might be situations where acoustic cues are insufficient or absent, requiring the exclusive use of visual cues for effective communication. As a result, it is vital for *human computer interaction* (HCI) systems to not only be capable of simultaneously handling both modalities, but also to adapt to unimodal situations. The straightforward approach is to have separate unimodal and multimodal solutions. However, it is more computationally effective to have versatile systems that can be adapted according to the available information. This flexibility ensures that HCI systems can accommodate a wide range of communication contexts and maintain their effectiveness in various real-world conditions.

Artificial intelligent (AI) methods have been explored for audio-visual learning to achieve high performance in either multimodal or unimodal scenarios [10], [11], [12]. Most conventional models rely on building separate modality-specific branches [13], [14], [15] or employing ensemble-like techniques [16], which can lead to convoluted and complex architectures. With the introduction of transformer frameworks [17], many solutions have been developed to avoid training with modality-specific strategies. These models even implement formulations that unify the cross-modality relationships [18], [19] into a single, comprehensive model.

Despite significant recent advancements in the construction of simpler multimodal models, there are still open research questions to be considered when constructing a unified multimodal model. One of these challenges is that current methods tend to solely focus on one type of *machine learning* (ML) task [20]. For example, they consider either classification or regression problems, without considering the potential need for versatility across different problem types. In emotion recognition, this problem is especially important, since emotions can be alternatively described with emotional attributes (e.g., arousal, valence, dominance) and categorical emotions (e.g., anger, happiness, sadness) [21].

Depending on the setting, therefore, it is important to have a model that can be utilized for either regression (emotional attributes) or classification (categorical emotions) settings

Manuscript received 10 May 2023; revised 23 May 2024; accepted 15 July 2024. Date of publication 25 July 2024; date of current version 7 March 2025. This work was supported by NSF under Grant CNS-2016719. Recommended for acceptance by F. Metze. *(Corresponding author: Carlos Busso.)*

The authors are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: goncalves@utdallas.edu; seong-gyun.leem@utdallas.edu; wei-cheng.lin@utdallas.edu; berrak.sisman@utdallas.edu; busso@utdallas.edu).

Digital Object Identifier 10.1109/TAFFC.2024.3433386

without making major changes in the architecture. Moreover, current methods often disregard the importance of maintaining distinct representations for each modality to capture modality-specific [22] features and characteristics within shared layers, which could affect performance in unimodal settings [13]. By failing to account for the unique characteristics of each modality, the models may not optimally leverage the strengths of each input type, ultimately limiting their effectiveness. Moreover, the complexity of many existing models can be a drawback, as it may lead to increased computational demands and reduced interpretability [23], making it more challenging for researchers and practitioners to analyze and adapt the models for various applications.

Our main contribution in this study is the proposal of a *versatile audio-visual learning* (VAVL) model, which unifies multimodal and unimodal learning in a single framework that can be used for emotion regression and classification tasks. Our approach utilizes branches that separately process each modality. In addition, we introduce shared layers, residual connections over shared layers, and a unimodal reconstruction task. The shared layers encourage learning representations that reflect the connections between the two modalities. The addition of the auxiliary reconstruction task and unimodal residual connections over the shared layers helps the model learn representations that reflect the heterogeneity of the modalities. Collectively, these components are added to our framework to help with the multimodal feature representation, which is a core challenge in multimodal learning [24]. We implement this audio-visual framework so it can be trained even when audio and visual paired data is not available for part of the data. The proposed approach is attractive as it enables the training of multimodal systems using incomplete information from multimodal databases (e.g., audio-only or visual-only information is available for some data points), as well as unimodal databases. The proposed framework also has the versatility to be used for either regression or classification tasks without changing the architecture or training strategy.

We quantitatively evaluate and compare the performance of our proposed model against strong baselines. The results demonstrate that our architecture achieves significantly better results on the CREMA-D [25], the CMU-MOSEI [22], and the MSP-IMPROV [26] corpora compared to the baselines. For example, VAVL achieves a new state-of-the-art result for the emotional attribute prediction task on the MSP-IMPROV corpus. Also, our architecture is able to sustain strong performance in audio-only and video-only settings, compared to strong unimodal baselines. Additionally, we conduct ablation studies into our model to understand the effects of each component on our model's performance. These results show the benefits of our proposed framework for audio-visual emotion recognition.

II. Background

Emotion recognition has been widely explored using speech [1], [27], [28], facial expressions [29], and multimodal solutions [30], [31], [32], [33]. However, most of these models have focused on solving either emotion classification tasks [33] or emotional attribute prediction tasks [27]. Furthermore, these studies are designed for either unimodal or multimodal solutions. Although some recent studies [15] have proposed approaches for audio-visual emotion recognition that can handle both tasks, they still rely on a single modality. There is a need for solutions that can handle multiple modality settings. Recent advancements in deep learning have resulted in the development of unified frameworks that aim to move away from unimodal implementations. This section focuses on relevant studies that address similar problems investigated in this paper.

A. Formulations for Emotion Recognition

Although an individual's cultural and ethnic background may impact her/his manner of expression, studies have indicated that certain basic emotions are very similar regardless of cultural differences [34]. These basic emotions are described as: happiness, sadness, surprise, fear, disgust, and anger. These basic emotional states are often the most widely explored task in emotion recognition, formulating the problem as a six-class classification problem [35], [36]. In fact, many of the affective corpora include all or a subset of these classes, including the CREMA-D [25], AffectNet [37], and AFEW [38] databases.

An alternative way to describe emotions is to leverage the continuous space of emotional attributes or sentiment analysis. The emotional attributes approach identifies dimensions to describe expressive behaviors. The most common attributes are arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) [39]. Several databases have explored the emotional attribute or sentiment analysis annotation of audio-visual stimuli, such as the IEMOCAP [40], MSP-IMPROV [26], and CMU-MOSEI [22] corpora. Studies have proposed algorithms to predict emotional attributes using audio-visual stimuli [41], [42], [43], facial expression [44], and speech [1], [45], [46].

B. Unified Multimodal Framework

Since the introduction of the transformer framework [17], many studies have explored multimodal frameworks that can still work even when only one modality is available [47]. Some proposed architectures employ training schemes that involve solely utilizing separate branches for different modalities, as seen in Baevski et al. [48]. These approaches construct branches containing unimodal information specific to each modality. In contrast, other studies explore the use of shared layers that encapsulate cross-modal information from multiple modalities [49], [50]. The aforementioned studies incorporate aspects that laid the foundation for modeling a more unified framework, but they still have some limitations. For example, common solutions require independent modality training, assume that all the modalities are available during inference, or fail to fully address the heterogeneity present in a multimodal setting. Recent studies have proposed audio-visual frameworks that can handle both unimodal and multimodal settings [51], but they still rely on having additional unimodal network branches. Gong et al. [19] proposed a unified audio-visual framework for classification, which incorporates some effective capabilities, such as independently processing audio and video, and including shared

audio-visual layers into the model. However, the model focused only on a classification task.

C. Versatile Task Modeling

For audio-visual unified models [19] and audio-visual models that have multiple modality branches and can handle unimodal scenarios [12], the final multimodal prediction is often obtained by averaging the outputs from different branches, receiving the contributions from all the modalities. In emotion classification tasks, the model output is a probability distribution over a set of discrete categories. Averaging predictions from multiple unimodal branches can help capture different aspects of the data, leading to improved classification accuracy. However, this approach may not be effective for emotion attribute regression tasks, as regression involves predicting a continuous numerical value rather than a discrete categorical emotion label. Therefore, the average of the predictions from different unimodal branches may not be appropriate. This is particularly true when considering the performance gap that often exists between speech and visual models for predicting arousal, valence, and dominance, as facial-only features are generally less effective than acoustic features [11]. Simply averaging these two results could lead to suboptimal outcomes. A potentially appealing alternative is to use a weighted combination of the predictions [52], [53]. However, this approach also has its limitations, as the prediction discrepancy may not be consistent across all data points. To address this issue, our study proposes training an audio-visual prediction layer, in addition to visual and acoustic prediction layers, used when only data from one modality is available. The audio-visual prediction layer is optimized only when both modalities are available. This layer can automatically learn appropriate weights to combine the modality representations and fuse them to generate a single audio-visual prediction. This approach works well for both emotion classification and regression settings.

D. Relation to Prior Work

While the implementation of our approach is with transformers, which have been used in the past in multimodal processing, our proposed VAVL model represents a significant contribution in comparison to previous works. The proposed architecture with shared layers, residual connections, reconstruction auxiliary tasks and separate audiovisual prediction layers is novel, where each of its components is carefully motivated to address a fundamental challenge in multimodal machine learning. We employ conformer layers [54] as our encoders, which are transformer-based encoders augmented with convolutions. The closest study to our framework is the work of Gong et al. [19], which proposed a unified audio-visual model for classification. Their framework involved the independent processing of audio and video features, with shared transformer and classification layers for both modalities. Prior to the shared transformer layers, they have modality-specific feature extractors and optional modality-specific transformers for each modality. In contrast, our framework employs conformer layers instead of vanilla transformers. Moreover, we introduce three important components to our framework. First, an audio-visual prediction layer is utilized solely when audio-visual modalities are available during inference, enhancing prediction accuracy and model versatility for both regression or classification tasks. Second, a unimodal reconstruction task is implemented during training to ensure that the shared layers capture the heterogeneity of the modalities. Third, residual connections are incorporated over the shared layers to ensure that the unimodal representations are not forgotten in the shared layers. This strategy maintains high performance even in unimodal settings. These novel additions represent significant contributions to the field of audio-visual emotion recognition. Additionally, our proposed model comprises approximately 86 million parameters. In comparison, TLSTM has 485,000 parameters, SFAV has 642,000 parameters, MulT has 749,000 parameters, and Auxformer has 1,226,000 parameters. While our model is more complex than these earlier models, it offers a significant reduction in complexity compared to the UAVM framework, which has 117 million parameters. This comparison indicates a reduction in complexity of approximately 26% relative to UAVM, highlighting the efficiency of our model in handling sophisticated multi-modal tasks.

The proposed approach is significantly difference from our previous work on audio-visual emotion recognition [12]. Goncalves and Busso [12] proposed a method that incorporates two main components: the use of auxiliary unimodal networks and the use of modality dropout during training. The auxiliary networks ensure unimodal representations are separately embedded and not lost during the training of shared spaces within the cross-modal layers. The use of modality dropout enhances the ability of the model to retain performance when parts of the input data are missing. This model uses audiovisual paired data for training and performs averaging of prediction heads to obtain final predictions. In contrast, our approach utilizes distinct branches for processing each modality and introduces shared layers with residual connections, complemented by an unimodal reconstruction task. These shared layers are pivotal in facilitating the learning of representations that capture the interplay between audio and visual modalities. The shared layers take either visual or acoustic information, but not both, providing rich information to leverage cross-modality information. Furthermore, the incorporation of an auxiliary reconstruction task alongside unimodal residual connections enhances the model's ability to understand the diverse characteristics inherent in each modality. A key innovation of our approach lies in its ability to effectively train with incomplete multimodal data, accommodating scenarios where only audio or visual information is available for certain data points. This flexibility addresses a significant challenge in multimodal learning and broadens the applicability of our model, making it a substantial contribution to the field. The approach allows us to use partial audio or visual information from multimodal and unimodal databases.

In summary, in relation to previous works, the primary novelty of our framework lies within the unique organization and training methodology of its layers. Our approach strategically structures its layers in a manner that optimizes their functionality and efficacy for our specific application for audio-visual and unimodal (visual or acoustic) emotion recognition.

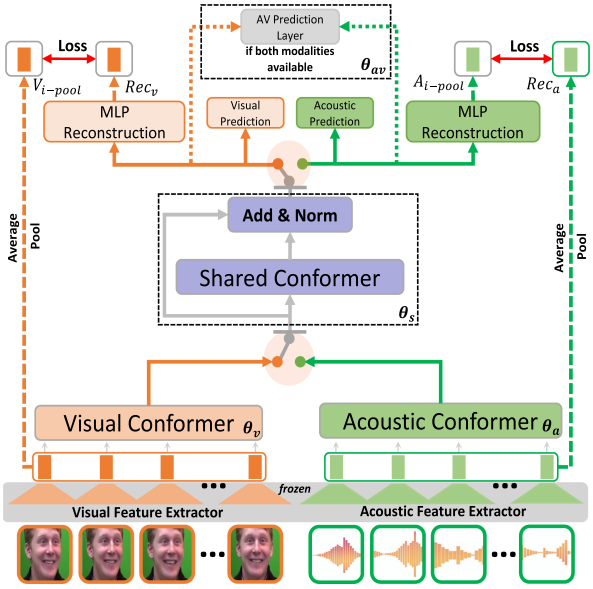


Fig. 1. Overview of our proposed *versatile audio-visual learning* (VAVL) framework. The orange branches represent the visual information and are parameterized by the weights θ_v. The green branches represent the acoustic information and are parameterized by the weights θ_a. The purple modules are the shared layers where both modalities flow through, which are parameterized by the weights θ_s. Finally, the gray module is the audio-visual prediction layer, which is parameterized by the weights θ_{av}.

III. Proposed Approach

Fig. 1 illustrates the architecture of the proposed approach, which consists of four major components. Each of these components is represented in the figure by the set of weights θ_v, θ_a, θ_s, and θ_{av}. The orange branches represent parts of the model where only visual information flows. This region is parameterized by the weights θ_v, and includes the visual conformer encoder layers that encode the visual input representations, the visual prediction layer for visual-only predictions, and the MLP reconstruction layers for the averaged visual input features. The green branches represent parts of the model where only acoustic information flows. This region is parameterized by the weights θ_a, and consists of the acoustic conformer encoder layers that encode the acoustic input representations, the acoustic prediction layer for acoustic-only predictions, and the MLP reconstruction layers for the averaged acoustic input features. The shared layers, depicted in purple in Fig. 1, are parameterized by the weights θ_s. This block processes both acoustic and visual inputs to learn intermediate audio-visual representations from both modalities. It also contains residual connections from the unimodal branches to the shared layers to preserve information from the unimodal representations. Lastly, the audio-visual prediction layer is parameterized by the weights θ_{av}, which focuses on processing audio-visual representations from the shared layers when the model receives paired audio-visual data for training or inference. The following sections describe these components in detail.

A. Acoustic and Visual Layers

As shown in Fig. 1, the acoustic and visual layers mirror each other. Both layers share the same basic structure and training mechanism. The main components of these layers are conformer encoders [54], which process all the sequential video or acoustic frames in parallel, depending on the modality available during either training or inference. To obtain inputs for our model, we utilize pre-trained feature extractors that produce a 1,408D feature vector for each visual frame and a 1,024D feature vector for each acoustic frame. Section IV-B describes these feature extractor modules in detail. We apply a 1D temporal convolutional layer at the frame level before entering the feature vectors to their corresponding conformer encoder layers. This step ensures that each element in the input sequences is aware of its neighboring elements. This approach also allows us to project both feature vectors into a 50D feature representation, matching their dimensions. The dimensionality of the projection was determined based on preliminary experiments. These experiments indicated that a 50D representation strikes an effective balance: it retains essential information from the original feature vectors while keeping the model's complexity manageable.

As seen in the top region of Fig. 1, there are two additional components separately implemented for each modality (acoustic and visual) that are implemented to (1) predict the emotional attributes, and (2) reconstruct the unimodal feature representations. For the prediction of the emotional attributes, the acoustic and visual layers contain a set of prediction layers, referred to as visual prediction and acoustic prediction in Fig. 1. They are responsible for generating unimodal predictions, which are utilized when our model operates in a unimodal setting. For the reconstruction of the unimodal feature representation, the model has a *multilayer perceptron* (MLP) head that serves as an auxiliary reconstruction task. The reconstruction task is used to have the model reconstruct the average-pooled input representations for the modality being used during training at that moment. The input of the reconstruction is the average-pooled representations obtained at the output of the shared layers (Section III-B). Equation 1 shows the total loss function of the model,

$$\mathcal{L} = \mathcal{L}_{pred}(y, pred) + \alpha \mathcal{L}_{MSE}\left(x_{pool}, Rec_M\right), \qquad (1)$$

where $\mathcal{L}_{pred}$ is the emotion task-specific loss (i.e., cross-entropy – (2), or *concordance correlation coefficient* (CCC) – (3)), y is the input label, $pred$ is the emotional prediction of the model, α is the scaling weight for the reconstruction loss, $\mathcal{L}_{MSE}$ is *mean squared error* (MSE) loss between the average-pooled input value x_{pool} and the reconstructed input Rec_M for modality M. The terms x_{pool} and Rec_M are specific to each modality; x_{pool} represents the pooled features of the input (audio or video), while Rec_M represents the reconstructed features for the corresponding modality. This differentiation ensures that the reconstruction loss appropriately guides the model to retain and reconstruct modality-specific information.

The reconstruction task is included in our model to promote the learning of more general features that can be applied to both modalities while preserving separate information for each modality, as the reconstruction from the shared layers requires the model to retain information from both modalities.

B. Shared Layers

The shared layers, depicted in purple in Fig. 1, mainly comprise a conformer encoder, following a similar structure to the acoustic and visual layers. These shared layers are modality-agnostic, meaning that during training and inference, the features from both modalities will pass through these layers whenever each modality is available. The purpose of this block is to ensure that the shared layers maintain information from both modalities incorporated in our model.

We also introduce a residual connection over the shared layers (gray arrow in model θ_s shown in Fig. 1). This residual connection from the unimodal branches over the shared layers ensures that the model retains unimodal separable information. This mechanism complements the reconstruction tasks mentioned in Section III-A to increase the robustness of the system when only one modality is available. The approach allows the gradients to flow directly from the shared layers to the unimodal branches, preserving modality-specific information in our model.

C. Audio-Visual Prediction Layer

When only one modality is available, the system will use the visual or acoustic prediction blocks. When both modalities are available for training or inference, we use the audio-visual prediction layer, highlighted in gray in Fig. 1. This layer plays a crucial role in making our approach versatile. Unlike other methods that rely on averaging predictions from different branches in their models, the audio-visual layers effectively utilize the contributions of acoustic and visual inputs in the audio-visual space. The audio-visual prediction layer consists of two fully connected layers that feed into a final audio-visual prediction head. We added this simple structure to ensure that representations from both modalities, obtained from the average-pooled output of each modality from the shared layers, are properly combined to obtain a final audio-visual prediction. During training, when audio-visual data is available, all other layers of the model are frozen after updating with acoustic and visual data. Only this layer is separately optimized to learn the required weights for audio-visual predictions. This layer ensures proper combination of audio-visual representations for robust predictions in both classification or regression settings.

D. Versatile Model Training

We train our proposed VAVL model using Algorithm 1. The training process can involve either a single modality or both modalities at any given iteration. If both modalities are available, the model first backpropagates the errors using the acoustic predictions to optimize the acoustic and shared weights (θ_a, θ_s). Then, it backpropagates the errors using the visual predictions to optimize the visual and shared weights (θ_v, θ_s). Next, it freezes all the parameters of the models other than the audio-visual prediction block (i.e., θ_v, θ_a, and θ_s are frozen). Then, it backpropagates the errors using the audio-visual predictions to optimize the audio-visual prediction weights (θ_{av}). This method facilitates the use of unpaired and paired audio-visual data for training. If only one modality is available, we only update either the visual and shared weights (θ_v, θ_s) or acoustic and shared weights (θ_a, θ_s). Essentially, most of the blocks in the proposed architecture are trained with either acoustic features or visual features. The only block that requires paired data is the audio-visual prediction layer. The shared conformer layer processes each modality separately when full audiovisual information is available, so the dimension of the input to this block has always consistent dimension. The outputs are then combined in the audio-visual prediction layer.

At inference time, the model utilizes the available information to make predictions, and in cases where both modalities are present, the model outputs the predictions from the audio-visual prediction layer. When only one modality is available during inference, the model outputs the prediction from the corresponding visual or acoustic prediction blocks.

Algorithm 1: - VAVL (Training and Inference).

Require: $\mathcal{D} = \{A, V, Label\}$, $\mathcal{M} = \{\theta_a, \theta_v, \theta_s, \theta_{av}\}$

Training$(\mathcal{D}, \mathcal{M}, \alpha)$

while $i < max\,train\,iterations$ **do**
  sample a batch of $\{A_i, V_i, Label_i\}$
  if $A_i \neq$ None **then**
    $Pred_a, A_{i-pool}, Rec_a = \mathcal{M}_{\{\theta_a, \theta_s\}}(A_i)$
    $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_a) + \alpha \mathcal{L}_{MSE}(A_{i-pool}, Rec_a)$
    backpropagate and update $\{\theta_a, \theta_s\}$
  if $V_i \neq$ None **then**
    $Pred_v, V_{i-pool}, Rec_v = \mathcal{M}_{\{\theta_v, \theta_s\}}(V_i)$
    $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_v) + \alpha \mathcal{L}_{MSE}(V_{i-pool}, Rec_v)$
    backpropagate and update $\{\theta_v, \theta_s\}$
  if $A_i \neq$ None **and** $V_i \neq$ None **then**
    $freeze\ \{\theta_a, \theta_v, \theta_s\}$
    $Pred_{av} = \mathcal{M}_{\{\theta_a, \theta_v, \theta_s, \theta_{av}\}}(A_i, V_i)$
    $\mathcal{L} = \mathcal{L}_{pred}(Label_i, Pred_{av})$
    backpropagate and update $\{\theta_{av}\}$
return $\mathcal{M}$

Inference $(\{A, V\}, \mathcal{M})$

if $A \neq$ None **and** $V \neq$ None **then**
  $Pred = \mathcal{M}_{\{\theta_a, \theta_v, \theta_s, \theta_{av}\}}(A, V)$
else if *one modality is not available* **then**
  $Pred = \mathcal{M}_{\{\theta_a, \theta_s\}}(A)$ *when* $V == None$
  $Pred = \mathcal{M}_{\{\theta_v, \theta_s\}}(V)$ *when* $A == None$
return $Pred$

E. Complexity Analysis

An analysis of the complexity of the VAVL method reveals a framework with 86 M trainable parameters, underscoring its capacity for handling complex multi-modal tasks. Central to the framework, we have the acoustic and visual conformers, each equipped with 34 M parameters, adeptly processing audio

and visual inputs. The shared conformer, integral to the model's functionality, contains 15 M parameters, facilitating seamless integration of different modalities. A crucial aspect of our model is the 1D temporal convolutional layers, containing 650 k parameters. These blocks are important in projecting both visual and acoustic features into a representation with the same dimension. Additionally, the model features three specialized prediction heads for acoustic, audio-visual, and visual processing, each with 390 K parameters, further enhancing its processing capabilities. Lastly, the reconstruction MLP layers contain a combined total of 1.4 million parameters. To evaluate the practical efficiency, we measured the inference latency and *real-time factor* (RTF) on a 3090 RTX GPU. The results are as follows: audio-only mode had an average latency of 0.003499 seconds (RTF: 0.000005), video-only mode had an average latency of 0.001759 seconds (RTF: 0.000003), and audiovisual mode had an average latency of 0.003770 seconds (RTF: 0.000006).

IV. Experimental Settings

A. Emotional Corpora

In this study, we use the CREMA-D [25], the MSP-IMPROV [26], and the CMU-MOSEI [22] corpora. The CREMA-D corpus is an audio-visual corpus with high-quality recordings from 91 ethically and racially diverse actors (48 male, 43 female). Actors were asked to convey specific emotions while reciting sentences. Videos were recorded against a green screen background, with two directors overseeing the data collection. One director worked with 51 actors, while the other worked with 40 actors. Emotional labels were assigned by at least seven annotators. In total, 7,442 clips were collected and rated by 2,443 raters. We use the perceived emotions from the audio-visual modality in the CREMA-D corpus for our classification task. We consider six emotional classes: anger, disgust, fear, happiness, sadness, and neutral state.

The MSP-IMPROV corpus [26] is the second audio-visual database used in this study. The corpus was collected to study emotion perception. The corpus required sentences with identical lexical content but conveying different emotions. Instead of actors reading sentences, a sophisticated protocol elicited spontaneous target sentence renditions. The corpus includes 20 target sentences with four emotional states, resulting in 80 scenarios and 652 speaking turns. It also contains the interactions that prompted the target sentence (4,381 spontaneous speaking turns), natural interactions during breaks between the recording of dyadic scenarios (2,785 natural speaking turns), and read recordings expressing the target emotional classes (620 read speaking turns). In total, the MSP-IMPROV corpus contains 7,818 non-read speaking turns and 620 read sentences. The corpus was annotated using a crowdsourcing protocol, monitoring the quality of the workers in real-time. Each sentence was annotated for arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) by five or more raters using a five-point Likert scale. We employ these three emotional attributes for our regression task.

The CMU-MOSEI [22] comprises review video clips of movies sourced from YouTube. Each clip is annotated by human experts with a sentiment score ranging from -3 to 3. We retrieved the data from the author's SDK and obtained a total of 22,859 files of which, based on their standard splits, 16,326 are used for training, 1,871 are used for development, and 4,659 are used for testing. This database consists of in-the-wild audio-visual recordings.

B. Acoustic and Visual Features

For the CREMA-D and MSP-IMPROV corpora, we have access to the raw videos and audio recordings, so we can extract audio and visual features. Our acoustic feature extractor is based on the "wav2vec2-large-robust" architecture [55], which has shown superior emotion recognition performance compared to other variants of the Wav2vec2.0 model [56], as demonstrated in the study by Wagner et al. [45]. The downstream head of our model consists of two fully connected layers with 1,024 nodes, layer normalization, and *rectified linear unit* (ReLU) activation function, followed by a linear output layer with three nodes for predicting emotional attribute scores (arousal, dominance, and valence). We import the pre-trained "wav2vec2-large-robust" model from the Hugging Face library [57]. We use this wav2vec2 model specifically pre-trained for emotion recognition tasks before its integration to ensure the representations used are optimal for emotion recognition. We aggregate the output of the transformer encoder using average pooling and feed them to the downstream head. To regularize the model and prevent overfitting, we utilize dropout with a rate of $p = 0.5$ applied to all hidden layers. To fine-tune the model, we use the training set of the MSP-Podcast corpus (release of v1.10) [58]. The ADAM optimizer [59] is employed with a learning rate set to 0.0001. We update the model with mini-batches of 32 utterances for 10 epochs. With the fine-tuned "wav2vec2-large-robust" model, we extract acoustic representations with a 25 ms window size and a 20 ms stride from the given audio, which are then used as the acoustic features. This strategy creates 50 frames per second.

To obtain visual features, we used the *multi-task cascaded convolutional neural network* (MTCNN) face detection algorithm [60] to extract faces from each image frame in the corpora using bounding boxes. Following the extraction of bounding boxes, we resize the images to a predetermined dimension of $224 \times 224 \times 3$. After face extraction, we utilize the pre-trained EfficientNet-B2 model [29] to extract emotional feature representations. At the time of our study, this model was among the top performers on the AffectNet corpus [37]. Similarly to the acoustic feature extractor, the EfficientNet-B2 model is pre-trained for emotion recognition tasks before its integration. This approach helps ensure the representations obtained from videos are optimal for emotion recognition. The representation obtained from the EfficientNet-B2 model is retrieved from the last fully connected layer before classification, with an array dimension of 1,408. This representation is then concatenated row-wise with all other frames within each clip from the datasets, serving as the input for the visual branch of our framework.

The CMU-MOSEI dataset offers pre-extracted features instead of raw data. Specifically, for the visual modality, it provides 35 facial action units, while the audio data includes features such

as *Mel-frequency cepstral coefficients* (MFCCs), pitch tracking, glottal source, and peak slope parameters, totaling 74 features. Consequently, we could not use the same set of features used to conduct experiments on the CREMA-D and MSP-IMPROV corpora (i.e., wav2vec2 and EfficientNet-B2 based features). Instead, we use the features provided with the release, which also facilitates the comparison with other methods using this corpus.

C. Implementation Details

We implemented the conformer blocks with an encoder hidden layer set to 512D, with 8 attention heads. We set a dropout rate to $p = 0.1$. The number of layers in the acoustic, visual, and shared conformer layers were set to three, three, and two, respectively. The acoustic/visual prediction layers and the reconstruction layers are implemented using a fully-connected structure with three layers. The first two layers are implemented with 512 nodes and 256 nodes, respectively. For the acoustic and visual prediction modules, the third layer is the output layer, where the size depends on the emotion recognition task (i.e., classification or regression). For the reconstruction task, the third layer is the target representation to be reconstructed, so it has 1,024 nodes for the acoustic features and 1,408 nodes for the visual features. The fully connected layers are implemented with dropout, with the rate set to $p = 0.2$. Similarly, the audio-visual prediction layers have a mostly identical fully-connected hidden layer structure to the unimodal prediction layers. The only difference is that the input layers are now 1,024D, as they need to take both unimodal representations from the shared layers as inputs. We trained the models for 20 epochs using the ADAM optimizer, with the ReLU as the activation function. The learning rate is set to 5e-5. We used a batch size of 32 for the CREMA-D and CMU-MOSEI dataset experiments, and a batch size of 16 for the MSP-IMPROV dataset experiments, since the MSP-IMPROV has longer sentences. The model was implemented in PyTorch and trained using an Nvidia Tesla V100.

The recordings in each database are divided into train, development, and test sets, with approximately 70%, 15%, and 15% of the data in each set, respectively. The splits were carried out in a speaker-independent manner, ensuring that no speaker appeared in more than one set. We trained each model five times, with different splits each time. All models were trained using the train set, and the best-performing model on the development set was selected and used to make predictions on the test set.

D. Cost Functions and Evaluation Metrics

Three of the most prominent tasks in affective computing are classification tasks for categorical emotions, regression models for emotional attributes, and sentiment analysis which can be approached as either a classification or a regression task, depending on the specifics of the problem and the nature of the sentiment scores. Our model is designed to be utilized and evaluated in both formulations (i.e., classification or regression tasks).

For emotion classification, our training objective is based on the multiclass cross-entropy loss, as seen in (2),

$$\mathcal{L}_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{2}$$

where M is number of classes, log is the natural log, c is the correct label for observation o, and p is the predicted probability that observation o belong to class c. We predict the emotional state/value for each input sequence in our test set and report the 'micro' and 'macro' F1-scores. The 'micro' F1-score is computed by considering the total number of true positives, false negatives, and false positives, making it sensitive to class imbalance. The 'macro' F1-score calculates the F1-score for each class separately and aggregated the scores with equal weight, so the performance on the minority class is as important as the performance in the majority class.

For emotional attributes regression models, we use the *concordance correlation coefficient* (CCC), which measures the agreement between the true and predicted emotional attribute scores. Equation 3 illustrates the CCC measurements,

$$\mathcal{L}_{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{3}$$

where μ_x and μ_y are the means of the true and predicted scores, σ_x and σ_y are the standard deviation of the true and predicted scores, and ρ is their Pearson's correlation coefficient. We train our model to maximize CCC so that the predicted scores have a high correlation with the true scores, while reducing their errors in the prediction. We use the CCC metric for model evaluation of the predictions of arousal, valence, and dominance.

We formulate the prediction of sentiment analysis as the prediction of a continuous number between -3 to 3. We employ the *mean absolute error* (MAE) as the primary training objective. This *L1 loss* function is defined as the average of the absolute differences between the predicted sentiment scores and the true sentiment scores. The L1 loss is formulated as shown in (4):

$$\mathcal{L}_{L1} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{4}$$

where N is the number of samples, y_i is the true sentiment score, and $\hat{y}_i$ is the predicted sentiment score for the i-th sample. The L1 loss is particularly suitable for sentiment analysis as it robustly handles outliers and ensures a linear error penalization. Our model aims to minimize this loss during training, thereby reducing the average magnitude of errors in sentiment score predictions. This loss function is instrumental in achieving high accuracy in predicting the nuanced sentiment scores across the specified range.

We conducted each experiment in this work five times with different partitions or seeds and reported the average metrics. Additionally, we performed statistical analyses to evaluate our model's performance against the baseline models. We used a two-tailed t-test, asserting a significance level at p-value<0.05.

E. Baselines

To evaluate the performance of our proposed model, we conducted experiments with three strong audio-visual frameworks and one unimodal framework. Baseline models 1, 2, 3, and 4 were implemented utilizing the code available in their respective repositories. For Baseline 5, the implementation was carried out in accordance with the specifications outlined in its associated

paper. Baseline 6 represents a unimodal variant of our model, specifically crafted for comparisons in a unimodal setting.

1) Baseline 1: UAVM: Gong et al. [19] proposed a unified audio-visual framework for classification, which independently processes audio and video features. The framework includes a shared transformer component and a classification layer, whose weights are shared between the modalities.

2) Baseline 2: AuxFormer: Goncalves and Busso [51] proposed an audio-visual transformer-based framework that creates cross-modal representations through transformer layers, sharing representations from query inputs from one modality to the keys and values of another modality. Additionally, their architecture includes unimodal auxiliary networks and modality dropout during training to enhance the robustness to missing modalities, allowing the model to be used in both audio-visual and unimodal settings.

3) Baseline 3: MulT: Tsai et al. [61] proposed a multimodal transformer architecture for human language time-series data. Their model uses a cross-modal transformer framework that generates pairs of bimodal representations, where the keys and values of one modality interact with the queries of a target modality. Vectors with similar target modalities are concatenated and passed to another transformer layer that generates representations used for prediction. The original model considered textual, visual, and acoustic features. We adapt the model from the original study by removing the dependencies on the textual branch, focusing only on visual and acoustic features.

4) Baseline 4: SFAV: Chumachenko et al. [62] present an architecture for audiovisual emotion recognition, particularly addressing the challenge of incomplete data from either modality during inference. Their model is designed to learn from both audio and visual data and incorporates robust fusion mechanisms that perform well even when one modality is absent. The authors explore fusion techniques, such as late transformer fusion and intermediate transformer fusion to more effectively integrate features from the audio and visual branches. In this paper, we refer to this baseline as *SFAV* in agreement with their methodology approach which uses self-attention fusion for audio-visual emotion recognition.

5) Baseline 5: TSLTM: Huang et al. [63] propose a multimodal transformer architecture for continuous emotion recognition, leveraging the Transformer's ability to model long-term temporal dependencies with self-attention mechanisms. Their model combines audio and visual modalities through model-level fusion without an encoder-decoder structure. It employs multi-head attention to learn emotional temporal dynamics and fuses audio-visual modalities into a shared semantic space, outperforming traditional fusion methods. Additionally, the architecture integrates *long short-term memory* (LSTM) networks to further improve performance. In our study, we refer to the model as *TLSTM* consistent with their method of combining the transformer model and LSTM for emotion recognition.

6) Baseline 6: Unimodal Acoustic and Visual Model: The unimodal baseline model has a similar structure to the model proposed in this study. Like the acoustic and visual layers of our model, it uses conformer encoder layers [54] to process all sequential video or acoustic frames in parallel. The output of the conformer layers is then average-pooled and fed into a network that contains two fully connected layers to generate the prediction. We build the unimodal network with five conformer layers to match the structure of the full VAVL network.

V. Experimental Results

A. Comparison With Baselines

This section compares our proposed model with the audiovisual and unimodal baselines explored in this study. All models were trained using the details discussed in Section IV-C. Tables I and II presents the average results for all the models across the five trials on each corpus. The models' performances are evaluated based on three modalities: audio-visual, acoustic, and visual. For the MSP-IMPROV dataset, the performance is assessed using the CCC predictions for *arousal* (Aro.), *valence* (Val.), and *dominance* (Dom.). For the CREMA-D dataset, we report the *F1-Macro* (F1-Ma) and *F1-Micro* (F1-Mi) scores. For the CMU-MOSEI dataset, we report *Mean Absolute Error* (MAE). Upon examining the performance metrics presented in the tables, the VAVL model demonstrates notable superiority across various tasks when compared to the baseline models on CREMA-D, MSP-IMPROV, and CMU-MOSEI datasets. This superiority is quantified by the asterisks indicating statistical significance.

In the audio-visual modality on the CREMA-D dataset, the VAVL model achieves an F1-Macro score of 0.779 ± 0.025, which is significantly higher than that of the strongest baseline, the TSLTM model, at 0.667 ± 0.012. This pattern is consistent in the F1-Micro score, where VAVL scores 0.826 ± 0.015, outperforming the second-best SFAV model's score of 0.810 ± 0.010.

The VAVL model's performance in acoustic modality is also strong, with a MAE of 0.829 ± 0.034 on the CMU-MOSEI dataset, which is lower than the best-performing baseline model (AuxFormer) with an MAE of 0.830 ± 0.032. Lower MAE indicates better performance, and these numbers underscore the predictive accuracy of the VAVL model in capturing sentiment changes. On the visual modality of the CMU-MOSEI dataset, The MAE of the VAVL model is 0.795 ± 0.028, which again is the lowest error rate compared to the baselines. The closest competitor is the MulT model with an MAE of 0.815 ± 0.028, indicating that VAVL is more precise in interpreting visual data for sentiment analysis.

The MSP-IMPROV dataset further showcases the VAVL's robustness, particularly in the audio-visual category for arousal (Aro.), with a score of 0.856 ± 0.110, significantly surpassing the UAVM model which has a score of 0.471 ± 0.257. Similarly, in dominance (Dom.), VAVL scores 0.814 ± 0.095, while the next best is the TSLTM model at 0.782 ± 0.145. The results for the audio-visual setting on the MSP-IMPROV corpus reinforce our hypothesis that the use of averaging unimodal predictions, as employed by some baselines, might not work well for regression tasks. These metrics highlight the strong capability of the VAVL framework in detecting the intensity and control aspects of emotions conveyed through both audio and visual cues.

In summary, the VAVL model not only consistently outperforms the baseline models in terms of CCC and macro and

TABLE I
COMPARISON BETWEEN THE VAVL MODEL AND THE AUDIO-VISUAL AND UNIMODAL BASELINES

	CREMA-D						CMU-MOSEI		
	Audio-Visual		Acoustic		Visual		Audio-Visual	Acoustic	Visual
Model	F1-Ma.	F1-Mi.	F1-Ma.	F1-Mi.	F1-Ma.	F1-Mi.	MAE	MAE	MAE
VAVL	**0.779±.025**	**0.826∗±.015**	0.628±.013	**0.701∗±.015**	**0.738∗±.033**	**0.787∗±.020**	**0.792±.029**	0.829±.034	**0.795∗±.028**
UAVM	0.737±.018	0.804±.008	0.552±.034	0.607±.033	0.691±.017	0.748±.013	0.802±.029	0.862±.031	0.803±.026
AuxFormer	0.743±.019	0.791±.017	0.504±.084	0.578±.046	0.692±.033	0.739±.032	0.808±.030	0.830±.032	0.806±.028
SFAV	0.771±.020	0.810±.010	**0.658±.015**	0.699±.014	0.656±.019	0.709±.015	0.804±.024	0.821±.032	0.813±.032
TLSTM	0.667±.012	0.746±.012	0.503±.046	0.543±.042	0.574±.020	0.651±.025	0.811±.022	0.821±.032	0.815±.028
MulT	0.714±.015	0.762±.015	—	—	—	—	0.810±.028	—	—
Uni. (A)	—	—	0.625±.024	0.690±.015	—	—	—	**0.821±.028**	—
Uni. (V)	—	—	—	—	0.725±.039	0.783±.028	—	—	0.859±.032

The table reports the average performance metrics across five trials. The symbol * indicates that the VAVL model is significantly better than the other baselines on the CREMA-D and CMU-MOSEI datasets.

TABLE II
COMPARISON BETWEEN THE VAVL MODEL AND THE AUDIO-VISUAL AND UNIMODAL BASELINES. THE TABLE REPORTS THE AVERAGE PERFORMANCE METRICS ACROSS FIVE TRIALS

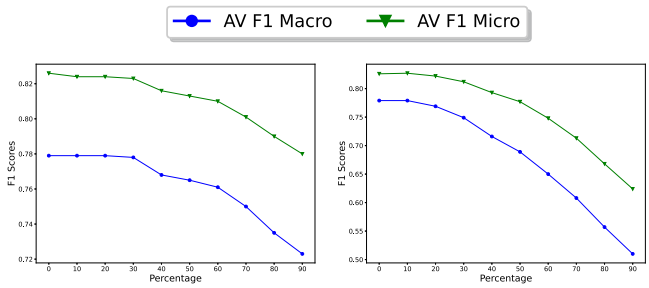
	MSP-IMPROV								
	Audio-Visual			Acoustic			Visual		
Model	Aro.	Val.	Dom.	Aro.	Val.	Dom.	Aro.	Val.	Dom.
VAVL	**0.856∗±.110**	**0.876∗±.095**	**0.814∗±.151**	**0.853∗±.104**	0.858±.111	0.783±.155	**0.422∗±.133**	**0.631∗±.032**	0.375±.112
UAVM	0.471±.257	0.687±.329	0.544±.309	0.578±.316	0.705±.339	0.637±.384	0.274±.145	0.522±.260	0.296±.163
AuxFormer	0.672±.134	0.820±.071	0.652±.149	0.722±.218	0.789±.142	0.730±.162	0.363±.153	0.581±.071	0.293±.141
SFAV	0.786±.103	0.761±.119	0.721±.147	0.745±.081	0.699±.073	0.628±.140	0.345±.192	0.572±.175	**0.473±.148**
TLSTM	0.832±.095	0.843±.103	0.767±.155	0.835±.117	0.765±.188	**0.841±.125**	0.172±.140	0.574±.094	0.201±.066
MulT	0.775±.088	0.761±.056	0.778±.136	—	—	—	—	—	—
Uni. (A)	—	—	—	0.841±.095	**0.878±.108**	0.820±.158	—	—	—
Uni. (V)	—	—	—	—	—	—	0.383±.065	0.598±.093	0.321±.070

The table reports the average performance metrics across five trials. The symbol * indicates that the VAVL model is significantly better than the other baselines on the MSP-IMPROV dataset.

micro F1 scores, it also maintains lower MAE across datasets, showcasing its superior ability to accurately recognize and predict emotions. The numerical superiority of the VAVL model across various datasets which have been collected under diverse scenarios emphasizes its potential for practical applications in emotion recognition systems.

B. Random Masking of Features

This section evaluates the system's performance in scenarios with absent features, simulating missing data by randomly zeroing out either visual or acoustic data at the frame level. To analyze the system's resilience to incomplete information, we incrementally mask available frames from 0% to 90% in 10% increments. For instance, in the 30% condition, 30% of the frames from the modality being masked are replaced with zeros. This evaluation helps us understand the model's performance variations when audio-visual data is available but partially incomplete. The rationale behind our approach of simulating missing modalities by randomly dropping frames at varying percentages is to closely mimic conditions where input data might be incomplete or degraded due to practical issues. In real-world scenarios, data loss can occur randomly due to face occlusions, hardware intermittent malfunctions, or other environmental factors. Randomly dropping frames aims to replicate these unpredictable disruptions. We focus on the MSP-IMPROV and CREMA-D corpora, which allow us to use our entire framework–from processing raw inputs to predicting emotions - a process not possible with the CMU-MOSEI corpus due to the absence of raw data. The MSP-IMPROV and CREMA-D corpora provide consistency in frame-wise facial feature extraction and synchronize acoustic feature extraction timing, employing a 25 ms window size and a 20 ms stride for both corpora.



(a) Audio Masking (b) Video Masking

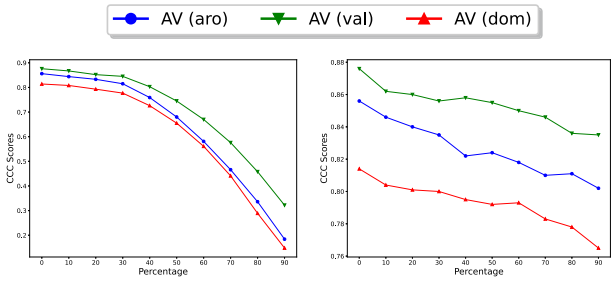
Fig. 2. Performance of the proposed model under audio-visual settings on the CREMA-D corpus with partial visual or acoustic information. The figure plots the micro F1-scores as a function of the percentage of the frames included for the masked modality (the other modality is assumed to be complete).

Fig. 2(a) and (b) show the average results on the CREMA-D corpus for masking audio and visual inputs, respectively, obtained from the model's audio-visual prediction heads. The results indicate that removing either modality impacts the model. With random audio masking, performance remains stable when we mask up to 30% of the frames. The model reaches an F1 Macro score of approximately 0.72 when only 10% of the

TABLE III
ABLATION ANALYSIS OF THE PROPOSED VAVL MODEL USING THE MSP-IMPROV AND CREMA-D CORPORA

	MSP-IMPROV									CREMA-D					
	Audio-Visual			Acoustic			Visual			Audio-Visual		Acoustic		Visual	
Model	Aro.	Val.	Dom.	Aro.	Val.	Dom.	Aro.	Val.	Dom.	F1-Ma	F1-Mi	F1-Ma	F1-Mi	F1-Ma	F1-Mi
VAVL	**0.856**	**0.876**	**0.814**	**0.853**	0.858	**0.783**	**0.422**	**0.631**	**0.375**	**0.779**	**0.826**	**0.628**	**0.701**	**0.738**	**0.787**
Ablt. 1	0.716	0.779	0.705	0.682	0.753	0.712	0.176	0.341	0.133	0.751	0.807	0.397	0.528	0.727	0.783
Ablt. 2	0.810	0.873	0.788	0.791	0.864	0.780	0.172	0.577	0.227	0.761	0.816	0.604	0.684	0.718	0.775
Ablt. 3	0.711	0.843	0.659	0.784	**0.870**	0.776	0.374	0.617	0.265	0.762	0.814	0.629	0.691	0.713	0.764

Ablation 1 removes the residual connections over the shared layers. Ablation 2 removes reconstruction step from the framework. Ablation 3 removes audio-visual prediction layer, estimating the output by averaging the unimodal predictions.



(a) Audio Masking (b) Video Masking

Fig. 3. Performance of the proposed model under audio-visual settings on the MSP-IMPROV corpus with partial visual or acoustic information. The figure plots the CCC scores as a function of the percentage of the frames included for the masked modality (the other modality is assumed to be complete).

audio frames are present. In contrast, visual masking results in a sharper drop in performance when we mask 30% of the frames. The approach has a low F1 Macro score of about 0.51 when only 10% of the video frames are present. This score is significantly lower than those in Table I using only the acoustic head, implying that the audio-visual head may get confused with zero-masked frames. When the percentage of missing visual frames is higher than 60-70%, it is better to rely solely on the acoustic modality.

Fig. 3(a) and (b) depict average results on the MSP-IMPROV corpus for masking audio and visual inputs. The results from the audio-visual prediction heads show that the performance with random audio masking is steady up to 30%, then drops sharply, reaching scores of about 0.4 CCC for valence, and 0.2 CCC for arousal and dominance. This result suggests that the emotional attribute models rely more on the acoustic modality to achieve high performance. Masking the audio frames can lead to lower performances than using only the visual head for prediction. When we miss more than 60% of the acoustic features, it is better to focus solely on visual features, using the results from the visual head. Conversely, masking visual features results in a smaller performance drop. Even when we mask 90% of the visual features, we still achieve scores of 0.85 CCC for valence, 0.82 CCC for arousal, and 0.75 CCC for dominance. These findings confirm that the model is more dependent on acoustic features than visual features to predict emotional attributes.

C. Ablation Analysis of the VAVL Framework

Table III presents the results of an ablation analysis of the VAVL model on the MSP-IMPROV and CREMA-D corpora, comparing the performance of the full model with its ablated versions. In these experiments, we focus on the on the MSP-IMPROV and CREMA-D corpora, since with these corpora we were able to use our entire framework from raw inputs to emotion predictions, which was not possible with CMU-MOSEI since raw datapoints are not provided. Three ablated models are considered: Ablation 1 (Ablt. 1) removes the residual connections over the shared layers, Ablation 2 (Ablt. 2) removes the reconstruction step from the framework, and Ablation 3 (Ablt. 3) removes the audio-visual prediction layer, using the average of the unimodal predictions as the audio-visual prediction.

In the MSP-IMPROV dataset, the full VAVL model outperforms all the ablated versions in terms of arousal, valence, and dominance for the audio-visual and visual modalities. In the acoustic modality, VAVL shows the best performance for arousal and valence. However, Ablation 2 and 3 slightly surpass the VAVL results for dominance. These results indicate that the residual connections, reconstruction step, and audio-visual prediction layer all contribute to the strong performance of the VAVL model.

On the CREMA-D dataset, VAVL consistently achieves the highest F1-Macro and F1-Micro scores across multimodal and unimodal settings, indicating that the full model is superior to its ablated versions. Ablation 1 has the lowest performance in the acoustic modality, while Ablations 2 and 3 show competitive results in some cases, albeit not surpassing the results of the full VAVL model. These results are consistent with the MSP-IMPROV ablation results and suggest that the combination of all components in the VAVL model leads to the best performance, with each component playing a significant role in the overall success of the model.

D. Shared Embedding Analysis

In this section, we perform analysis on the output embedding representations for each modality at the output of the shared layers from our VAVL model and the baselines. This experiment is conducted to investigate whether the shared layers of a multimodal model are capable of producing different representations for each modality. Using shared layers in a multimodal model allows the architecture to learn a common representation across different modalities, which can help in tasks that require integrating information from multiple sources. However, maintaining distinct representations for each modality in a multimodal model is crucial for several reasons. First, it allows the model to capture modality-specific information, which is essential for optimal performance. Second, it enhances interpretability by making it easier to analyze each modality's contribution to the final output

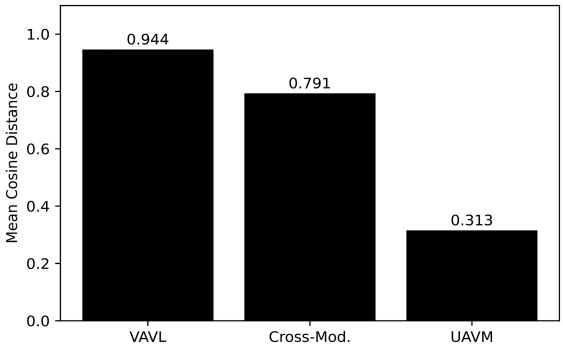


Fig. 4. Comparison of average cosine distances between acoustic and visual representations generated by shared layers of the VAVL, AuxFormer, and UAVM models.

or decision. Lastly, distinct representations ensure the model's adaptability, enabling it to perform well even when some modalities are missing or incomplete. To verify our proposed model's capability of generating distinct separable representations for each modality from the shared layers, we obtain separate output embeddings from the shared layers of the trained model using either acoustic or visual information. Then, we calculate the cosine distance between their respective embeddings. On the one hand, a high cosine distance indicates that the embeddings are more dissimilar, and thus capturing modality-specific information. On the other hand, a low cosine distance implies that the embeddings are more similar, potentially indicating that the model is not effectively capturing the unique features of each modality, and thus not effectively utilizing the multimodal input.

Fig. 4 presents a comparison of the average cosine distances between acoustic and visual representations generated by the shared layers of our proposed framework (VAVL). For comparison, we use the shared models from the AuxFormer architecture [51], which are implemented with cross-modal layers (the same type of layers used in the MulT model [61]), and from the UAVM model [19]. Based on the average cosine distance values obtained, the VAVL shared layers is the most effective model in generating distinct representations for the acoustic and visual modalities. The cross-modal layers show moderate effectiveness, while the UAVM shared layers seem to struggle in separating the modalities. These results correlate with our model's superior overall performance in unimodal settings compared to the baselines.

VI. Conclusion

This study introduced the VAVL model, a novel versatile framework for audio-visual, audio-only, and video-only emotion recognition. The proposed approach is built with shared layers, residual connections over shared layers, and a unimodal reconstruction task. It attains high emotion recognition performance and can be trained even when paired audio and visual data is unavailable for part of the recordings. Furthermore, the model allows for direct application across both emotion regression and classification tasks. In the MSP-IMPROV corpus, the VAVL model consistently outperforms the CCC predictions of all multimodal baselines across all emotional attributes, achieving 0.856, 0.876, and 0.814 for arousal, valence, and dominance in audio-visual settings. Our results demonstrate that our method is able to better leverage visual representations to improve both unimodal and audio-visual performance, particularly in the visual-only setting. In the CREMA-D corpus, VAVL consistently surpasses baselines across all settings, achieving the highest F1-Macro (0.779) and F1-Micro (0.826) scores in the audio-visual modality. Under more naturalistic environment with the CMU-MOSEI corpus, we achieve better audio-visual and visual MAE scores than the strong baselines explored for sentiment score prediction. In addition to high performance on emotion recognition tasks, we show through the experimental evaluation our models' capabilities on handling missing modalities at the audio-visual setting and we also ran experiments to show that the shared layers of the proposed architecture are able to better capture the distinct features contained in acoustic and visual inputs, when compared to the shared layers of the baseline models.

While our training pipeline offers flexibility by accommodating both single and dual modalities, we acknowledge potential limitations. Specifically, training can become unstable when dealing with extremely highly imbalanced data from the two modalities. To mitigate this problem, we employ careful batch management and learning rate adjustments to ensure stability. Additionally, the iterative freezing and unfreezing of weights introduces some complexity to the training process, but these steps are crucial for effective learning from both unpaired and paired data.

Future work for this study is to utilize this framework on other audio-visual tasks such as sound recognition [64], scene classification [65], event localization [66], and audio-visual speech recognition [67]. Furthermore, extension to this work can include considering additional modalities, such as *electroencephalogram* (EEG) [68]. Integrating EEG data with facial expressions could provide complementary information, enriching multi-modal emotion analysis and opening new research avenues. This approach is significantly less complex compared to other methods, such as cross-modal learning [51], [61]. In cross-modal learning, bimodal context representation vectors are developed, and the introduction of each additional modality requires the addition of $M * (M - 1)$ context layers to the model, where M represents the number of modalities. We expect that our approach will be easier to computationally scale to incorporate new modalities.

References

[1] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.
[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.
[3] S. Marioryad and C. Busso, "Facial expression recognition in the presence of speech using blind lexical compensation," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 346–359, Fourth Quarter 2016.
[4] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *Proc. 7th Int. Seminar Speech Prod.*, Ubatuba-SP, Brazil, 2006, pp. 549–556.
[5] C. Busso et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, State College, PA: ACM Press, 2004, pp. 205–211.

[6] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, Apr. 2015.

[7] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-supervised learning with cross-modal transformers for emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Shenzhen, China, 2021, pp. 381–388.

[8] Y. L. Bouali, O. B. Ahmed, and S. Mazouzi, "Cross-modal learning for audio-visual emotion recognition in acted speech," in *Int. Conf. Adv. Technol. Signal Image Process.*, Sfax, Tunisia, 2022, pp. 1–6.

[9] M. Tran and M. Soleymani, "A pre-trained audio-visual transformer for emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Singapore, 2022, pp. 4698–4702.

[10] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Automat. Comput.*, vol. 18, pp. 351–376, Jun. 2021.

[11] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. Int. Conf. Multimodal Interaction*, Utrecht, The Netherlands, 2020, pp. 400–404.

[12] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, Fouth Quarter 2022.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 689–696.

[14] R. Arandjelović and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 609–617.

[15] P. Antoniadis, I. Pikoulis, P. Filntisis, and P. Maragos, "An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 3638–3644.

[16] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, and P. Xiao, "Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features," *Neurocomputing*, vol. 391, pp. 42–51, May 2020.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[18] H. Akbari et al., "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Proc. Conf. Neural Inf. Process. Syst.*, 2021, pp. 24206–24221.

[19] Y. Gong, A. H. Liu, A. Rouditchenko, and J. Glass, "UAVM: Towards unifying audio and visual models," *IEEE Signal Process. Lett.*, vol. 29, pp. 2437–2441, 2022.

[20] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[21] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[22] A. Zadeh, P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, pp. 5642–5649.

[23] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," in *Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 1955–1960.

[24] P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," Sep. 2022, *arXiv:2209.03430*.

[25] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Fourth Quarter 2014.

[26] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, First Quarter 2017.

[27] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 4995–4999.

[28] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 951–955.

[29] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Fourth Quarter 2022.

[30] W.-C. Lin, L. Goncalves, and C. Busso, "Enhancing resilience to missing data in audio-text emotion recognition with multi-scale chunk regularization," in *Proc. ACM Int. Conf. Multimodal Interaction*, Paris, France, 2023, pp. 207–215.

[31] J. Li et al., "Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers," Mar. 2023, *arXiv:2303.09164*.

[32] J. Yu et al., "A dual branch network for emotional reaction intensity estimation," Mar. 2023, *arXiv:2303.09210*.

[33] L. Goncalves and C. Busso, "Learning cross-modal audiovisual representations with ladder networks for emotion recognition," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes island, Greece, 2023, pp. 1–5.

[34] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Social Psychol.*, vol. 17, no. 2, pp. 124–129, Mar. 1971.

[35] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Proc. Interspeech*, Incheon, South Korea, 2022, pp. 1168–1172.

[36] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE MultiMedia*, vol. 27, no. 1, pp. 37–48, First Quarter 2020.

[37] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, First Quarter 2019.

[38] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Third Quarter 2012.

[39] M. Bradley and P. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[40] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *J. Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[41] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, Jun. 2021.

[42] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 4482–4486.

[43] J.-H. Hsu and C. H. Wu, "Applying segment-level attention on Bi-modal transformer encoder for audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3231–3243, Fourth Quarter 2023.

[44] A. Wasi, K. Šerbetar, R. Islam, T. Rafi, and D.-K. Chae, "ARBEx: Attentive feature extraction with reliability balancing for robust facial expression learning," May 2023, *arXiv:2305.01486*.

[45] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," Mar. 2022, *arXiv:2203.07378*.

[46] S. Upadhyay et al., "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, Cambridge, MA, USA, 2023, pp. 1–8.

[47] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4651–4664.

[48] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[49] Y. Dai et al., "One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code," May 2022, *arXiv:2205.06126*.

[50] N. Shvetsova et al., "Everything at once-multi-modal fusion transformer for video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 19988–19997.

[51] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Singapore, 2022, pp. 7357–7361.

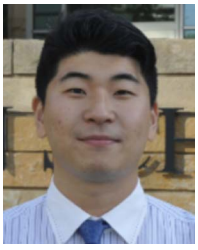
[52] X. Guo, B. Zhu, L. Polanía, C. Boncelet, and K. Barner, "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions," in *Proc. ACM Int. Conf. Multimodal Interaction*, Boulder, CO, USA, 2018, pp. 635–639.

[53] A. Khan, Z. Li, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong, "Group-level emotion recognition using deep models with a four-stream hybrid network," in *Proc. ACM Int. Conf. Multimodal Interact.*, Boulder, CO, USA, 2018, pp. 623–629.

[54] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 5036–5040.

[55] W.-N. Hsu et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," Apr. 2021, *arXiv:2104.01027*.
[56] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.
[57] T. Wolf et al., "HuggingFace's transformers: State-of-the-art natural language processing," Oct. 2019, *arXiv: 1910.03771v5*.
[58] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.
[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–13.
[60] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[61] Y.-H. Tsai, S. Bai, P. Liang, J. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 6558–6569.
[62] K. Chumachenko, A. Iosifidis, and M. Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Montreal, QC, Canada, 2022, pp. 2822–2828.
[63] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 3507–3511.
[64] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 721–725.
[65] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Toronto, ON, Canada, 2021, pp. 626–630.
[66] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany:Springer, 2018, pp. 252–268.
[67] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-AVSR: Audio-visual speech recognition with automatic labels," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.
[68] X. Huang et al., "Multi-modal emotion analysis from facial expressions and electroencephalogram," *Comput. Vis. Image Understanding*, vol. 147, pp. 114–124, Jun. 2016.

Lucas Goncalves (Student Member, IEEE) received the BS degree in electrical engineering from the University of Wisconsin - Platteville, in 2018. He is currently working toward the PhD degree in electrical engineering with the University of Texas at Dallas. At UTD, he is a research assistant with the Multimodal Signal Processing (MSP) laboratory. In 2022 and 2023, he was awarded the Excellence in Education Doctoral Fellowship from the Erik Jonsson School of Engineering and Computer Science. His research interests include areas related to affective computing, deep learning, and multimodal signal processing. He is also a student member of the *IEEE Signal Processing Society* (SPS) and *International Speech Communication Association* (ISCA).



Seong-Gyun Leem (Student Member, IEEE) received the BS and MS degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2018 and 2020, respectively. He is currently working toward the PhD degree in electrical engineering with the University of Texas at Dallas. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.

Wei-Cheng Lin (Student Member, IEEE) received the BS degree in communication engineering from the National Taiwan Ocean University (NTOU), Taiwan, in 2014, the MS degree in electrical engineering from the National Tsing Hua University (NTHU), Taiwan, in 2016, and the PhD degree in electrical engineering from the University of Texas at Dallas (UTD), in 2023. His research interests include human-centered behavioral signal processing (BSP), deep learning, and multimodal/speech signal processing. He is also a student member of the IEEE Signal Processing Society (SPS) and International Speech Communication Association (ISCA).

Berrak Sisman (Member, IEEE) received the PhD degree in electrical and computer engineering from National University of Singapore, in 2020, fully funded by Singapore International Graduate Award (SINGA). She is an assistant professor with the Electrical and Computer Engineering Department, University of Texas at Dallas. She is leading the Speech and Machine Learning (SML) laboratory which is part of the Center for Robust Speech Systems (CRSS). Prior to joining UT Dallas, she was a tenure-track faculty member with the Singapore University of Technology and Design (2020-2022). She was a postdoctoral research fellow with the National University of Singapore (2019-2020). She was an exchange PhD student with the University of Edinburgh and a visiting scholar with the Centre for Speech Technology Research (CSTR), University of Edinburgh (2019). She was also a visiting researcher with RIKEN Advanced Intelligence Project in Japan (2018). Her research is focused on machine learning, deep learning, emotion, speech synthesis and voice conversion. She has served as the General Coordinator of the Student Advisory Committee (SAC) of the International Speech Communication Association (ISCA) and is currently serving as the General Coordinator of the ISCA Postdoc Advisory Committee (PECRAC). She served as the Area Chair at INTERSPEECH 2021, INTERSPEECH 2022, IEEE SLT 2022 and IEEE ICASSP 2023. She also served as the Publication Chair of ICASSP 2022. She is elected as a member of the IEEE Speech and Language Processing Technical Committee (SLTC) in the area of Speech Synthesis for the term from 2022 to 2024.

Carlos Busso (Fellow, IEEE) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an incoming Professor of the Language Technologies Institute at Carnegie Mellon University (January 2025). He is currently a Professor at the University of Texas at Dallas's Electrical and Computer Engineering Department, where he is also the director of the Multimodal Signal Processing (MSP) Laboratory [http://msp.utdallas.edu]. His research interests include human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) with the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently an associate editor of the *IEEE Transactions on Affective Computing*. He is a member of AAAC and a senior member of ACM. He is an an ISCA fellow.