The Role of Facial and Speech Features in Emotion Classification

Loïc Houmard[1], **Ard Kastrati**[1], **Dushan Vasilevski**[2], **Roger Wattenhofer**[1]

[1]ETH Zurich [2]Magic Leap
[1]{lhoumard, akastrati, wattenhofer}@ethz.ch, [2]dvasilevski@magicleap.com

Abstract

In this study, we investigate the role of facial and speech features in improving the accuracy of emotion classification methods. First, we evaluate the performance of state-of-the-art models in speech and facial emotion classification, ranging from computationally intensive ones to those employing efficient feature extraction techniques. We also evaluate the utility of partial facial features (eyes and mouth) when a complete video feed is not feasible. Our empirical results reveal that incorporating both modalities consistently improves emotion classification accuracy, surpassing single-modality benchmarks. In particular, for efficient models, we observe the most significant improvement, with nearly a 10% increase in multi-modal accuracy. This insight is especially useful for devices with limited video feed and computing power, like head-mounted displays (HMDs). Interestingly, our findings indicate that speech is more adept at classifying certain emotions, whereas facial features excel in others. This also highlights the intrinsic advantages of a multi-modal approach, offering a more comprehensive understanding of emotion classification.

Introduction

Emotion classification is the task of identifying discrete emotional states such as anger, surprise, fear, sadness, etc. It has gained interdisciplinary interest, drawing contributions from fields such as psychology (Mandler 1997), medicine (Nyquist and Luebbe 2020; Greco et al. 2021), and computer science (Ewart J. de Visser and Shaw 2018). It offers transformative capabilities in sectors like social media (Andalibi and Buss 2020), automobile safety (Zepf et al. 2020), and human-computer interaction (Cowie et al. 2001). Particularly, in the expanding domains of Augmented Reality (AR) and Virtual Reality (VR), accurate emotion classification is important for creating immersive and interactive experiences through head-mounted displays (HMDs) (Marín-Morales et al. 2020).

There has been substantial recent work aimed at improving the accuracy of classifying the emotional state, with approaches generally falling into three main categories: the use of advanced deep learning models (Li and Deng 2018; Rajamani et al. 2021a; Minaee, Minaei, and Abdolrashidi 2021; Rajamani et al. 2021b), the application of multi-view learning techniques (Tompkins et al. 2023), and the incorporation of multi-modal data (Luna Jiménez et al. 2021). The first approach capitalizes on advancements in deep learning to construct larger or more sophisticated models (Pepino, Riera, and Ferrer 2021; Rizos et al. 2020). Utilizing pre-trained models like Wav2Vec2 (Baevski et al. 2020) for speech and ResNet (He et al. 2016) for images has been shown to significantly improve emotion classification accuracy. The second approach, multi-view learning, trains models to predict both emotional categories and continuous attributes like valence and arousal (Cowen and Keltner 2017). This provides a richer training signal and enhances the overall reliability of emotion classification (Tompkins et al. 2023). The third approach combines multi-modal data sources, to offer a more nuanced understanding of emotional states, thereby potentially increasing performance (Luna Jiménez et al. 2021). Various features can be employed for emotion recognition, including speech (Schuller 2018; Wani et al. 2021; Akçay and Oğuz 2020), facial expressions (Canal et al. 2022), EEG signals (Li et al. 2022), and text (Chowanda et al. 2021).

In this study, our focus is centered on emotion classification through speech and facial features. We aim to offer a thorough overview of the interaction between modalities and the impact of various feature extraction methods on emotion classification accuracy. This includes examining efficient techniques as well as pre-trained large models, illuminating the synergies and trade-offs when used for emotion classification with limited facial visibility and computational resources, as is typical with Head-Mounted Displays (HMDs). Our study provides multiple contributions:

1. We reproduce and extend the results of existing research (Luna Jiménez et al. 2021), thereby establishing a robust benchmark for each modality within the current state of the art.
2. Inspired by the TIMNet architecture (Ye et al. 2022), known for enhancing speech emotion recognition, we run experiments with multi-modal setup by integrating facial features directly into a unified model, moving away from conventional ensemble-based approaches (Luna Jiménez et al. 2021).
3. We explore feature extraction methods balancing efficiency and accuracy, ranging from accurate but

computationally intensive pre-trained models, like Wav2Vec2 (Baevski et al. 2020) and ResNet (He et al. 2016), to efficient methods, like key-point distances for eyes and Log Mel Spectrogram for speech.

4. Our work also delves deeper into facial features, segmenting them into distinct categories — specifically, features around the eyes and the mouth — to provide a more detailed analysis when a complete video feed is not feasible (e.g. in HMDs).
5. Performance assessments on the RAVDESS dataset (Livingstone and Russo 2018) show that the multi-modal model consistently outperforms single-modality approaches. The relative improvement is always present but substantially larger when efficient feature extraction methods are used.
6. Our analysis reveals that certain modalities are particularly adept at decoding specific emotional states, emphasizing the advantages of a multi-modal approach for the classification of emotions. For instance, while emotional states like happiness and sadness are more accurately classified through facial features, surprise and anger is better identified through speech data.

Related Work

Speech Emotion Recognition The study of paralanguage, which includes non-lexical elements of the voice like tone and pitch, underpins the idea that emotional states can be intentionally or subconsciously conveyed through speech. Schuller and Batliner provides a comprehensive framework that forms the basis for emotion recognition in speech (Schuller and Batliner 2013). Traditionally, feature engineering has been the cornerstone of speech emotion recognition. For instance, Ancilin and Milton reports an accuracy of 64.31% by using MFCCs (Ancilin and Milton 2021), and Bhavan et al. reaches an overall accuracy of 72.91% by passing MFCCs and spectral centroids to a bagged ensemble of support vector machines (Bhavan et al. 2019). With the advent of deep learning, the field has seen a shift towards models capable of operating on raw audio data or pre-trained features. Singh et al. reaches an accuracy of 81.2% on RAVDESS using a hierarchical DNN classifier (Singh et al. 2021), while Pepino, Riera, and Ferrer combines eGeMAPS features with embeddings from Wav2Vec2.0, achieving an accuracy of 77.5% (Pepino, Riera, and Ferrer 2021). A key challenge is the lack of standardized metrics, especially for datasets like RAVDESS (Livingstone and Russo 2018). Recent work has highlighted the need for a uniform evaluation framework for effective model comparison (Luna Jiménez et al. 2021).

Facial Emotion Recognition Facial emotion recognition leverages multiple techniques to discern emotional states from facial expressions. Tools like the dlib library use facial landmarks to capture important cues related to emotional expressions (King 2009). These landmarks serve as a foundation for a more descriptive system known as the Facial Action Coding System (FACS), which further categorizes facial movements into Action Units (AUs) (Ekman and Friesen 1978). Sanchez-Mendoza, Masip, and Lapedriza utilizes 12 AUs and achieves a 90% recognition rate on the Cohn–Kanade (CK) database using decision trees (Sanchez-Mendoza, Masip, and Lapedriza 2014). Yao et al. employs a two-stage approach that involves predicting AUs and then using an SVM to recognize emotions, reaching an average accuracy of 92% on the CK database (Yao et al. 2021). On the deep learning front, Minaee, Minaei, and Abdolrashidi proposes the Deep-Emotion model, which employs an architecture with an attention mechanism (Minaee, Minaei, and Abdolrashidi 2021).

Multi-modal Emotion Recognition In the field of multi-modal emotion recognition, Deng and Leung utilizes an early fusion approach by combining features from the T5 transformer and audio model TRILL trained with an unsupervised triplet loss. Contrastingly, Sun et al. used late fusion by independently training bi-LSTM models with attention layers for audio, video, and text modalities and fusing the posteriors to predict arousal and valence (Sun et al. 2021). Further supporting late fusion's effectiveness, Luna Jiménez et al. applies ensemble methods to integrate the posteriors from aural and visual models, achieving an 86.7% accuracy rate on the RAVDESS dataset (Luna Jiménez et al. 2021).

Methods

Dataset We utilize the RAVDESS dataset's speech set (Livingstone and Russo 2018), comprised of 24 actors (12 female, 12 male) vocalizing two statements, each embodying eight emotions, with 60 samples per actor. Samples vary in duration from 2.9 to 5.2 seconds, presented in 720x1080 pixels video at 30fps and audio at a 48,000Hz sampling rate.

Training Samples We use OpenFace (Baltrusaitis et al. 2018) to compute head pose, action unit activations, and landmark locations for each video frame. These landmarks serve as guides for extracting patches around the mouth and eyes. Specifically, bounding boxes calculated around designated landmarks facilitate the extraction of mouth (size 110x180) and eye patches (size 128x128). To have training samples with the same shape, only the central 88 frames and their corresponding audio segments are retained. This standardization is accomplished by trimming initial and final frames from longer videos, and aligning all samples with the dataset's minimum video duration (2.9s).

Feature Extraction We use two types: pre-trained large models and classical feature engineering.

- *Pretrained Models:* Employed for extracting features from speech (Wav2Vec2, base and large, yielding 768 and 1024 features respectively (Baevski et al. 2020)) and visual elements (ResNet (He et al. 2016), producing 512

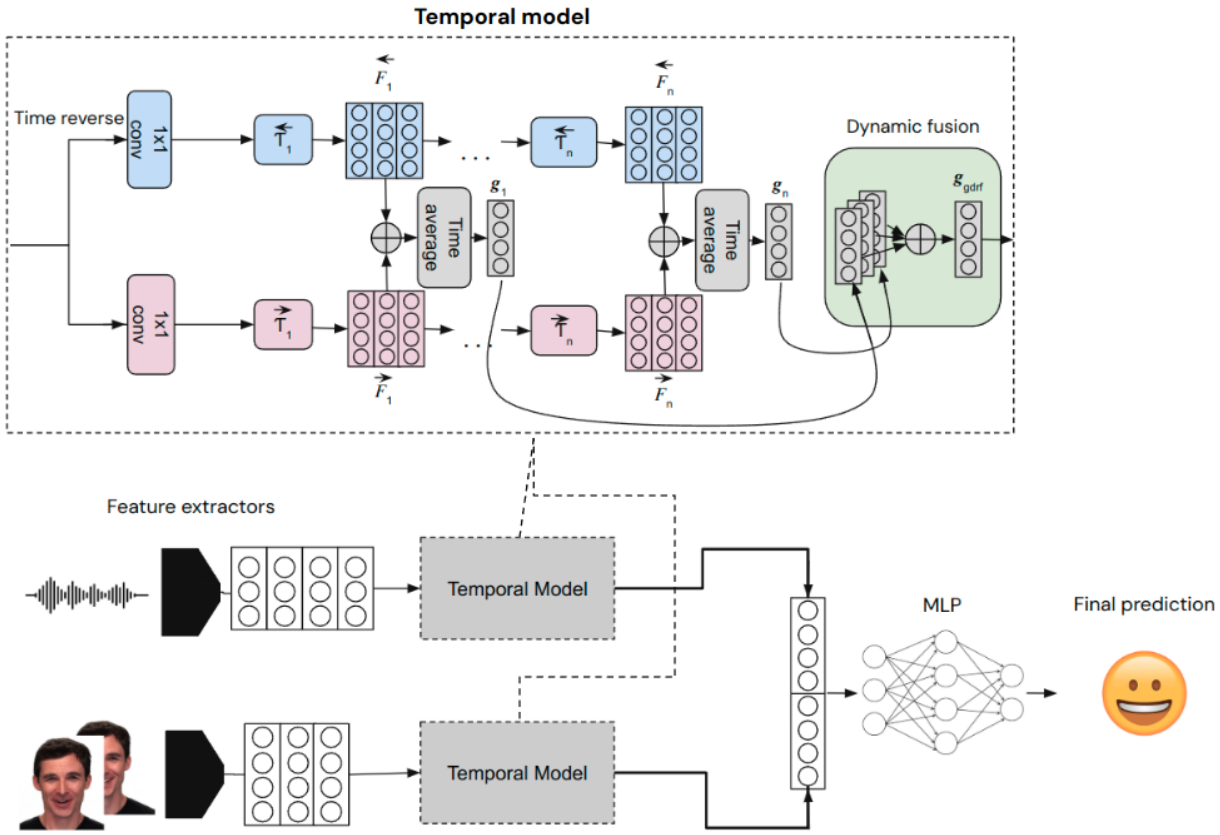

Figure 1: Illustration of the multi-modal model. Each temporal model is based on the TIM-Net architecture.

features for mouth, eyes, and face with averaged eye features). Due to computational limits, only 20 evenly-intervalled frames are utilized.

- *Feature Engineering:* For efficient feature extraction, Log Mel Spectrogram for speech is utilized, producing an 88x24 feature matrix, with parameters including 24 coefficients, a window length of 4800, a hop length of 1600, and a maximum frequency of 17kHz. Visual features involve computed and normalized key point distances, resulting in 780-dimensional vectors for faces and 179-dimensional vectors for eyes per frame. Normalization ensures consistent facial attribute sizes across individuals. Other features include head position (pitch, yaw, roll angles, and 3D location), and action units (binary and intensity, normalized to 0-1).

All details related to data preparation of each modality used are provided in the supplementary material.

Classification Model We run experiment with a multi-modal setup drawing inspiration from TIM-Net (Ye et al. 2023). TIM-Net is designed to discern the temporal structure within features extracted from each modality, aiming to classify inputs into one of eight emotions. Our multi-modal setup ingests time-series features per frame, processed through the TIM-Net backbone. These features are then concatenated and funneled through a multi-layer perceptron (MLP) with 3 layers for final emotion prediction. The model is trained end-to-end; when pre-trained models serve as the feature extraction backbone, these are also subject to fine-tuning. Refer to Fig. 1 for the model's depiction. In scenarios utilizing a single modality, our multi-modal setup simplifies to TIM-Net (Ye et al. 2023).

Experiments

We assess various models for emotion classification through binary and multiclass classification methods to explore the efficacy of different features, modalities, and the impact of pre-trained versus engineered features. Models were trained and evaluated through cross-validation with 5 splits, following the methodology outlined in (Luna Jiménez et al. 2021). All details regarding the hyperparameters used are provided in the supplementary material.

Binary Classification Binary classification predicts the presence or absence of emotions. Due to the dataset's imbalance, we use the macro F1 score for evaluation and employ weighted binary cross-entropy loss during training. Table 1 shows that pre-trained models, such as Wav2Vec2 for audio and ResNet for video, consistently outperformed traditional feature engineering techniques. The highest F1 score achieved is 0.93 for the emotion "disgust" using

Modality	Type	Features	Emotion							
			Angry	Calm	Disgust	Fear	Happy	Neutral	Sad	Surprise
Audio	Speech	LMS	0.83	0.83	0.85	0.76	0.79	0.70	0.68	0.81
		Wav2Vec2-B	**0.91**	**0.88**	**0.93**	**0.87**	0.85	**0.82**	0.75	**0.88**
Video	Face	ResNet	0.81	0.82	0.84	0.76	**0.90**	0.75	0.75	0.70
		KPD	0.78	0.82	0.85	0.73	0.89	0.79	**0.76**	0.67
		AU	0.77	0.78	0.79	0.67	0.87	0.63	0.62	0.63
	Eyes	ResNet	0.68	0.72	0.77	0.75	0.75	0.68	0.69	0.65
		KPD	0.69	0.69	0.69	0.73	0.73	0.73	0.69	0.59
	Mouth	ResNet	0.80	0.77	0.80	0.67	0.88	0.77	0.63	0.68
	Head Pose	RPY-3D	0.65	0.65	0.56	0.54	0.55	0.55	0.53	0.53

Table 1: Macro F1 Score for different emotions, modalities, features. LMS = Log Mel Spectrogram, Wav2Vec2-b = Wav2Vec2-Base, KPD = Key-Point Distances, AU = Action Units, RPY-3D = Roll, Pitch, Yaw and 3D location.

Type→ ↓		Speech		
	Feature→ ↓	None	LMS	Wav2Vec-L
Eyes	ResNet	53.18	53.18	82.48
	KPD	52.35	70.30	81.42
Mouth	ResNet	61.58	59.97	82.04
Eyes + Mouth	ResNets	63.48	60.98	83.51
Face	ResNet	71.20	69.58	**83.57**
	KPD	67.13	**76.82**	81.63
	AU	58.63	70.93	81.65

Table 2: Accuracy for multi-modal experiments. "None" in the first numerical row and column indicates single-modality cases; other entries are multi-modal. LMS = Log Mel Spectrogram, Wav2Vec2-L = Wav2Vec-Large, KPD = Key-Point Distances, AU = Action Units.

speech features extracted through the Wav2Vec2 model. Interestingly, head pose information also contributes positively to emotion classification performance, especially for emotions "angry" and "calm". Notably, the experiments reveal a variation in the predictability of different emotions using visual and auditory features. While emotions like "angry" and "surprise" are best predicted through audio features, others, namely "happy" and "sad" are more accurately classified using visual features. For emotion "happy", even when only using mouth-related visual features, the models still surpassed the speech-only performance, underscoring the significance of visual cues in detecting certain emotions.

Multiclass Classification Multiclass classification is the task of identifying one of the 8 possible emotions. The models are evaluated based on accuracy metric and trained with unweighted cross-entropy loss, as the dataset is balanced across different emotions. Table 2 presents the accuracy of various models, that combine audio and visual modalities with different feature extraction techniques, revealing that the integration of audio and visual modalities enhances classification accuracy.

As in the binary classification task, pre-trained models continue to dominate, yielding superior performance compared to feature engineering methods. Combining Log Mel Spectrogram and Key-Point Distance features elevates accuracy from 65.6% and 67.13% to 76.82%, a significant improvement especially beneficial for head-mounted displays (HMDs). The highest accuracy (83.57%) was achieved by combining whole-face and speech features with pre-trained models, highlighting the synergy of audio-visual cues in emotion classification. In contrast to the binary classification tasks where the base version was used, this analysis utilized the large version of Wav2Vec2 (Baevski et al. 2020) for comparison with the 86.7% accuracy reported by (Luna Jiménez et al. 2021).

Conclusion

This study underscores the importance of multi-modal approaches in enhancing emotion classification accuracy. The integration of speech and facial features is particularly effective, offering a nearly 10% accuracy increase in efficiently engineered models, a significant advancement vital for resource-limited devices such as Head-Mounted Displays (HMDs). Furthermore, the findings reveal that speech and facial features are selectively proficient at identifying different emotions. This distinction not only supports the efficacy of a multi-modal strategy but also provides a nuanced understanding of emotion classification. This, in turn, has the potential to benefit a wide range of applications and industries, making emotion classification an even more valuable tool in human-computer interaction and beyond.

References

Akçay, M. B.; and Oğuz, K. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116: 56–76.

Ancilin, J.; and Milton, A. 2021. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, 179: 108046.

Andalibi, N.; and Buss, J. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. CHI '20, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 59–66. IEEE.

Bhavan, A.; Chauhan, P.; Hitkul; and Shah, R. R. 2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184: 104886.

Canal, F. Z.; Müller, T. R.; Matias, J. C.; Scotton, G. G.; de Sa Junior, A. R.; Pozzebon, E.; and Sobieranski, A. C. 2022. A Survey on Facial Emotion Recognition Techniques: A State-of-the-Art Literature Review. *Inf. Sci.*, 582(C): 593–617.

Chowanda, A.; Sutoyo, R.; Meiliana; and Tanachutiwat, S. 2021. Exploring Text-based Emotions Recognition Machine Learning Techniques on Social Media Conversation. *Procedia Computer Science*, 179: 821–828. 5th International Conference on Computer Science and Computational Intelligence 2020.

Cowen, A. S.; and Keltner, D. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38): E7900–E7909.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1): 32–80.

Deng, J. J.; and Leung, C. H. C. 2021. Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition. In *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings*, 179–188. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-86992-2.

Ekman, P.; and Friesen, W. V. 1978. Facial action coding system: a technique for the measurement of facial movement.

Ewart J. de Visser, R. P.; and Shaw, T. H. 2018. From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, 61(10): 1409–1427.

Greco, C.; Matarazzo, O.; Cordasco, G.; Vinciarelli, A.; Callejas, Z.; and Esposito, A. 2021. Discriminative Power of EEG-Based Biomarkers in Major Depressive Disorder: A Systematic Review. *IEEE Access*, 9: 112850–112870.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. 770–778.

King, D. E. 2009. Dlib-Ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.*, 10: 1755–1758.

Li, S.; and Deng, W. 2018. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, PP.

Li, X.; Zhang, Y.; Tiwari, P.; Song, D.; Hu, B.; Yang, M.; Zhao, Z.; Kumar, N.; and Marttinen, P. 2022. EEG Based Emotion Recognition: A Tutorial and Review. *ACM Computing Surveys*, 55(4): 1–57.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5): 1–35.

Luna Jiménez, C.; Kleinlein, R.; Griol, D.; Callejas, Z.; Montero, J.; and Fernández-Martínez, F. 2021. A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. *Applied Sciences*, 12: 327.

Mandler, G. 1997. *The Psychology of Facial Expression*. Studies in Emotion and Social Interaction. Cambridge University Press.

Marín-Morales, J.; Llinares, C.; Guixeres, J.; and Alcañiz, M. 2020. Emotion Recognition in Immersive Virtual Reality: From Statistics to Affective Computing. *Sensors*, 20(18): 5163.

Minaee, S.; Minaei, M.; and Abdolrashidi, A. 2021. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors*, 21(9).

Nyquist, A.; and Luebbe, A. 2020. An Emotion Recognition–Awareness Vulnerability Hypothesis for Depression in Adolescence: A Systematic Review. *Clinical Child and Family Psychology Review*, 23.

Pepino, L.; Riera, P.; and Ferrer, L. 2021. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Proc. Interspeech 2021*, 3400–3404.

Rajamani, S.; Rajamani, K.; Mallol-Ragolta, A.; Liu, S.; and Schuller, B. 2021a. A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition. 6294–6298.

Rajamani, S.; Rajamani, K.; Mallol-Ragolta, A.; Liu, S.; and Schuller, B. 2021b. A Novel Attention-Based Gated Recurrent Unit and its Efficacy in Speech Emotion Recognition. 6294–6298.

Rizos, G.; Baird, A.; Elliott, M.; and Schuller, B. 2020. Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3502–3506.

Sanchez-Mendoza, D.; Masip, D.; and Lapedriza, A. 2014. Emotions classification using facial action units recognition. *Frontiers in Artificial Intelligence and Applications*, 269: 55–64.

Schuller, B.; and Batliner, A. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley Publishing, 1st edition. ISBN 1119971365.

Schuller, B. W. 2018. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM*, 61(5): 90–99.

Singh, P.; Srivastava, R.; Rana, K.; and Kumar, V. 2021. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, 229: 107316.

Sun, L.; Mingyu, X.; Lian, Z.; Liu, B.; Tao, J.; Wang, M.; and Cheng, Y. 2021. Multimodal Emotion Recognition and Sentiment Analysis via Attention Enhanced Recurrent Model. 15–20.

Tompkins, D.; Emmanouilidou, D.; Deshmukh, S.; and Elizalde, B. 2023. Multi-View Learning for Speech Emotion Recognition with Categorical Emotion, Categorical Sentiment, and Dimensional Scores. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Wani, T. M.; Gunawan, T. S.; Qadri, S. A. A.; Kartiwi, M.; and Ambikairajah, E. 2021. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*, 9: 47795–47814.

Yao, L.; Wan, Y.; Ni, H.; and Xu, B. 2021. Action unit classification for facial expression recognition using active learning and SVM. *Multimedia Tools and Applications*, 80.

Ye, J.; Wen, X.; Wei, Y.; Xu, Y.; Liu, K.-H.; and Shan, H. 2022. Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition.

Ye, J.; Wen, X.-C.; Wei, Y.; Xu, Y.; Liu, K.; and Shan, H. 2023. Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; and Picard, R. 2020. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Computing Surveys*, 53: 1–30.

Supplementary Material

Data preparation

In this section, we outline the steps taken to prepare each modality for our experiments, ensuring the reproducibility of our results.

- *ResNet:* ResNet served as the feature extractor for eyes, mouth, and face in our experiments. All images were resized to 224x224 pixels. For facial images, each frame was first center-cropped to 720x720 pixels to remove the borders, which did not contain facial information. These cropped frames were then scaled from 0.0 to 1.0 and normalized using mean=[0.485, 0.456, 0.406] and standard deviation (std)=[0.229, 0.224, 0.225], following the pre-training procedures outlined in PyTorch's documentation. In unimodal and binary experiments where fine-tuning was performed, we selected 20 equally spaced samples from the full video. For multimodal experiments, we employed the weights derived from the unimodal experiments, froze the ResNet extractor, and utilized the 88 middle frames of each video. We experimented with using 88 center frames for the unimodal experiments (with a frozen ResNet fine-tuned using 20 equally spaced samples from the video) to ensure that the superiority of our multimodal results was not merely due to the increased number of frames. However, this approach yielded slightly inferior results (0.3-1% reduced accuracy), and therefore, these findings are not reported in our tables.
- *Video key-points:* For video key-points, we used OpenFace to extract 40 3D landmarks from the initial 68, as some landmarks were closely positioned and didn't offer substantial additional information, hence were omitted to simplify the feature space. We calculated the pairwise distance between each landmark, constructing a 780-dimensional feature vector for each of the 88 middle frames in the video. These distances were then normalized against those from the most neutral frame—identified as the frame with the lowest activated action units—from the first repetition of a neutral video where the actor vocalized the same sentence. This process ensures a consistent and comparative basis for analysis across different frames and videos.
- *Eyes key-points:* For eyes key-points, we adopted an approach similar to the one used for video key-points. We extracted 12 specific key-points within each eye, designated as eye_lmk_i in OpenFace, and computed the pairwise distances within each eye individually. Additionally, pairwise distances between 10 chosen key-points located on both eyes and the eyebrows were concatenated to the initial distances to incorporate inter-eye values. These distances were normalized using the same procedure previously described for video key-points. The average gaze angles of both eyes across two axes were then calculated and appended to the data, resulting in a 179-dimensional vector.
- *Action units (AU):* OpenFace supplies both binary action units (either activated or not, for 18 AUs) and a continuous intensity measure (ranging from 0 to 5, for 17 AUs). As these two sets of values are generated through different models, there might be inconsistencies in their correspondence. To address this, we integrated both types of values. The continuous intensities were first rescaled to a range between 0 and 1. Following this, we concatenated the rescaled intensities with the binary action units to form a combined 35-dimensional vector for each of the 88 frames.
- *Audio:* We applied straightforward preprocessing to the audio. This involved extracting a segment of 140,800 units in length from the center of the audio, which is equivalent to the duration of 88 frames or approximately 2.9 seconds. The stereo signal was then converted into a mono signal by averaging the two channels. In the case of the Wav2Vec experiments, the audio signal was resampled from 48KHz down to 16KHz and normalized by subtracting the mean and dividing by the standard deviation.

Hyperparameters

In this section, we offer a comprehensive overview of the architecture's hyperparameters and provide detailed training information necessary for reproducing our results. The complete set of hyperparameters, along with specific training details are shown in Table 3.

In our experiments, the MutiNet network utilized three layers within its temporal block, compared to the original TIM-Net's two, and employed a kernel of size 2 for 1D convolutions. We experimented with both 32 and 64 convolution channels, with 64 yielding superior results in most cases (the exceptions being the multi-class eyes ResNet and LMS + Eyes KPD experiments, reported with 32 channels only). The dilation factors used for multi-scale feature extraction were always powers of two, consistent with the original paper. When scales equal n, it indicates the use of n different temporal blocks with dilation factors ranging from 2^0 to 2^{n-1}. These factors were selected to ensure the largest temporal block dilation factor was smaller than the total feature temporal length (hence for 88 frames, we chose 7 so that $2^6 = 64 < 88$). We always used 4 folds for training and 1 for evaluating our models and reported the average over the 5 evaluation folds. Lastly, we selected model checkpoints with the lowest training loss for the results, which outperformed the final checkpoints.

Model	Experiment	Learning Rate	Batch Size	Epochs	Scales
Wav2Vec2-L Wav2Vec2-B	both	0.00005	16	100	8
LMS	both	0.001	64	100	7
Video ResNet Mouth ResNet Eyes ResNet	both	0.0002	8	60	5
Face KPD Eyes KPD	multi-class	0.0005	64	100	7
Face KPD Eyes KPD	binary	0.0002	64	100	7
AU	both	0.0002	64	40	7
Multimodal with Wav2Vec-L + ...	multi-class	0.0001	64	100	7
Multimodal with LMS + ...	multi-class	0.002	64	100	7

Table 3: Hyperparameters used in the experiments. Each experiment type is categorized as binary, multi-class, or both; "both" is used when identical hyperparameters were applied to both experiment types. "Scales" denotes the count of distinct scales at which each modality's features were extracted in the temporal model before being fused by the dynamic fusion module.