In-depth Study and Application Analysis of Multimodal Emotion Recognition Methods: Multidimensional Fusion Techniques Based on Vision, Speech, and Text

Yuke Lei

Hainan International College, Communication University of China, Lingshui, China

leiyuke@cuc.edu.cn

Abstract. Emotion recognition technology, pivotal in fields such as medical health, game entertainment, and human-computer interaction, benefits significantly from multimodal approaches. This paper delves into the techniques and applications of multimodal emotion recognition, focusing on fusion methods that integrate visual, speech, and text data. Emotion recognition through single modalities often faces limitations such as susceptibility to noise and low accuracy, whereas multimodal systems exhibit enhanced performance by leveraging combined data sources. The primary fusion techniques discussed include feature-level, decision-level, and model-level integrations. Feature-level fusion amalgamates multiple data types early in the processing stage, improving detection robustness. Decision-level fusion, on the other hand, involves synthesizing results from separate analyses, offering flexibility and ease of integration. Model-level fusion allows for deep interactions between modalities, potentially capturing more complex emotional states. This study confirms that multimodal emotion recognition systems generally surpass the performance of their single-modal counterparts, advocating for further exploration into sophisticated fusion techniques to boost accuracy and applicability across various domains.

Keywords: Multimodal, emotion recognition, fusion.

1. Introduction

The significance of emotion recognition in enhancing interactions across diverse fields such as healthcare, entertainment, and human-computer interface is profound. Emotions, universally expressed across cultures through consistent yet complex patterns, are pivotal in shaping human interactions and decision-making processes. Historically, emotion recognition was confined to unimodal methodologies which analyze singular aspects such as facial expressions, voice intonation, or textual sentiment. However, these methods face challenges like high susceptibility to noise and context variability, prompting a shift towards more robust multimodal approaches that integrate multiple streams of emotional data.

Recent advancements have significantly broadened the scope of emotion recognition. Research now extends beyond traditional unimodal approaches to incorporate multimodal techniques, which synergize inputs from visual, auditory, and textual data to enhance accuracy and reliability. Current methodologies in multimodal emotion recognition employ advanced machine learning techniques, including deep learning frameworks that effectively amalgamate features from diverse sources. Despite these advancements, the field still grapples with challenges such as data fusion complexity, the need for extensive computational resources, and maintaining high accuracy across varied and dynamic real-world settings [1].

This paper primarily explores multidimensional fusion techniques in multimodal emotion recognition, aiming to refine the integration of visual, speech, and text data to improve both efficacy and efficiency in emotion detection. The study examines various fusion strategies like feature-level, decision-level, and model-level fusion, each offering distinct advantages and suited for different application scenarios. Feature-level fusion merges various data types at an early stage, enhancing model robustness against noise [2]. Decision-level fusion synthesizes separate analytical results at the end stage, offering flexibility and ease of model integration. Model-level fusion, meanwhile, fosters deeper interaction between modalities, potentially capturing more nuanced emotional expressions. This research not only contributes to theoretical advancements in emotion recognition but also underscores practical implementations, particularly in enhancing interactive systems and creating empathetic human-computer interactions.

2. Fundamental Theories and Methods of Unimodal Emotion Recognition

Emotion recognition and analysis is mainly through the facial expression, voice and other ways to express emotions to analyze the emotions of people at that time [3]. At present, the current mainstream single-modal emotion analysis mainly includes emotion analysis based on visual, speech and text emotion information. Emotion recognition based on different modals can be roughly divided into three steps: information collection, feature extraction and classification. In the following, the principle, technology and application of the three different modals are described respectively.

2.1. Principles and techniques of visual emotion recognition

In daily life, people's emotions are expressed through facial expressions, which play a very important role in people's communication. Visual emotion recognition basically refers to facial emotion recognition (FER), containing three stages: faces detection, feature extraction and classification [4].

The traditional FER method uses geometric and optical flow based feature extraction methods. In addition to classical methods, many studies use Neural Network based (NNB) approaches, and good results have been obtained. Figure 1 shows different classical methods for FER, as well as the framework of integrated CNN, which was offered by Lu et al [5]. This method based on CNN-integrated facial expression recognition designs three different structured subnetworks in a set of CNNS, respectively consisting of 3, 5, and 10 convolutional layers [6].

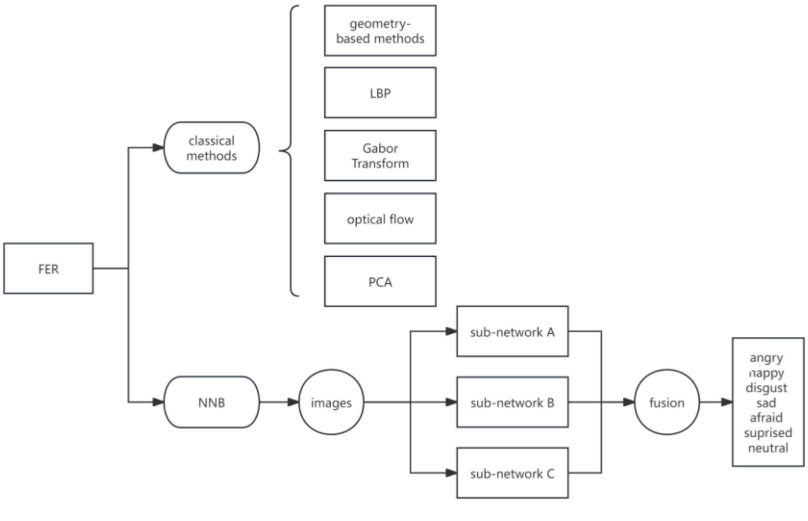

Figure 1. Different methods of FER (Photo credit: Original).

In order to take advantage of the advantages of LBP in traditional methods such as rotation invariance and insensitivity to illumination, a feature fusion FER method based on VGG-NET was proposed by Li et al, which fed LBP features and features extracted from CNN convolutional layer into the network connection layer of improved VGG-16 for weighted fusion [7]. Finally, the fused features are sent to Softmax classifier to obtain the probability of various features and complete the basic 6 expressions classification [8].

Methods based on deep learning make up for the shortcomings of classic approaches in facial expression feature extraction and improve the recognition effect. At the same time, there are also some problems, such as unstable models and inaccuracy data sets.

2.2. Core Technologies and challenges in speech emotion recognition (SER)

In general, features extracted from speech include prosodic, energy, language, vocal tract information and speaker information [9]. To recognize emotion from speech, acoustic features such as prosody, sound quality and spectral features are used. Figure 2 represents the scope of SER system.

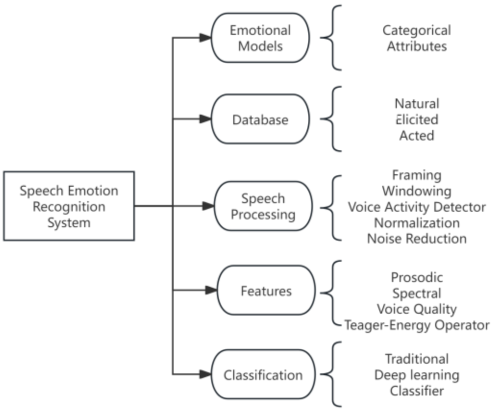
Speech Emotion Recognition System
Emotional Models
Categorical Attributes
Database
Natural Elicited Acted
Speech Processing
Framing Windowing Voice Activity Detector Normalization Noise Reduction
Features
Prosodic Spectral Voice Quality Teager-Energy Operator
Classification
Traditional Deep learning Classifier

Figure 2. General SER system [10].

Based on traditional machine learning: Traditional methods for speech emotion recognition include SVM, HMM, and Gaussian Mixture Model (GMM) among others. Lee used LDA, KNN and SVM for emotion classification and principal component analysis for feature extraction, whose results show that LDA has the best effect. J. et al used DBN (Deep Belief Network) and SVM (Support Vector Machine) for classification experiments, during which DBN achieved an accuracy of 94.6%, while SVM performed worse with the 84.54% accuracy [11].

Based on deep learning: Yang proposed a novel approach based on the WavLM representation [12]. Previous methods usually obtain emotional features using a trained model made up of interconnected layers and training data, but they overlook the significance of contextual information.

Although there is a lot of research and application of SER, challenges still remain: a. Feelings are complex, mixed and difficult to define; b. Feature selection cost is too high.

2.3. Algorithms and applications of text-based emotion analysis

Text-Based Emotion analysis refers to the search for information that can express opinions and emotions from the text. Unlike audio and images, text is relatively lacking in the richness of voice tone and facial expressions that express emotion [13]. Past research has proposed several approaches to emotion recognition using natural language processing (NLP).

Key word and lexicon-based approach: This method firstly extracts emotion words, and then estimates emotion according to the emotion polarity of the words and related words in emotion dictionary. However, they still found that full emotion keywords and respect for word semantics in meaning were missing [14]. They used BERT (bidirectional encoder representation from transformers) to generate sentence-specific representations of sentiment analysis, which enabled the model to make more accurate predictions.

Machine leaning approach: In traditional machine learning methods, a training set is created and annotated firstly, then relevant text features are extracted, and a model is constructed through machine learning methods to analyze the features [15]. The classification models used in the analysis include logistic regression, support vector machine, random forest, maximum entropy classification and so on.

Text sentiment analysis has a wide range of applications, including obtaining user satisfaction information, recommending products based on user emotions, and predicting emotions [16]. In this digital world, there is a great demand for people to exchange information, comment and publish articles online, which makes real-time text sentiment detection a future trend.

3. Fusion Techniques in Multimodal Emotion Recognition

Single-modal emotion recognition is easy to be affected by noise, and has limitations such as low accuracy and poor stability [17]. In order to make up for the defects of single modal, the researchers propose the method of multimodal emotion recognition, and find that the accuracy of most multimodal emotion recognition systems is higher than that of the corresponding optimal single-modal systems [18].

3.1. Implementation and optimization of feature-level fusion

Feature-level fusion is also called early feature-based approach, which is to extract the feature of each mode data and then fuse it into a unified feature set, as shown in figure 3.

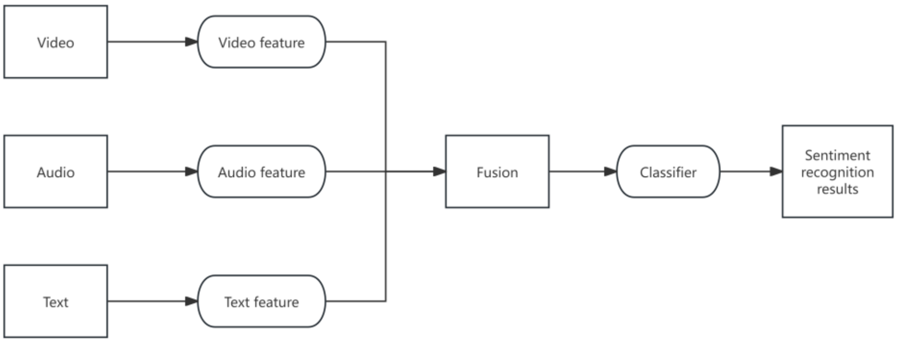

Figure 3. Overall framework of feature-level fusion (Photo credit: Original).

In the research based on feature-level fusion, a method to multimodal sequence modeling and analysis which is proposed by Louis-Philippe et al is to represent multimodal feature vectors as higher-order tensors and extract hidden states and transition probabilities using tensor decomposition methods [19].

Another method offered by Kumar et al enhances multimodal sentiment analysis by leveraging both video and text data [20, 21]. This selective learning approach fuses feature vectors, which are then integrated using a cross-attention mechanism. Then, Bi-GRU is used to extract deep feature vectors from each modality. Finally, these deep multimodal feature vectors are concatenated with softmax for sentiment analysis purposes.

Within the textual features, the text is segmented into three sections: the upper left context, the upper right context, and the target entity. Three LSTMs are employed to extract the contextual information and emotional attributes from these segments. For visual features, ResNet is used to extract visual attributes, while an attention mechanism determines the weight of each section. Subsequently, a gate recurrent unit (GRU) is incorporated to minimize image noise. Finally, the features from both modalities are amalgamated and fed into a softmax function for emotion analysis through feature-level fusion. This method makes comprehensive use of the information of each modal, but the more complex network structure makes the running time longer.

3.2. Strategies and applications of decision-level fusion

Such models are typically lightweight and adaptable. If any one modality is absent, decisions can still be made using the remaining modalities. Figure 4 illustrates the framework for a multimodal model based on late-stage decision-level fusion [22].

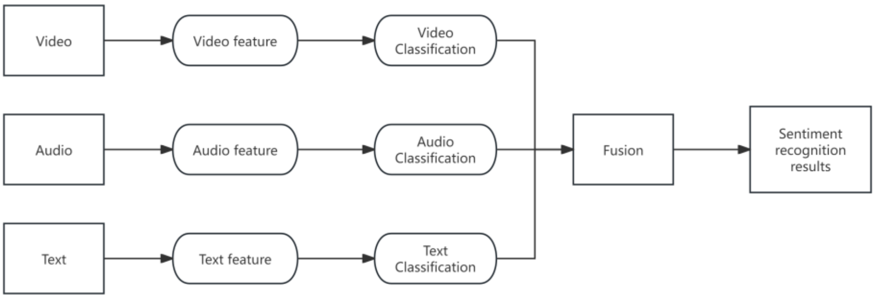

Figure 4. Overall framework of decision-level fusion (Photo credit: Original).

Although this model is simple and easy to apply, it cannot handle the interaction between different modalities. The final stage involves a fully connected layer that performs the classification.

In addition, Yu et al introduced a new loss function regression model called SDL (speaker-distribution loss) and a temporal selective attention model (TSAM), which consists of an attention module, an encoding module, and a speaker distribution loss function [23]. The attention module uses an LSTM to preprocess the sequence, and the encoding stage uses a BiLSTM to encode the sequence observations and weightedly combine them as the module's output. Finally, the output is sent to the SDL for sentiment analysis [24].

The statistical rules and probability theory used in the fusion process rely on the assumption that all classifiers are independent of each other, but this is obviously not practical. Lu et al adopted a fusion strategy called fuzzy integration. Fuzzy integration is the integral of a real function over a fuzzy measure [25]. The best accuracy of this method is 87.59%, which is shows the function of fuzzy integral in improving the accuracy of emotion recognition.

Let μ be a fuzzy measure on X. The discrete Choquet integral of a function $f: X \to \mathbb{R}^+$with respect to μ, is

$$\begin{aligned} &\mathcal{C}_\mu\left(f\left(x_1\right), f\left(x_2\right), \ldots, f(x_n)\right) \\ &:= \textstyle\sum_{i=1}^{n} \left[f\left(x_{(i)}\right) - f\left(x_{(i-1)}\right)\right] \mu\left(A_{(i)}\right) \end{aligned} \tag{1}$$

Where $\cdot_{(i)}$ presents the permuted indices to satisfy $0 \leq f\left(x_{(1)}\right) \leq f\left(x_{(2)}\right) \leq \cdots \leq f\left(x_{(n)}\right) \leq 1$. Also $f\left(x_{(0)}\right) = 0$ and $A_{(i)} := \left\{x_{(i)}, x_{(i+1)}, \ldots, x_{(n)}\right\}$.

3.3. Design and case studies of model-level fusion

The process allows for the selection of the site for modality feature fusion, facilitating intermediary interactions [26]. Figure 5 illustrates the framework of a multimodal model that utilizes model-level fusion.

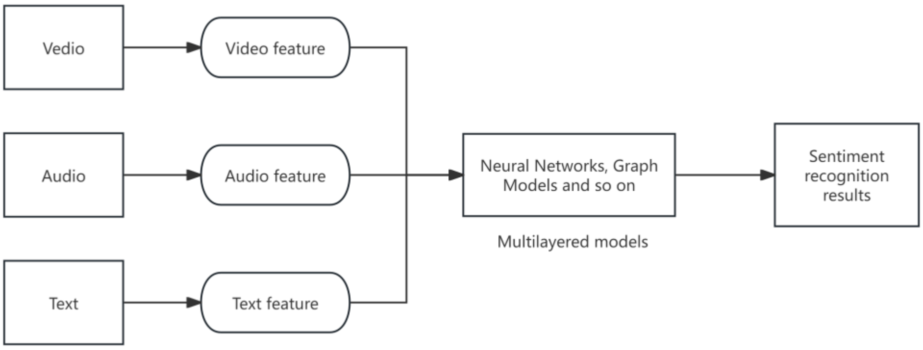

Figure 5. Overall framework of model-level fusion (Photo credit: Original).

The current model-level fusion mainly takes the strategy of building deep neural network models with multiple layers, where each layer learns increasingly complex transformations to match increasingly complex features and enhance non-linear expressive capabilities. Zhang et al used CNN (convolutional neural networks) and 3DCNN (three dimensional convolutional neural networks) to generate fragment features from audiovisual data. It is then fused into a deep confidence network [27].

Although the self-supervised learning method makes the method have high accuracy and robustness, the result is still affected by the quality of data annotation.

4. Application Scenarios and Case Studies of Multimodal Emotion Recognition

4.1. Applications and practices in the healthcare sector

Doctors or medical institutions can use emotion recognition systems to identify patients' emotional states more accurately, so as to obtain more targeted and professional diagnosis and treatment plans.

For example, doctors can better identify patients' emotional states in medical records and symptom descriptions based on the Multi-modal conversational emotion analysis system, so as to better interact with patients and make more accurate diagnosis and treatment. At the same time, the recording of mood changes also provides a reference for the design of subsequent treatment course.

Especially in mental health, multimodal emotion recognition can provide early intervention, personalized treatment, and remote monitoring.

4.2. Emotion recognition and feedback mechanisms in human-computer interaction

The multimodal emotion recognition system can enhance the user's participation, experience and satisfaction with the product. The machine can adjust the feedback and interaction with the user based on the emotion it recognizes.

At the same time, taking emotions into account in the feedback mechanism of human-computer interaction can provide more personalized and humane services. This approach also makes human-computer interaction more promising in education, gaming, and healthcare applications.

4.3. Innovative applications in the entertainment and gaming industries

Through multimodal emotion recognition, game designers can understand the game experience and personality preferences of players based on their facial expressions, voices, and emotional changes, so as to set game difficulty changes and level design. This technology can help the game bring more immersive experience and emotional resonance to the player, making the player more happy and satisfied during the game.

5. Conclusion

This paper has comprehensively explored the methodologies and practical applications of multimodal emotion recognition, focusing on advanced fusion techniques that amalgamate data from visual, speech, and text modalities. The investigation has underscored the superiority of multimodal systems over unimodal approaches, primarily due to their enhanced robustness against noise, higher accuracy, and the ability to capture more nuanced emotional states through the integration of diverse data sources. Through detailed analysis of feature-level, decision-level, and model-level fusion techniques, this study has illuminated their distinct advantages and suitability for various real-world applications, ranging from healthcare diagnostics to interactive gaming and human-computer interaction systems.

The field of multimodal emotion recognition presents several avenues for further research. Key among these is the development of more sophisticated fusion algorithms that can seamlessly integrate increasingly complex data streams while maintaining computational efficiency. There is also a significant need for the creation of extensive, diverse, and well-annotated datasets that can train more generalized models capable of functioning accurately across different cultures and contexts. Additionally, future research should address the ethical implications of emotion recognition technology, particularly concerns related to privacy, consent, and the potential for misuse. As the technology advances, ensuring that these systems are used responsibly and ethically will become paramount. By continuing to refine the technologies and address these challenges, multimodal emotion recognition can profoundly impact a wide range of sectors, enhancing both machine understanding and responsiveness to human emotional states.

References

[1] Canal F. Z., Müller T. R., Matias J. C., Scotton G. G., de Sa Junior A. R., Pozzebon E., Sobieranski A. C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences, 2022, 582: 593–617. https://doi.org/10.1016/j.ins. 2021.10.005.

[2] LU J. H., ZHANG S. M., ZHAO J. L. Facial expression recognition based on CNN ensemble. Journal of Qingdao University (Engineering & Technology Edition), 2020, 35(2): 24-29.

[3] LI X. L., NIU H. T. Facial expression recognition using feature fusion based on VGG-NET. Computer Engineering & Science, 2020, 42(3): 500-509.

[4] Khaireddin Y., Chen Z. Facial emotion recognition: State of the art performance on FER2013. arXiv preprint arXiv:2105.03588, 2021.

[5] Gaonkar A., Chukkapalli Y., Raman P. J., Srikanth S., Gurugopinath S. A comprehensive survey on multimodal data representation and information fusion algorithms. 2021 International Conference on Intelligent Technologies (CONIT). https://doi.org/10.1109/conit51480.2021.9498415.

[6] Chul Min Lee, Narayanan S. S., Pieraccini R. Classifying emotions in human-machine spoken dialogs. IEEE International Conference on Multimedia and Expo, n.d. https://doi.org/10.1109/icme. 2002.1035887.

[7] Zhang W., Zhao D., Chai Z., Yang L. T., Liu X., Gong F., Yang S. Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. Softw., Pract. Exper., 2017, 47(8): 1127–1138. doi: 10.1007/978-3-319-39601-9_5.

[8] Lee J., Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition. Proc. 16th Annu. Conf. Int. Speech Commun. Assoc., 2015: 1–4.

[9] Yang J., Liu J., Huang K., Xia J., Zhu Z., Zhang H. Single-and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding. Electronics, 2024, 13(7): 1380.

[10] Wani T. M., Gunawan T. S., Qadri S. A. A., Kartiwi M., Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access, 2021, 9: 47795-47814.

[11] Seal D., Roy U. K., Basak R. Sentence-level emotion detection from text based on semantic rules. Information and Communication Technology for Sustainable Development. Springer, Singapore, 2020: 423–430.

[12] CHEN F., YUAN Z., HUANG Y. Multi-source data fusion for aspect-level sentiment classification. Knowledge-Based Systems, 2020, 187: 104831.

[13] Wang Y. R. A review of multimodal sentiment analysis algorithms. Computer programming skills and maintenance, 2021.

[14] Zhu X., Xu H., Zhao Z., Wang X., Wei X., Zhang Y., & Zuo J. 2021 An environmental intrusion detection technology based on WiFi (Wireless Personal Communications) 119(2): 1425-1436

[15] Bharti T. S. K., Varadhaganapathy S., Gupta R. K., Shukla P. K., Bouye M., Hingaa S. K., Mahmoud A. Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience, 2022, 1: 2645381.

[16] Morency L. P., Mihalcea R., Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th international conference on multimodal interfaces, 2011: 169–176.

[17] Zhu X., Huang Y., Wang X., & Wang R. 2024 Emotion recognition based on brain-like multimodal hierarchical perception (Multimedia Tools and Applications) 83(18): 56039-56057

[18] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. International Journal of Multimedia Information Retrieval, 2024, 13(4): 39.

[19] YU J., JIANG J., XIA R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 429-439.

[20] Nojavanasghari B., Gopinath D., Koushik J., Baltrušaitis T., Morency L. P. Deep multimodal fusion for persuasiveness prediction. Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016: 284–288.

[21] Wang H., Meghawat A., Morency L. P., Xing E. P. Select-additive learning: Improving generalization in multimodal sentiment analysis. 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017: 949–954. IEEE.

[22] Zhu X., Guo C., Feng H., Huang Y., Feng Y., Wang X., & Wang R. 2024 A Review of Key Technologies for Emotion Analysis Using Multimodal Information (Cognitive Computation) 1(1): 1-27

[23] LU Y., ZHENG W., LI B., et al. Combining eye movements and EEG to enhance emotion recognition. Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence, 2015: 1170−1176.

[24] Lai S., Hu X., Xu H., Ren Z., Liu Z. Multimodal sentiment analysis: A survey. Displays, 2023: 102563.

[25] Siriwardhana S., Reis A., Weerasekera R., Nanayakkara S. Jointly fine-tuning “BERT-like” self-supervised models to improve multimodal speech emotion recognition. arXiv preprint arXiv:2008.06682, 2020.

[26] Zhang S., Zhang S., Huang T., Gao W., Tian Q. Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 3030–3043. https://doi.org/10.1109/tcsvt.2017.2719043.

[27] Kalateh S., Estrada-Jimenez L. A., Hojjati S. N., Barata J. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. IEEE Access, 2024.