applied sciences

MDPI

Article

HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention

Moung-Ho Yi [1], Keun-Chang Kwak [1] and Ju-Hyun Shin [2,*]

[1] Department of Electronic Engineering, Chosun University, Gwangju 61452, Republic of Korea; audgh3710@gmail.com (M.-H.Y.); kwak@chosun.ac.kr (K.-C.K.)
[2] Department of New Industry Convergence, Chosun University, Gwangju 61452, Republic of Korea
* Correspondence: jhshinkr@chosun.ac.kr

Abstract: Emotion recognition is becoming increasingly important for accurately understanding and responding to user emotions, driven by the rapid proliferation of non-face-to-face environments and advancements in conversational AI technologies. Existing studies on multimodal emotion recognition, which utilize text and speech, have aimed to improve performance by integrating the information from both modalities. However, these approaches have faced limitations such as restricted information exchange and the omission of critical cues. To address these challenges, this study proposes a Hybrid Multimodal Transformer, which combines Intermediate Layer Fusion and Last Fusion. Text features are extracted using KoELECTRA, while speech features are extracted using HuBERT. These features are processed through a transformer encoder, and Dual Cross Modal Attention is applied to enhance the interaction between text and speech. Finally, the predicted results from each modality are aggregated using an average ensemble method to recognize the final emotion. The experimental results indicate that the proposed model achieves superior emotion recognition performance compared to existing models, demonstrating significant progress in improving both the accuracy and reliability of emotion recognition. In the future, incorporating additional modalities, such as facial expression recognition, is expected to further strengthen multimodal emotion recognition capabilities and open new possibilities for application across diverse fields.

Keywords: KoELECTRA; HuBERT; Dual Cross Modal Attention; Hybrid Multimodal Transformer; emotion recognition

Academic Editor: Alexander N. Pisarchik

Received: 30 December 2024
Revised: 19 January 2025
Accepted: 20 January 2025
Published: 21 January 2025

Citation: Yi, M.-H.; Kwak, K.-C.; Shin, J.-H. HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention. *Appl. Sci.* **2025**, *15*, 1053. https://doi.org/10.3390/app15031053

1. Introduction

Emotion recognition has emerged as a critical element in the field of human–computer interaction (HCI), enabling precise understanding of user emotions and appropriate responses. Recently, with the widespread adoption of non-face-to-face environments, conversational AI systems have become essential across various industries, including remote education, healthcare, and emotional labor support. These systems actively interact with users, provide information, solve problems, and deliver personalized experiences tailored to individual needs and preferences, thereby enhancing user satisfaction. In the field of remote education, emotion recognition plays a vital role in understanding learners' emotional states and motivations in real time, enabling the provision of the most suitable learning environment. By analyzing student focus or engagement levels, it becomes possible to offer personalized learning content, maximizing learning outcomes. A study developed instructional design principles incorporating multimodal elements to improve student concentration in remote learning contexts [1]. Another study explored changes in teachers' perceptions of online classes following the COVID-19 pandemic, systematically identifying the strengths and weaknesses of online education and seeking new educational possibilities based on these insights [2]. In healthcare, the potential applications of emotion recognition technology are significant. By continuously monitoring patients' psychological and emotional states, this technology can propose optimized healthcare and treatment strategies tailored to individuals. Research on IoT-based smart healthcare systems has demonstrated their ability to collect and analyze personalized health data, enabling customized medical care and status monitoring [3]. Additionally, AI-based well-being support systems for the elderly have been shown to induce psychological stability, alleviate depression, and improve quality of life in older adults [4]. Emotion recognition also plays a crucial role in systems supporting emotional labor. Due to the inherently emotionally demanding nature of their work, emotional laborers require appropriate feedback to manage stress and fatigue effectively. One study introduced an AI-powered smart healthcare exercise management system that monitors individuals' health in real time and provides solutions to reduce fatigue [5]. Research on personalized health management systems for home patients has also highlighted their effectiveness in monitoring daily health conditions and facilitating medical intervention when necessary, creating a more efficient environment for both patients and healthcare providers [6]. These prior studies underscore the importance of emotion recognition models capable of accurately capturing user emotions. The model proposed in this study reflects this necessity and is expected to contribute significantly to enhancing user-centric services across various domains, including remote education, healthcare, and emotional labor support.

Existing studies on multimodal emotion recognition, which simultaneously utilize text and speech, have aimed to enhance the accuracy and reliability of emotion recognition by complementing subtle emotional signals that are difficult to capture with a single modality. For example, Seunghyun Yoon's study [7] extracted features such as MFCC (mel-frequency cepstral coefficients), energy, and pitch from speech data, while text was processed as transcriptions using a Bag-of-Words approach, with feature vectors from both modalities simply concatenated. Young-Jun Kim's study [8] converted speech signals into spectrograms, extracting features through CNN, while the text modality employed word embeddings processed via a CNN-LSTM architecture to learn sequential information. By averaging the output values of the two models, the study effectively combined time–frequency domain features with linguistic contextual information to predict emotions. Similarly, H. Park's study [9] combined text embeddings generated by KoBERT with CNN-based speech analysis, using a weighted average ensemble method to comprehensively understand user states. These multimodal approaches share the common goal of improving learning performance by integrating complementary information provided by text and speech. However, they often face limitations, such as restricted information exchange between modalities or the omission of critical cues, leading to reduced performance improvements. To overcome these challenges, this study proposes the HyFusER (Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention) model. Unlike previous studies that relied on simple concatenation or independent learning structures between modalities, the HyFusER model strengthens interactions between text and speech to effectively integrate information without losing key cues. Specifically, the Dual Cross Modal Attention mechanism focuses on maximizing the exchange of information by reinterpreting text and speech data in each other's contexts. This approach enables the model to fully exploit the complementary characteristics of the two modalities, paving the way for more accurate and reliable emotion recognition.

The proposed model in this study begins by independently processing text and speech data through separate transformer encoders for each modality. Subsequently, the Dual

Cross Modal Attention mechanism is employed to strengthen interactions between text and speech. In the first stage, text is set as the primary input (Query), while speech serves as the auxiliary input (Key and Value). This configuration allows text to effectively learn the acoustic characteristics and speech patterns embedded in the speech data. In the second stage, the roles are reversed. speech becomes the primary input (Query), and text serves as the auxiliary input (Key and Value). This enables speech to learn the linguistic context and semantic structure of the text data. This iterative process enhances the interaction between text and speech, facilitating efficient exchange and fusion of modality-specific information. As a result, the model maximizes integrated representation learning. Finally, the learned features of text and speech are combined using an average ensemble method to predict the final emotional state. The proposed HyFusER model leverages the strengths of both text and speech, significantly improving the accuracy and reliability of emotion recognition. It is anticipated to make meaningful contributions to various applications, including emotion-based conversational AI, healthcare, remote education, and emotional labor support. Future research aims to expand the scope by incorporating additional modalities, such as facial images, to enhance the performance of multimodal emotion recognition technologies capable of identifying more complex and diverse emotional states. The main strengths of the proposed HyFusER model are as follows:

- Integration of Complementary Information: By employing Dual Cross Modal Attention, the model effectively combines information from text and speech, enabling recognition of complex emotional signals that are difficult to capture using a single modality.
- High Learning Efficiency: The design allows each modality to learn features within the context of the other, ensuring that critical cues essential for emotion recognition are not missed.
- Improved Prediction Accuracy and Reliability: By fully leveraging the strengths of both text and speech, the model significantly enhances the accuracy and reliability of emotion prediction.
- Applicability to Various Domains: The model demonstrates high performance in practical applications requiring emotion recognition, including emotion-based conversational AI, healthcare, remote education, and emotional labor support.

The structure of this paper is as follows. Section 2 defines and reviews the contents of previous studies related to multimodal emotion recognition. Section 3 explains the details of the technologies used in this study. Section 4 describes the experimental procedures of the proposed emotion recognition model. Section 5 evaluates the experimental results of the proposed model. Finally, Section 6 presents the conclusion of this study and discusses its significance and future research directions.

2. Related Work

This chapter covers recent trends and key insights in multimodal emotion recognition studies utilizing text and speech. Methods based on single modalities are limited in capturing subtle emotional signals due to the inherent constraints of each modality. In contrast, multimodal approaches that integrate various modalities—such as linguistic, acoustic, or visual information—can significantly improve the accuracy of emotion recognition by leveraging the complementary strengths of different data types. Consequently, recent studies in emotion recognition have actively explored multimodal approaches to more precisely identify emotional states. Among these, this paper focuses on multimodal emotion recognition that combines text and speech. The complementary interaction between linguistic context and non-verbal acoustic signals allows for the recognition of complex emotional states that are difficult to capture with a single modality.

Seunghyun Yoon's study [7] proposed a simple concatenation approach for combining text and speech modalities. Acoustic features such as MFCC, energy, and pitch were extracted from speech data, while transcription data from text was processed using the Bag-of-Words method to extract key emotion-relevant keywords. The resulting feature vectors from both modalities were concatenated and fed into a multilayer perceptron (MLP)-based model for emotion classification. This study demonstrated higher accuracy in emotion recognition compared to single-modality approaches, highlighting the complementary relationship between speech and text. Young-Jun Kim's study [8] converted speech signals into spectrograms, which were processed using CNN to extract acoustic features. For the text modality, word embeddings were generated and fed into a CNN-LSTM architecture to capture sequential information. The final outputs of the text and speech models were combined using an average ensemble method for emotion prediction. This process, which considers both temporal-frequency domain features of speech and contextual information from text, showed that more refined emotion recognition is achievable compared to simple concatenation.

H. Park's study [9] proposed an algorithm combining text-based and speech-based emotion recognition to comprehensively understand user states. Text data underwent preprocessing steps such as tokenization, neutral data removal, and label mapping, followed by embedding using KoBERT. Speech data features were extracted using MFCC and analyzed with a CNN-based model. The results from both modalities were combined using a weighted average ensemble method, reflecting the importance of each modality. This integration significantly improved emotion recognition accuracy and provided higher reliability in assessing actual user states. Sung-Sik Kim's study [10] utilized RoBERTa for sentence-level text embeddings to incorporate rich linguistic context and applied MFCC and HuBERT for deep acoustic feature extraction in the speech modality. These feature vectors were combined to construct multimodal representations, and the final emotion class was predicted. The study demonstrated superior performance compared to single-modality approaches by maximizing the interaction between text embeddings that convey diverse contextual meanings and speech embeddings that reflect speaker tone, pitch, and prosody.

Yuchul Byun's study [11] proposed Early Fusion, Late Fusion, and Hybrid Fusion methods for combining text and speech modalities in multimodal emotion recognition, comparing and analyzing their performance. Text embeddings were extracted using LLaMa2, and acoustic features were obtained using HuBERT. In the Late Fusion method, softmax probability values from modality-specific models were averaged to predict the final emotion. This method achieved over 11% higher accuracy than single-modality approaches, effectively combining information from text and speech to produce excellent results in complex emotion recognition scenarios. Jayashree Agarkhed's study [12] focused on the interaction between modalities in multimodal emotion recognition. Text embeddings were generated using a Recurrent Text Encoder (TRE), capturing sentence-level information and sequential characteristics, while speech features were learned through an Audio Recurrent Encoder (ARE). The extracted text and speech vectors were concatenated into a unified vector, which was fed into an RNN-based final network for emotion classification. This approach achieved significant improvements in emotion recognition recall and precision by finely integrating the non-verbal elements of speech with the contextual information from text.

Makhmudov's study [13] proposed a method to enhance multimodal emotion recognition by applying an attention mechanism to a BERT and CNN architecture. The text modality utilized a BERT-based model to extract linguistic features, while the speech modality employed a CNN to analyze acoustic features. An attention mechanism was then used to effectively fuse information from both modalities for emotion classification. Lin Feng's study [14] introduced a model leveraging multi-scale MFCCs and a multi-view attention mechanism. In the speech modality, rich emotional features were extracted using MFCCs at different scales. For interaction with the text modality, emotional features were comprehensively fused across four aspects: speech self-attention, text self-attention, text-guided speech attention, and speech-guided text attention. Additionally, gender recognition was incorporated as an auxiliary task to enhance learning of gender-specific emotional cues. A combined loss function, integrating softmax cross-entropy loss and center loss, further improved emotion recognition accuracy.

Jiachen Luo's study [15] proposed the Cross-Modal RoBERTa (CM-RoBERTa) model for utterance-level emotion recognition. This model dynamically captured interactions between speech and text modalities using parallel self-attention and cross-attention mechanisms, effectively learning intra-modal and inter-modal features. The use of mid-level fusion and residual connections enabled modeling of long-term contextual dependencies and modality-specific patterns, achieving superior performance. Rutherford Agbeshi Patamia's study [16] suggested a multimodal emotion recognition model employing pre-trained transformer frameworks for self-supervised learning of speech and text representations, augmented by motion capture data. Speech features were extracted using wav2vec 2.0, text features using BERT, and motion capture data added non-verbal behavioral cues. These features were fused at the feature level for emotion classification, demonstrating excellent results.

Peiying Wang's study [17] proposed a novel approach for multimodal emotion recognition by leveraging label information. Representative label embeddings were generated for both text and speech modalities, and label-token and label-frame interactions were employed to learn label-enhanced representations for each utterance. A label-based attention fusion module was designed to integrate these enriched text and speech representations, achieving superior classification performance. Zhen Wu's study [18] conducted an empirical analysis of the impact of fusion strategies on the performance of audio and text modalities in multimodal emotion recognition. The study explored various fusion methods, identifying effective approaches while preserving the unique emotional expression characteristics of each modality. To further enhance model performance, a Perspective Loss function was introduced to maintain modality-specific emotional features during fusion.

Recent multimodal emotion recognition studies have been progressing toward modeling interactions between modalities more precisely to improve the accuracy of emotion recognition. Initially, research focused on processing each modality independently or combining text and speech modalities through simple concatenation or ensemble techniques. However, these approaches have limitations in adequately capturing the unique characteristics of each modality, leading to missed critical information exchanged between modalities. In particular, failure to effectively model the information exchange occurring at intermediate layers can result in performance constraints at the final emotion classification stage. Recently, methods combining attention mechanisms with self-supervised learning have enabled more precise emotion recognition while enhancing data efficiency. Furthermore, studies leveraging label information or improving fusion mechanisms between modalities have opened new possibilities for recognizing complex emotional states. In line with these advancements, this study proposes a hybrid multimodal emotion recognition model that combines Intermediate Layer Fusion and Last Fusion using Dual Cross Modal Attention to integrate the text and speech modalities, enabling precise emotion recognition.

3. Background

3.1. Text Feature Extraction: The KoELECTRA Modal

KoELECTRA is a transformer-based pre-trained language model developed to effectively process and analyze Korean text. It is built upon ELECTRA (Enhanced Representation through Clustering) [19] and is specifically optimized to reflect the grammatical features and lexical diversity of the Korean language. Unlike traditional masking techniques employed by models like BERT, ELECTRA adopts a training strategy that involves two models: a generator and a discriminator. This approach generates fake data and performs a process where the discriminator distinguishes between real and fake tokens, leading to more refined representation learning. Specifically, during training, the generator predicts masked tokens in the input sentence, as shown in Equation (1), while the discriminator differentiates between tokens restored by the generator and actual tokens, as shown in Equation (2). As this process is repeated, the discriminator becomes increasingly adept at identifying fake tokens, and the generator learns to produce natural tokens that can deceive the discriminator. Consequently, the ELECTRA architecture facilitates a deeper understanding of the contextual meaning of text, enabling the learning of rich linguistic representations.

$$L_G = -\sum_{i=1}^{n} logp_G(x_i|\overline{x}) \tag{1}$$

$$L_D = -\sum_{i=1}^{n} [y_i logp_D(y_i|\overline{x}) + (1-y_i)\log(1-p_D(y_i|\overline{x}))] \tag{2}$$

KoELECTRA is a language model optimized to effectively process the grammatical characteristics and lexical diversity of the Korean language, building upon the architecture of ELECTRA. Its core structure, like ELECTRA, consists of two models: a Generator and a Discriminator. The model achieves refined contextual understanding by predicting masked words in text data and determining whether the predicted words match the original ones. Inheriting this structure, KoELECTRA is designed to handle the complex contextual and grammatical structures of Korean effectively. For instance, in the sentence '내가 실력이 부족한 건 맞아' ('It is true that I lack skill'), KoELECTRA masks specific words, allowing the Generator to restore the masked words based on the context, while the Discriminator assesses the authenticity of the restored words. Since this study is conducted in Korean, the examples are provided to align with Korean text embeddings. Figure 1 visually summarizes the structure and operation of the Generator and Discriminator in KoELECTRA.

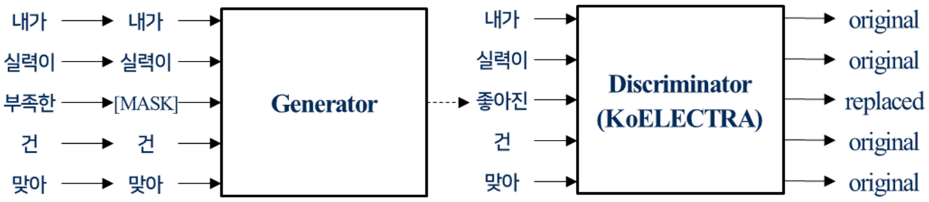

Figure 1. Structure and Operation of KoELECTRA's Generator and Discriminator.

KoELECTRA places a strong emphasis on reflecting the unique linguistic characteristics and grammatical structures of the Korean language, enabling high performance in Korean-specific tasks such as emotion recognition. As an agglutinative language, Korean exhibits variations in the meaning of words depending on context, particles, and verb endings, which directly influence the emotions conveyed in a sentence. KoELECTRA addresses these features by effectively deconstructing sentences through morpheme analysis and incorporating embeddings that capture complex particle and verb transformations, resulting in high accuracy in Korean emotion recognition. Furthermore, KoELECTRA is pre-trained on a large-scale Korean emotion dataset, allowing it to learn text patterns that represent a wide range of emotional states. This capability enables the model to accurately capture not only positive and negative emotions but also subtle emotional nuances arising from differences in expression. As a result, KoELECTRA demonstrates high accuracy and precision in analyzing emotions in Korean text and has established itself as a key tool for building Korean emotion analysis systems. In this study, the KoELECTRA Base-V3 model is utilized, and its performance is summarized in Table 1.

Table 1. KoELECTRA base model performance table.

	NSMC (acc)	**Naver NER (F1)**	**PAWS (acc)**	**KorNLI (acc)**	**KorSTS (spearman)**	**Question Pair (acc)**	**KorQuaD (Dev) (EM/F1)**	**Korean-Hate-Speech**
KoBERT	89.59	87.92	81.25	79.62	81.59	94.85	51.75/ 79.15	66.21
XLM-Roberta-Base	89.03	86.65	82.80	80.23	78.45	93.80	64.70/ 88.94	64.06
HanBERT	90.06	87.70	82.95	80.32	82.73	94.72	78.74/ 92.02	68.32
KoELECTRA-Base	90.33	87.18	81.70	80.64	82.00	93.54	60.86/ 89.28	66.09
KoELECTRA-Base-V2	89.56	87.16	80.70	80.72	82.30	94.85	84.01/ 92.40	67.45
KoELECTRA-Base-V3	**90.63**	**88.11**	**84.45**	**82.24**	**85.53**	**95.25**	**84.83/ 93.45**	**67.61**

3.2. Speech Feature Extraction: HuBERT Modal

HuBERT (Hidden Unit BERT) [20] is a self-supervised learning-based model designed to extract high-dimensional and multifaceted features from speech signals. It is specifically developed to efficiently learn both linguistic and non-linguistic information from speech representations. The core idea of HuBERT is to effectively capture the inherent structure and complex acoustic characteristics of speech signals, enabling it to learn rich representations that achieve superior performance in various downstream tasks. To achieve this, HuBERT undergoes a two-stage learning process and leverages CNNs and transformer encoders to extract both low-level and high-level features from input speech. Figure 2 illustrates the architecture of the HuBERT model.

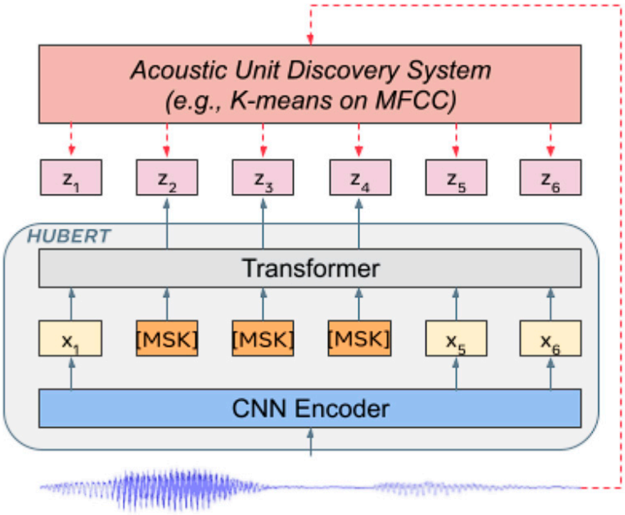

Figure 2. HuBERT model structure.

The first stage, Offline Clustering, involves dividing speech data into multiple clusters based on acoustic similarity using unsupervised learning techniques. The K-Means algorithm is commonly employed for this process, assigning segments with similar acoustic features to the same cluster, thereby automatically learning representative 'acoustic units (labels)' for each segment. These cluster labels are later used as learning targets during the masked prediction stage. The second stage, Masked Prediction Learning, is the core mechanism of HuBERT training. After setting the acoustic unit labels obtained through offline clustering as the learning targets, specific regions of the input speech data are intentionally masked, making them inaccessible to the model. The model is then trained to infer the cluster labels of the masked regions using only the information from the unmasked segments. This process is mathematically represented by Equation (3). Through this learning process, the model acquires the ability to reconstruct missing parts based on partial information, ultimately gaining a deeper understanding of the structural characteristics and contextual relationships within speech signals.

$$L_m(f; X,\ M,\ Z) = \sum_{t \in M} log p_f(z_t | \hat{X},\ t) \tag{3}$$

In terms of feature extraction, HuBERT employs a multi-stage approach. First, raw audio signals are transformed into low-level acoustic features using a Convolutional Neural Network (CNN). The CNN effectively captures local patterns along the time and frequency axes (e.g., energy distribution, frequency variations, and intensity of articulation), generating initial feature vectors. Next, a transformer-based encoder processes these low-level features, learning to infer high-level characteristics such as utterance context, speaker intent, emotion, and style. By leveraging multi-head attention mechanisms, the transformer encoder captures long-term dependencies and reconstructs rich representation vectors that reflect the interactions between various elements within the speech data. This learning strategy enables HuBERT to achieve high performance across a range of speech-based applications, including speech recognition, speaker classification, and emotion analysis. The acoustic units automatically obtained during the offline clustering stage serve as robust targets for self-supervised learning. During the masked prediction stage, the model develops the ability to infer the full characteristics of speech signals from partial information. As a result, HuBERT internalizes the multidimensional and complex features within speech data, making it an exceptionally useful tool for various downstream tasks related to speech processing.

3.3. Cross Modal Attention in Multimodal Environments

In a multimodal environment, effectively modeling the interactions between heterogeneous modalities such as text, speech, and images requires a structure that not only preserves the unique information of each modality but also facilitates their meaningful integration. One prominent technique for achieving this is Cross Modal Attention. This method uses 'query information' extracted from one modality (e.g., text) to combine the features (key and value) generated by another modality (e.g., speech), enabling the model to learn complex interactions and semantic structures that are difficult to capture with a single modality [21,22]. For instance, in learning the relationships between text and speech, text serves as the 'query' specifying what information is being sought, while detailed and non-verbal features from speech (e.g., intonation, emotion, pitch variations) act as the key and value. This enables a semantic fusion that organically integrates text and speech. However, in traditional structures where the query is fixed to text and the key and value are sourced from speech, a text-dominant bias can arise. This configuration, while advantageous for its fast learning and consistency, risks underrepresenting critical non-verbal information from speech in tasks like emotion recognition. Ideally, in a multimodal environment, text and speech should interact bidirectionally, with speech complementing text and text interpreting or realigning speech. Unfortunately, in a structure where text is always the query, speech is limited in its ability to rearrange or adjust textual representations actively. To address this bias, this paper proposes a Hybrid Multimodal Transformer structure that enables text and speech to play equal roles. By mitigating text-dominant bias, this approach ensures richer incorporation of non-verbal information embedded in speech, facilitating true cross modal learning and delivering more balanced and comprehensive multimodal integration.

4. Hybrid Multimodal Transformer for Emotion Recognition

4.1. Overall Process

This paper proposes the HyFusER (Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention) model. The proposed model is explained in three stages. In the Feature Extraction stage, text and speech data are transformed into feature vectors using KoELECTRA and HuBERT, respectively. In the Intermediate Layer Fusion stage, text and speech data are processed through separate transformer encoders, followed by complementary learning between the two modalities using Dual Cross Modal Attention. Finally, in the Last Fusion stage, the results learned in the previous stage are combined using average ensemble to recognize the final emotion. Figure 3 illustrates the overall structure of the proposed method.

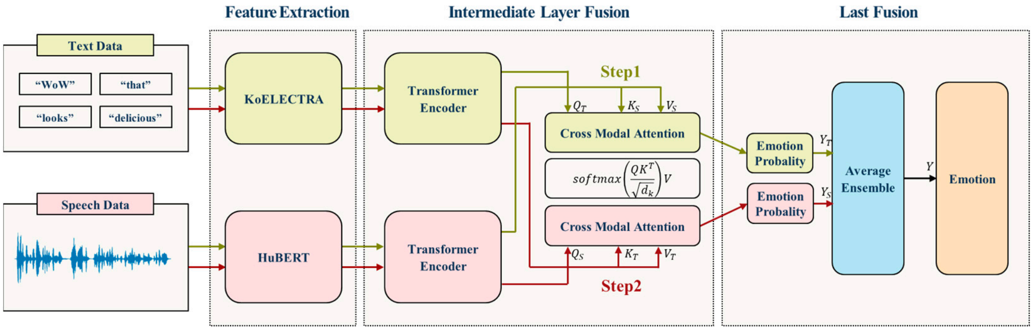

Figure 3. Research structure.

4.2. Extract Text and Speech Features

This section includes the preprocessing steps for embedding text and speech data to effectively learn the features of different modalities. This process enhances interaction by transforming both modalities to have the same dimensions and categories.

To accurately capture the grammatical and semantic context of the Korean language, text data are embedded using the KoELECTRA base-v3 model. KoELECTRA is a pre-trained language model specialized for Korean, transforming each word within a sentence into a 768-dimensional vector. Consequently, the embedded output for a given sentence is a vector with dimensions corresponding to the number of tokens in the sentence and the embedding dimension of KoELECTRA (768). For instance, a sentence with 10 tokens is converted to a (10,768) vector, while one with 30 tokens becomes a (30,768) vector. To train deep learning models, all sentences must have consistent dimensions. First, the maximum number of tokens across the dataset is determined. The text data are then tokenized using KoELECTRA’s tokenizer, and each sentence is padded to match the maximum token count. Padding involves adding [PAD] tokens to sentences with fewer tokens, ensuring uniform lengths. The maximum token count identified is 126, and all sentences are normalized to this length. As a result, the final text data are transformed into fixed-dimensional vectors of size (number of samples, 126,768), where 126 represents the number of tokens and 768 is the embedding dimension of KoELECTRA. This preprocessing ensures that the text data have a consistent input format for deep learning model training.

Speech data are embedded using the HuBERT ls960 model. HuBERT learns temporal and spectral characteristics of speech, generating refined speech representation vectors. The vectors extracted by HuBERT vary in dimensions depending on the temporal length of the speech data. To input both text and speech data into the same deep learning model, the embedding sizes of the two modalities must match. To achieve this, linear interpolation is applied, where data with insufficient length are expanded by filling data points between existing ones, and data with excess length are compressed by reducing intervals between data points. Consequently, the speech data are reshaped to match the dimensions of the text data, resulting in vectors of size (number of samples, 126,768). Through this process, both text and speech data are aligned to identical dimensions, ensuring compatibility for model training.

To accurately learn the interaction between text and speech, a scaling process is required to standardize the categories of the two data modalities. The proposed method employs Standard Scaling, which adjusts the mean to 0 and variance to 1. To verify the consistency of the three-dimensional modalities, the mean is calculated along the first dimension (number of samples N) and the second dimension (sequence length T). This calculation summarizes the data into a one-dimensional array for each feature dimension D in the third axis. The summarized results are then analyzed and visualized to extract meaningful information. The process of this calculation is expressed by Equation (4) in the paper. This ensures that the features from both text and speech are normalized, enabling precise and balanced learning of cross modal interactions.

$$\overline{X} = \frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T} X_{nt} \tag{4}$$

The analysis of each dataset reveals that the text data range from -6.4 to 9.97, while the speech data range from -1.91 to 0.94. If Standard Scaling is applied independently to each modality, the data will be adjusted based on their respective means and standard deviations, resulting in different ranges even after scaling. To address this issue, the proposed method merges the two datasets into a single combined set and then applies Standard Scaling as described in Equation (5). This ensures that both text and speech data are scaled uniformly, aligning their ranges and enabling effective cross modal learning.

$$\begin{gathered} X = \text{concat}(X_T,\ X_S) \\ S = \tfrac{(X-\mu)}{\sigma} \end{gathered} \tag{5}$$

Using this method, the two datasets share the same mean and standard deviation, resulting in standardized data with identical ranges and statistical properties. Consequently, both text and speech data are scaled to the range of -1 to 1. Finally, for input into the deep learning model, the text and speech data are separated again as described in Equation (6). The visualization results of each process are presented in Figure 4, illustrating the alignment and consistency achieved through this scaling approach. This ensures the model receives uniformly processed data, facilitating effective joint learning.

$$\begin{gathered} X'_T = S[:,\ 0 : m] \\ X'_S = S[:,\ m : n] \end{gathered} \tag{6}$$

The feature-extracted text and speech data are transformed into vectors of shape (number of samples, 126,768) and then passed into the transformer encoder within the proposed model. The text data, embedded with features learned through KoELECTRA, and the speech data, embedded with features learned through HuBERT, are individually processed by separate transformer encoders. Subsequently, the Dual Cross Modal Attention mechanism facilitates the learning of interactions between the two modalities, enabling effective integration of their complementary features.

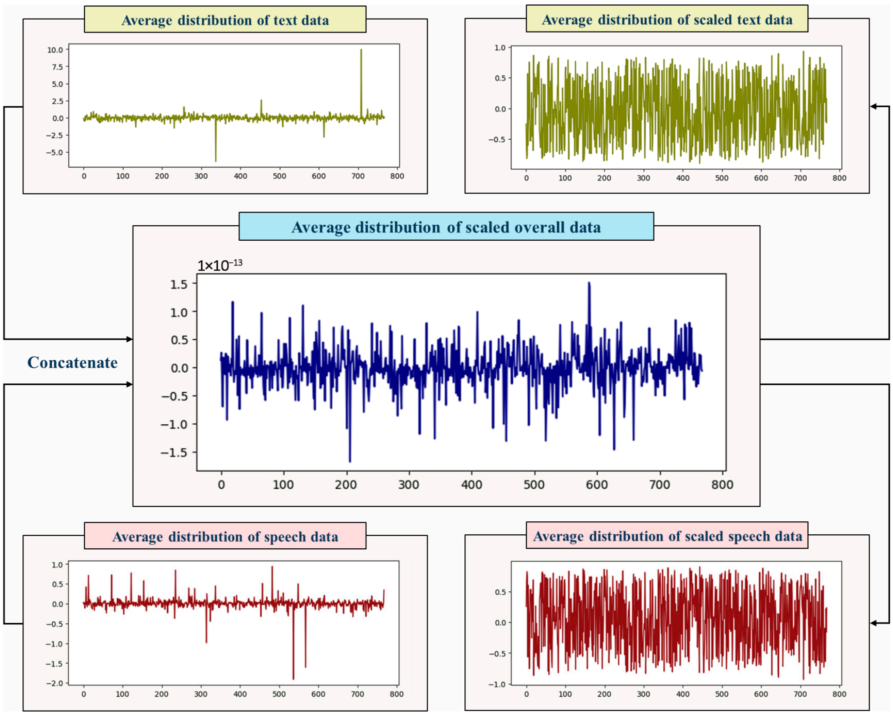

Figure 4. Each modality Standard Scaling process.

4.3. Hybrid Multimodal Transformer for Emotion Recognition Model

The objective of this section is to explain the process of recognizing emotions by effectively learning the features of different modalities, namely, text and speech data. The Multimodal Transformer utilizes Intermediate Layer Fusion and Last Fusion to optimize the interactions between text and speech, ultimately enhancing emotion recognition performance. The proposed model is a multimodal deep learning framework based on text and speech data. Through feature learning and integration at each stage, it overcomes the limitations of existing models and improves accuracy. By leveraging these fusion mechanisms, the model achieves a more refined understanding of cross modal features, contributing to superior emotion recognition capabilities.

4.3.1. Intermediate Layer Fusion

The embedded text and speech data are each passed through a transformer encoder, which is used to learn the contextual and structural features of the input data. The transformer encoder used in this study consists of three layers, with each layer comprising Multi-Head Attention and a Position-Wise Feed Forward Network. Multi-Head Attention performs multiple attention heads in parallel, learning the input data from various perspectives. This reduces dependency on specific parts of the data and enables the learning of more comprehensive features. The Position-Wise Feed Forward Network applies a non-linear transformation to the attention output to learn more complex patterns, which is applied independently to each sequence element. Through this process, the text and speech data maintain their unique characteristics while being effectively learned. The structure of the transformer encoder is shown in Figure 5.

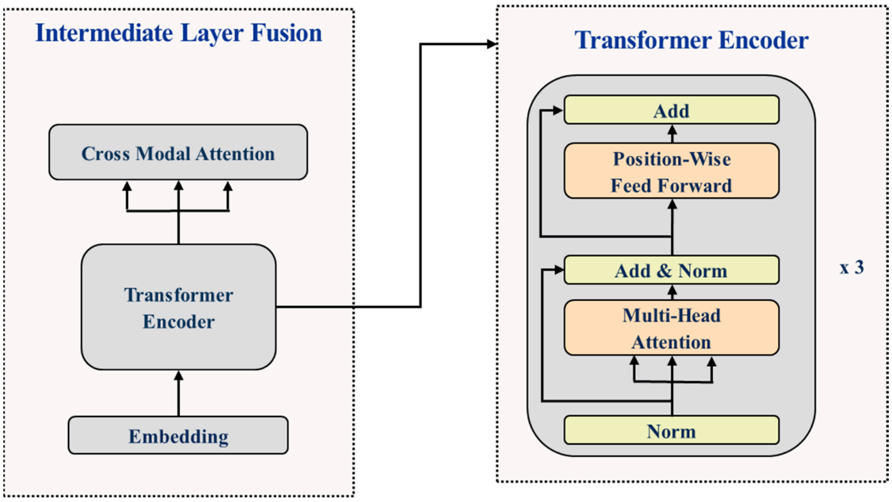

Figure 5. Transformer encoder structure.

Cross Modal Attention is utilized to learn the interactions between text and speech, extracting the most significant information from both modalities simultaneously. However, the learning direction in Cross Modal Attention depends on the configuration of Query (Q), Key (K), and Value (V). For example, if text is set as the Query and speech as the Key and Value, the model learns to interpret speech based on text information. This unidirectional learning approach may limit the complementary learning between text and speech. To address this limitation, a Dual Cross Modal Attention learning strategy is introduced, dividing the learning process into Step 1 and Step 2 to enable bidirectional and more effective cross modal learning.

1. Step 1: In this step, the text embeddings are set as the Query, and the speech embeddings are set as the Key and Value. The learning process, as described in Equation (7), complements the features of speech based on the contextual information from the text. This approach enhances emotion recognition performance by focusing on text-centric features.

$$\text{CrossModalAttention}(Q_T, K_S, V_S) = \text{softmax}(\frac{Q_T K_S^T}{\sqrt{d_k}})V_S \tag{7}$$

2. Step 2: In this step, the speech embeddings are set as the Query, and the text embeddings are set as the Key and Value. The learning process, as described in Equation (8), leverages the temporal and acoustic characteristics of speech to learn the semantic information from text. This approach improves emotion recognition performance by focusing on speech-centric features.

$$\text{CrossModalAttention}(Q_S, K_T, V_T) = \text{softmax}(\frac{Q_S K_T^T}{\sqrt{d_k}})V_T \tag{8}$$

The results learned through Step 1 and Step 2 are reflected as two distinct probability distributions: a text-centric probability distribution that primarily captures the nuances of emotions perceived through text, and a speech-centric probability distribution that embodies the unique characteristics of speech, such as tone, pitch, and intensity. Each of these probability distributions represents the unique perspectives of text and speech, acting complementarily to enhance emotion recognition performance. These distributions serve as key inputs for the subsequent stages, leveraging the strengths of both modalities to achieve a more comprehensive understanding of emotions.

4.3.2. Last Fusion

In the Last Fusion stage, the two outputs derived from the Intermediate Layer Fusion are combined to predict the final emotion. This stage aggregates the characteristic information from the text-centric probability distribution and the speech-centric probability distribution using a GlobalAveragePooling1D Layer and a Dense Layer, subsequently predicting emotions as probabilities. The final emotion is determined through an average ensemble of these probabilities. The GlobalAveragePooling1D Layer compresses the sequence information of each modality, extracting global features. The Dense Layer uses the softmax function to calculate probabilities for each emotion. This structure preserves the unique characteristics of each data type while providing suitable vector representations for emotion classification. The detailed architecture of the Last Fusion stage can be seen in Figure 6.

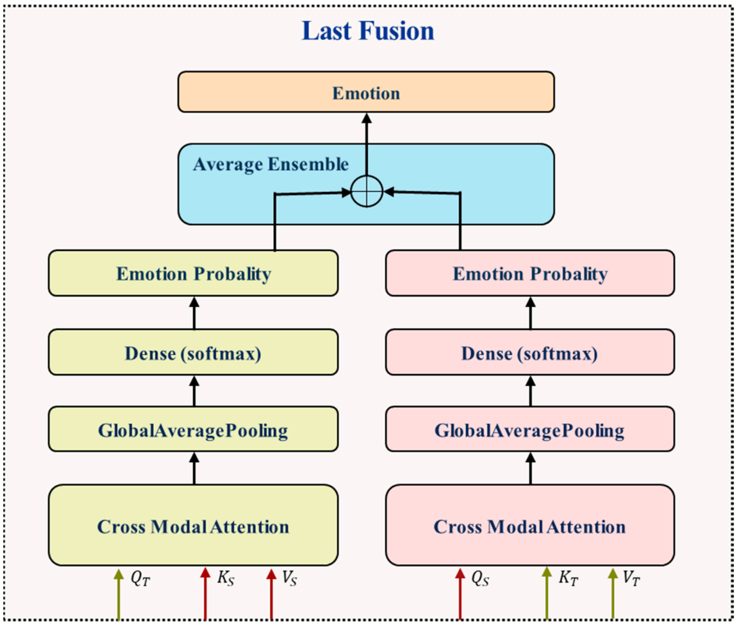

Figure 6. Last Fusion structure.

The average ensemble is used to merge the emotion predictions, calculated as probabilities, to produce more refined emotion predictions. This process integrates the learning results of text and speech data with equal weights, reflecting the strengths of both data modalities while minimizing biases from individual learning. The average ensemble is defined in Equation (9), which represents the process of generating the final prediction by taking the simple average of the emotion probability distributions derived from each modality. This approach ensures balanced contributions from both text and speech, leading to a more robust and accurate emotion recognition outcome.

$$Y = \operatorname{argmax}(\frac{Y_Y + Y_S}{2}) \tag{9}$$

This ensemble strategy is particularly effective in addressing discrepancies that may arise from the inherent characteristics of text and speech data. For example, text data often excel in capturing the semantic meaning of emotions, while speech data provide valuable paralinguistic cues such as tone and pitch. By equally weighting these two modalities, the model avoids over-reliance on either modality, ensuring a more holistic understanding of emotional expressions. Furthermore, this method enhances the generalization capability of the model, as it leverages complementary information from both modalities rather than treating them independently. This approach enhances the prediction consistency across various emotion classes and improves the model's stability. For instance, if the prediction probabilities for a specific emotion are 0.85 and 0.80 from text and speech, respectively, the average ensemble calculates the final prediction probability as (0.85 + 0.80)/2 = 0.825. Among the calculated probabilities for each emotion, the highest probability determines the final emotion. This simple yet effective technique not only smoothens prediction fluctuations across modalities but also ensures that less confident predictions from one modality are supplemented by the strengths of the other. Table 2 provides an example of the prediction results for a sentence. This method more accurately reflects the interaction between the two modalities, thereby improving the reliability and accuracy of emotion recognition. As such, the Last Fusion stage plays a pivotal role in synthesizing the outcomes of multimodal learning, maximizing the complementary learning between text and speech. For example, in cases where a speech signal may lack clarity due to noise, the text-based analysis can compensate by providing accurate semantic context, and vice versa. This cross modal interaction highlights the model's ability to handle real-world data variability and ensures that the final emotion prediction is both consistent and robust.

Table 2. Final emotion prediction example.

Emotion	Step 1 Model	Step 2 Model	Average Ensemble	Final Emotion
Angry	0.0110	0.0052	0.0081	neutral
Disgust	0.0012	0.0015	0.0013	
Fear	0.0079	0.0111	0.0095	
Happy	0.2358	0.3191	0.2775	
Neutral	0.5105	0.625	0.5677	
Sad	0.0196	0.0088	0.0142	
Surprise	0.2137	0.0290	0.1213	

Algorithm 1 explains the entire process of the proposed model.

Algorithm 1. Hybrid Multimodal Transformer Emotion Recognition Model

```
Input: Text data $D_T$, Speech data $D_S$
Output: Emotion Classification $Y$
//Modal Definition
$\mathrm{CrossModalAttention}(Q,\ K,\ V) : \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, $\mathrm{GlobalAveragePooling1D} : \frac{1}{T}\sum_{t=1}^{T} H_t$
//Feature Extraction
$X_T = \mathrm{KoELECTRA}(D_T)$, $X_S = \mathrm{HuBERT}(D_S)$
$\mathrm{X} = \mathrm{concat}(X_T,\ X_S)$, $\mathrm{S} = \frac{(X-\mu)}{\sigma}$
$X'_T = S[:,\ 0 : m]$, $X'_S = S[:,\ m : n]$
//Step 1: Intermediate Layer Fusion
$E_T = \mathrm{TransformerEncoder}(X'_T)$, $E_S = \mathrm{TransformerEncoder}(X'_S)$
$Q_T = E_T$, $K_S = E_S$, $V_S = E_S$, $H_T = \mathrm{CrossModalAttention}(Q_T,\ K_S,\ V_S)$
$G_T = \mathrm{GlobalAveragePooling1D}(H_T)$
$Y_T = \mathrm{softmax}(W_T G_T + b_T)$
//Step 2: Intermediate Layer Fusion
$E_T = \mathrm{TransformerEncoder}(X'_T)$, $E_S = \mathrm{TransformerEncoder}(X'_S)$
$Q_S = E_S$, $K_T = E_T$, $V_T = E_T$, $H_S = \mathrm{CrossModalAttention}(Q_S,\ K_T,\ V_T)$
$G_S = \mathrm{GlobalAveragePooling1D}(H_S)$
$Y_S = \mathrm{softmax}(W_S G_S + b_S)$
//Last Fusion
$Y = \mathrm{argmax}\left(\frac{Y_T + Y_S}{2}\right)$
```

4.3.3. Hyperparameter Settings

The proposed multimodal emotion recognition model was trained with a well-structured configuration to ensure robust performance and strong generalization capabilities. The Adam optimizer, with a learning rate of 0.0001 chosen through empirical evaluation, balanced convergence speed and model stability. A batch size of 32 ensured efficient training while maintaining the ability to capture representative gradients for parameter updates. The model was trained for a maximum of 30 epochs, with early stopping applied to prevent overfitting by halting training when validation accuracy failed to improve for three consecutive epochs. This approach reduced unnecessary computations while preserving generalization performance. Additionally, a ModelCheckpoint mechanism saved the model weights corresponding to the best validation accuracy, ensuring the optimal model was preserved for subsequent evaluations. Activation functions played a pivotal role in capturing the nonlinear relationships inherent in multimodal data. The ReLU activation function, applied within the feed-forward layers of the transformer encoder and CrossModalAttention components, introduced nonlinearity and facilitated the learning of complex patterns. In the output layer, a softmax activation function computed class probabilities for the seven emotion categories. These probabilities from both modalities were combined through an ensemble averaging approach, leveraging complementary information to produce the final emotion recognition outputs.

To stabilize training and mitigate overfitting, regularization techniques were integrated. Layer normalization, with an epsilon value of 0.000001, was applied after attention and feed-forward computations within the transformer encoder, ensuring numerical stability and consistent activation magnitudes. This contributed to improved training dynamics and performance reliability. The Sparse Categorical Crossentropy loss function was employed, as the emotion labels were represented as integers rather than one-hot encodings, aligning effectively with the dataset's multiclass classification structure. Architecturally, each modality—audio and text—was processed through three transformer encoder layers, enabling the extraction and refinement of hierarchical feature representations. A single attention head in each encoder simplified the model design while maintaining the capacity to capture essential cross modal interactions. The feed-forward network dimension was set to 768, providing sufficient capacity to represent rich multimodal features without incurring excessive computational overhead.

In summary, the proposed training configuration and architectural design collectively optimized the model's ability to recognize emotions from multimodal inputs. The thoughtful selection of hyperparameters, combined with robust regularization and optimization strategies, ensured the model achieved both efficiency and state-of-the-art performance. These results underscore the effectiveness of the proposed approach and provide a strong foundation for future extensions and real-world applications in multimodal emotion recognition. The detailed configuration is summarized in Table 3.

Table 3. Hyperparameter details.

Parameter	Value
Input Shape	(126, 768)
Feed-Forward Dimension	768
Number of Layers	3
Number of Attention Heads	1
Layer Normalization Epsilon	0.000001
Activation Function	ReLU
Optimizer	Adam (Learning Rate 0.0001)
Early Stopping Monitor	Val Sparse Categorical Accuracy
Early Stopping Patience	3

Table 3. *Cont.*

Parameter	Value
Loss	Sparse Categorical Crossentropy
Model Checkpoint Monitor	Val Sparse Categorical Accuracy
Batch Size	32
Epochs	30

5. Experimental Evaluation

5.1. Dataset

The data used in this study include the Korean multimodal emotion datasets KEMDy19 [23] and KEMDy20 [24], provided by the Electronics and Telecommunications Research Institute (ETRI). The KEMDy19 dataset was collected from 40 Korean voice actors, where pairs of participants participated in one session per pair to collect data. Similarly, KEMDy20 was collected from 80 adults aged 19–39 fluent in Korean speech, with pairs participating in one session each to collect data. The collected datasets include various modalities such as speech utterances, contextual meaning of the utterances, and physiological signals (e.g., skin conductance, pulse intervals, wrist skin temperature). However, this study only used text and speech data.

Upon analyzing the KEMDy20 dataset by emotion category, it was observed that there were 11,120 samples for the neutral emotion, while disgust and fear emotions had only 61 and 43 samples, respectively. This imbalance in emotion data poses challenges for accurate analysis. To address this issue, the KEMDy19 dataset was additionally used to augment the dataset. After merging the data, duplicate text entries and sentences with fewer than two syllables were removed. Furthermore, to resolve the imbalance among emotion categories, the number of neutral emotion samples was reduced to match the count of the emotion category with the highest number of samples.

The total number of samples for each emotion category is summarized in Table 4. To further evaluate the reliability of the proposed model, the Korean emotion recognition dataset provided by AI-HUB [25] was used, following the same preprocessing procedures.

Table 4. KEMDy data number by emotion.

Emotion	Collected Data	Final Data Used
Angry	1352	1352
Disgust	446	446
Fear	840	840
Happy	2643	2643
Neutral	15,082	2643
Sad	1028	1028
Surprise	1429	1429
Total	**22,820**	**10,381**

5.2. Experimental Results

In this section, the proposed HyFusER model was evaluated using 10,381 text and speech data samples from the KEMDy dataset. The dataset was split into an 8:2 ratio for training and validation, with a primary focus on assessing the model's performance using text data. The performance evaluation was conducted using metrics such as Accuracy, Recall, Precision, and F1-Score. To address the imbalance in the number of samples across classes, this study utilized a weighted average approach. This method calculates the average performance metric by weighting the metric of each class by its sample size and dividing by the total number of samples. The formula is presented in Equation (10), where w_i denotes the sample size of the i-th class, and m_i represents the performance metric of the i-th class. This approach ensures that the evaluation of model performance fairly accounts for the class imbalance within the dataset, allowing each class to contribute equitably to the overall assessment.

$$\text{WeightedAverage} = \frac{\sum_{i=1}^{n}(w_i \times m_i)}{\sum_{i=1}^{n} w_i} \tag{10}$$

To evaluate the performance of the proposed model, comparative experiments were conducted with existing emotion recognition methodologies. The comparative models included an LSTM model for text data; a CNN model for speech data; and several multimodal approaches: an average ensemble model that predicts the final emotion by averaging the results of single-modality models, a model utilizing Bidirectional Cross Modal Attention, and the KoHMT [26] model. The experimental results are summarized in Table 5. The results show that single-modality models demonstrated relatively lower performance. In contrast, multimodal models, particularly the proposed HyFusER model, achieved superior performance by leveraging the complementary learning between text and speech data. These results validate that the hybrid approach, combining Intermediate Layer Fusion and Last Fusion, is highly effective for emotion recognition. This approach provides critical insights and contributes to advancing the field of emotion recognition.

Table 5. Performance evaluation for each model.

Model	Accuracy (%)	Recall (Weighted)	Precision (Weighted)	F1-Score (Weighted)
LSTM (Text)	59.22	0.5922	0.6278	0.6046
CNN (Speech)	51.85	0.5185	0.6243	0.5584
Single modality average ensemble	60.85	0.6085	0.6660	0.6292
Single modality weighted average ensemble	61.00	0.6100	0.6638	0.6292
Bidirectional Cross Modal Attention	64.08	0.6408	0.6604	0.6419
KoHMT	65.62	0.6562	0.6739	0.6628
HyFusER	**67.83**	**0.6783**	**0.6890**	**0.6823**

To verify the reliability and generalization capability of the proposed HyFusER model, the Korean Emotion Recognition Dataset provided by AI-HUB, composed of text and speech, was additionally utilized. This dataset enabled the evaluation of the generalizability of the results obtained from the KEMDy dataset and allowed the measurement of model performance when applied to data from a different source. The experimental setup for the AI-HUB dataset was consistent with that used for the KEMDy dataset. The data were split into training and validation sets in an 8:2 ratio, and all comparative models, including the proposed HyFusER model, were trained under identical conditions. This ensured a fair performance comparison across the two datasets. Performance evaluation was conducted using metrics such as Accuracy, Recall, Precision, and F1-Score, with results presented in Table 6. Each metric was calculated as a weighted average to account for class imbalance. Notably, by comparing the performance of the proposed model with that of single-modality models and other multimodal models, the superiority and robustness of the HyFusER model were demonstrated. These results highlight its effectiveness in emotion recognition across datasets with different characteristics.

The Confusion Matrix analysis for evaluating the emotion recognition performance of the proposed model clearly demonstrates how the model classifies each emotional state. Figure 7 visually presents the accurate classifications and errors made by the model in recognizing seven primary emotional states: ‘angry’, ‘disgust’, ‘fear’, ‘happy’, ‘neutral’, 'sad', and 'surprise'. The diagonal elements of the Confusion Matrix represent the number of correctly classified instances for each emotion. For example, the model correctly classified the 'angry' emotion 154 times and the 'happy' emotion 410 times. These numbers indicate that the model shows relatively high accuracy in recognizing these emotions. In particular, the 'happy' and 'neutral' emotions were the most accurately recognized, possibly due to a sufficient amount of data for these emotions compared to others or because the model effectively learned the features of these two emotions.

Table 6. Comparative experimental performance evaluation.

Model	Accuracy (%)	Recall (Weighted)	Precision (Weighted)	F1-Score (Weighted)
LSTM (Text)	70.38	0.7038	0.7086	0.7039
CNN (Speech)	56.38	0.5638	0.5958	0.5703
Single modality average ensemble	70.52	0.7052	0.7232	0.7075
Single modality weighted average ensemble	71.52	0.7152	0.7247	0.7164
Bidirectional Cross Modal Attention	77.63	0.7763	0.7827	0.7770
KoHMT	77.45	0.7745	0.7780	0.7744
HyFusER	**79.77**	**0.7977**	**0.7975**	**0.7975**

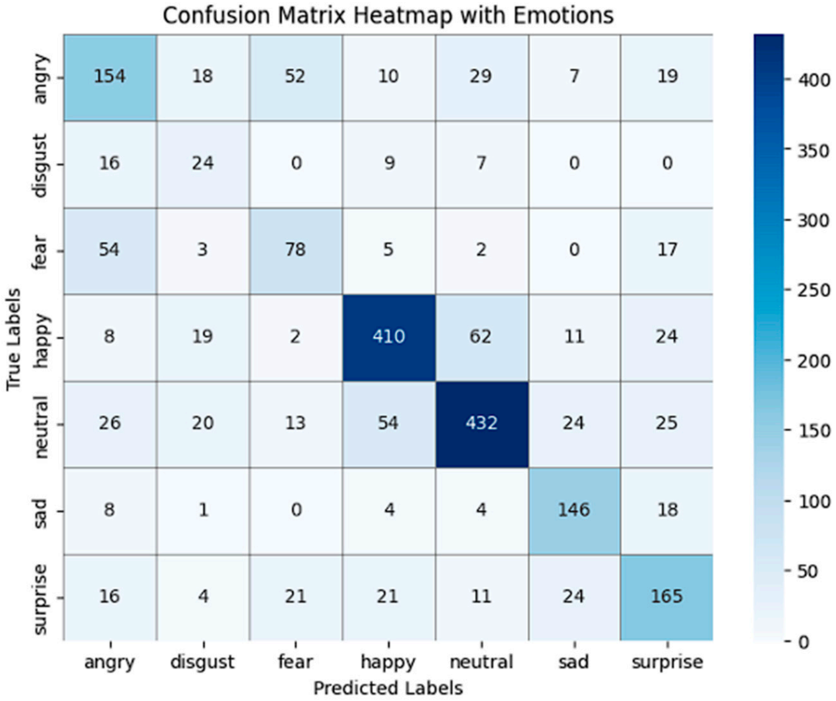

Figure 7. Proposed model Confusion Matrix.

On the other hand, the off-diagonal elements represent the number of misclassified instances for each emotion. For example, 'angry' was misclassified as 'disgust' 18 times, and 'happy' was misclassified as 'surprise' 24 times. These misclassifications suggest that certain emotions have overlapping characteristics or that the model has not sufficiently learned the distinguishing features between some emotions. Additionally, emotions such as 'disgust' and 'fear' showed relatively lower numbers of correct classifications, with only 24 and 78 correctly classified instances, respectively. Furthermore, 'disgust' was misclassified as 'angry' 16 times, and 'fear' was misclassified as 'angry' 54 times. This indicates that ‘disgust’ and ‘fear’ are either less distinct in their expression or that the model has not effectively learned the features necessary to differentiate these emotions. The high confusion between ‘fear’ and ‘angry’ emotions, with 54 instances of ‘fear’ being misclassified as ‘angry’, highlights the overlapping characteristics of these emotions in vocal and textual features.

For instance, trembling or heightened voices associated with ‘fear’ may be misinterpreted as the aggressive tone characteristic of ‘angry’. This overlap demonstrates that the model struggles to distinguish paralinguistic features such as tone modulation or energy levels, which often blur the line between these emotions. Additionally, ‘fear’ appears to be underrepresented compared to ‘angry’ in the dataset, leading to biased feature learning. As a result, the model may prioritize the dominant ‘angry’ emotion features, inadvertently reducing its ability to recognize ‘fear’ accurately. Strengthening spectral analysis can help the model better capture subtle differences in vocal patterns. For example, trembling vocal cords associated with ‘fear’ might display distinct frequency dynamics compared to the sharp pitch shifts of ‘angry’, allowing the model to separate overlapping paralinguistic cues more effectively. Furthermore, utilizing data augmentation methods to generate synthetic data that mimics the vocal and textual characteristics of ‘fear’ can mitigate dataset imbalance and provide a more comprehensive representation of underrepresented emotions. By implementing these strategies, the model’s ability to differentiate between similar emotions like ‘fear’ and ‘angry’ can be significantly improved, reducing misclassifications and enhancing the overall robustness of the emotion recognition system. The Confusion Matrix reveals lower classification accuracy for emotions like ‘disgust’ and ‘fear’, with frequent misclassifications such as ‘fear’ being mistaken for ‘angry’. These errors highlight challenges in distinguishing subtle emotional expressions and the limitations of imbalanced datasets. While emotions like ‘happy’ and ‘neutral’ benefit from sufficient data representation, underrepresented emotions like ‘fear’ and ‘disgust’ suffer from biased learning. Addressing these issues requires enhancing spectral analysis to better capture subtle vocal differences and applying data augmentation methods to generate synthetic examples of underrepresented emotions. These improvements can reduce misclassifications, balance learning across emotions, and enhance the overall robustness and accuracy of the emotion recognition system.

Figure 8 provides a comprehensive bar chart visualizing the performance of each multimodal emotion recognition model across four key metrics: Accuracy, Recall, Precision, and F1-Score. The performance values for these four key metrics were derived from Table 5. The models selected for comparison include the average ensemble of single models, the weighted average ensemble of single models, the Bidirectional Cross Modal Attention model, the KoHMT model, and the proposed HyFusER model. Each model is evaluated based on these performance metrics, enabling a direct comparison of their effectiveness. Notably, the HyFusER model achieved the highest scores across all key performance metrics, clearly outperforming existing multimodal models by a significant margin. The success of the HyFusER model is attributed to the effective fusion of text and speech data, advanced feature extraction techniques, and the application of efficient learning mechanisms. These results demonstrate that the HyFusER model represents a substantial advancement in the field of emotion recognition, suggesting its potential applicability in diverse domains such as emotion-based conversational AI, healthcare, and remote education. In conclusion, the performance comparison of multimodal emotion recognition models reaffirms the effectiveness and superiority of multimodal approaches in recognizing and interpreting complex emotional states. This analysis provides essential foundational data for the future development and improvement of emotion recognition technologies.

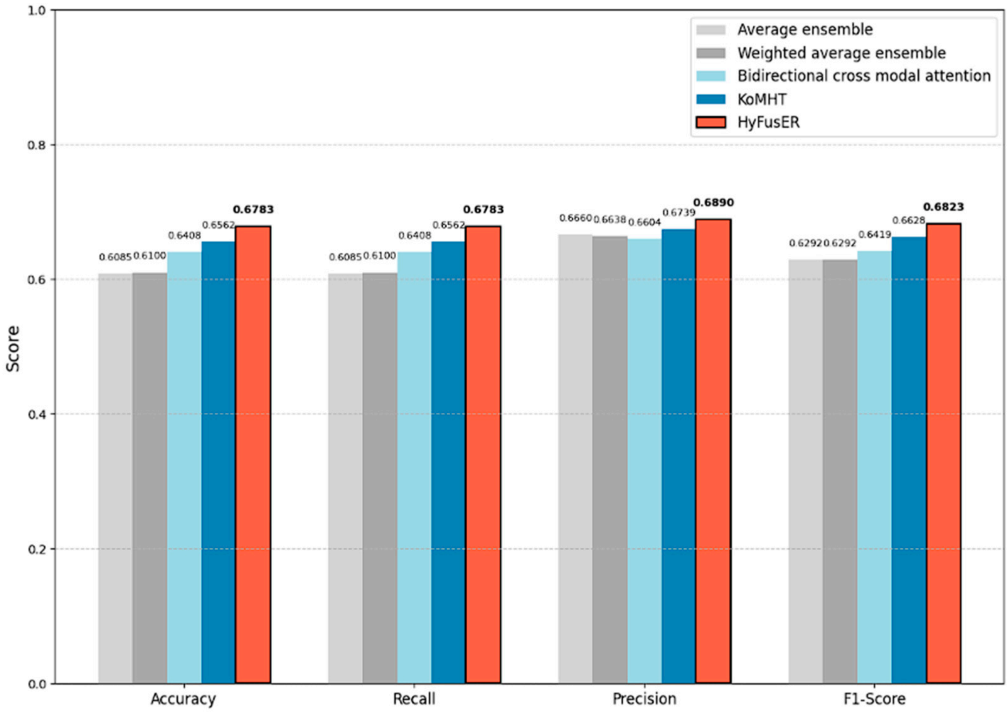

Figure 8. Comparison of performance metrics across models.

6. Conclusions

This paper proposes HyFusER: Hybrid Multimodal Transformer Emotion Recognition Model Combining Intermediate Layer and Last Fusion, a novel emotion recognition model. Moving beyond traditional approaches that process text and speech data independently, the model introduces an innovative hybrid fusion strategy that combines Intermediate Layer Fusion and Last Fusion to deeply and effectively integrate information from both modalities.

At the core of the HyFusER model, features extracted from each data modality are first processed through their respective transformer encoders. These features are then integrated during the Intermediate Layer Fusion stage using the Cross Modal Attention mechanism, which enables complementary learning of information between text and speech. This stage maximizes the strengths of both modalities through their interaction, allowing for more precise emotion recognition.

In the Last Fusion stage, the outcomes of the intermediate stage are combined using an Average Ensemble approach to produce the final emotion recognition results. The combination of these two fusion stages enhances the model's performance and improves the accuracy of emotion recognition. Experimental results demonstrate that the HyFusER model outperforms existing multimodal models.

This superior performance validates the effectiveness of the proposed hybrid fusion approach in integrating information from both modalities and minimizing potential information loss during the emotion recognition process, enabling more accurate predictions.

Future research will expand this model by incorporating additional modalities, such as facial expression recognition, to enable a deeper understanding of complex emotional states. Moreover, it will explore the use of diverse datasets, covering variations in language, cultural contexts, and emotional expressions, to enhance the model's robustness and applicability.

To achieve more nuanced and precise emotion detection, particularly in interactive AI systems, the integration of facial video datasets alongside text and speech data will be prioritized. Furthermore, research efforts will focus on ensuring the model's practical usability by developing techniques for robust emotion recognition in challenging real-world scenarios, such as noisy or dynamic environments.

Author Contributions: Conceptualization, M.-H.Y.; methodology, M.-H.Y. and K.-C.K.; software, M.-H.Y.; validation, J.-H.S.; resources, M.-H.Y. and K.-C.K.; writing—original draft preparation, M.-H.Y.; writing—review and editing, M.-H.Y. and K.-C.K. and J.-H.S.; supervision, J.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C1006149), and this study was supported by a research fund from Chosun University, 2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article.

Acknowledgments: The authors would like to acknowledge and thank all reviewers for their constructive and helpful reviews.

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

References

1. Choi, S.Y. A Development Study of Instructional Design Principles for Multi-Sensory Distance Learning in Elementary School. Master's Thesis, Seoul National University, Seoul, Republic of Korea, February 2022.
2. Lee, O.; Yoo, M.; Kim, D. Changes of Teachers' Perception after Online Distance Learning Experience Due to the COVID-19 Pandemic. *J. Educ. Technol.* **2021**, *37*, 429–458. [CrossRef]
3. Sim, J.-Y.; Seo, H.-G. Remote Medical Smart Healthcare System for IoT-Based Multi-Biometric Information Measurement. *J. Korea Converg. Soc.* **2020**, *11*, 53–61.
4. Cho, M.-G. A Study on Wellbeing Support System for the Elderly Using AI. *J. Converg. Inf. Technol.* **2021**, *11*, 16–24.
5. Ha, T.; Lee, H. Implementation of Application for Smart Healthcare Exercise Management Based on Artificial Intelligence. *J. Inst. Electron. Inf. Eng.* **2020**, *57*, 44–53.
6. Lee, N.-K.; Lee, J.-O. A Study on Mobile Personalized Healthcare Management System. KIPS Trans. *Comp. Commun. Syst.* **2015**, *4*, 197–204.
7. Yoon, S.; Byun, S.; Jung, K. Multimodal Speech Emotion Recognition Using Audio and Text. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; IEEE: New York, NY, USA, 2018; pp. 1–6.
8. Kim, Y.-J.; Roh, K.; Chae, D. Feature-Based Emotion Recognition Model Using Multimodal Data. In Proceedings of the 2023 Korean Computer Congress (KCC), Seoul, Republic of Korea, 21–23 June 2023; Korean Institute of Information Scientists and Engineers: Seoul, Republic of Korea, 2023; pp. 2169–2171.
9. Park, H. Enhancement of Multimodal Emotion Recognition Classification Model through Weighted Average Ensemble of KoBART and CNN Models. In Proceedings of the 2023 Korean Computer Congress (KCC), Seoul, Republic of Korea, 21–23 June 2023; Korean Institute of Information Scientists and Engineers: Seoul, Republic of Korea, 2023; pp. 2157–2159.
10. Kim, S.-S.; Yang, J.-H.; Choi, H.-S.; Go, J.-H.; Moon, N. The Research on Emotion Recognition through Multimodal Feature Combination. In Proceedings of the 2024 ASK Conference, Seoul, Republic of Korea, 15–17 May 2024; ASK Conference Proceedings: Seoul, Republic of Korea, 2024; pp. 739–740.
11. Byun, Y.-C. A Study on Multimodal Korean Emotion Recognition Using Speech and Text. Master's Thesis, Graduate School of Information and Communication, Sogang University, Seoul, Republic of Korea, 2024.
12. Agarkhed, J.; Vishalakshmi. Machine Learning-Based Integrated Audio and Text Modalities for Enhanced Emotional Analysis. In Proceedings of the 5th International Conference on Inventive Research in Computing Applications (ICIRCA 2023), Bengaluru, India, 20–22 July 2023; IEEE: New York, NY, USA, 2023; pp. 989–993.
13. Makhmudov, F.; Kultimuratov, A.; Cho, Y.-I. Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures. *Appl. Sci.* **2024**, *14*, 4199. [CrossRef]
14. Feng, L.; Liu, L.-Y.; Liu, S.-L.; Zhou, J.; Yang, H.-Q.; Yang, J. Multimodal Speech Emotion Recognition Based on Multi-Scale MFCCs and Multi-View Attention Mechanism. *Multimed. Tools Appl.* **2023**, *82*, 28917–28935. [CrossRef]
15. Luo, J.; Phan, H.; Reiss, J. Cross-Modal Fusion Techniques for Utterance-Level Emotion Recognition from Text and Speech. *arXiv* **2023**, arXiv:2302.02447.

16. Patamia, R.A.; Santos, P.E.; Acheampong, K.N.; Ekong, F.; Sarpong, K.; Kun, S. Multimodal Speech Emotion Recognition Using Modality-Specific Self-Supervised Frameworks. *arXiv* **2023**, arXiv:2312.01568.
17. Wang, P.; Zeng, S.; Chen, J.; Fan, L.; Chen, M.; Wu, Y.; He, X. Leveraging Label Information for Multimodal Emotion Recognition. *arXiv* **2023**, arXiv:2309.02106.
18. Wu, Z.; Lu, Y.; Dai, X. An Empirical Study and Improvement for Speech Emotion Recognition. *arXiv* **2023**, arXiv:2304.03899.
19. Clark, K.; Luong, M.-T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020.
20. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. *Audio Speech Lang. Process.* **2021**, *29*, 3459–3473.
21. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.
22. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. *arXiv* **2019**, arXiv:1906.00295.
23. KEMDy19 Dataset. Available online: https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR (accessed on 30 December 2024).
24. KEMDy20 Dataset. Available online: https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR (accessed on 30 December 2024).
25. AI-HUB Korean Emotion Recognition Dataset. Available online: https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=263 (accessed on 30 December 2024).
26. Yi, M.-H.; Kwak, K.-C.; Shin, J.-H. KoHMT: A Multimodal Emotion Recognition Model Integrating KoELECTRA, HuBERT with Multimodal Transformer. *Electronics* **2024**, *13*, 4674. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.