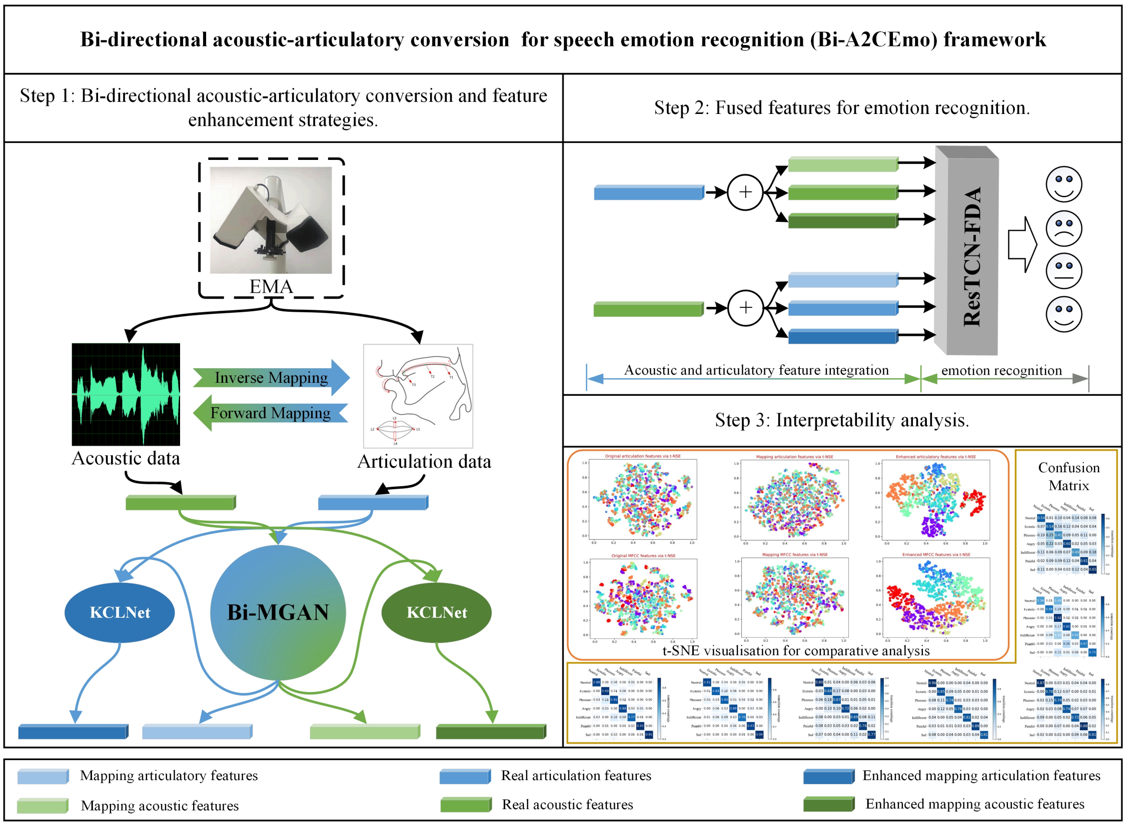
Graphical Abstract

Speech Emotion Recognition Based on Bi-directional Acoustic-Articulatory Conversion

Haifeng Li, Xueying Zhang, Shufei Duan, Huizhi Liang



Speech Emotion Recognition Based on Bi-directional Acoustic-Articulatory Conversion

Haifeng Li[a], Xueying Zhang[a,*], Shufei Duan[a], Huizhi Liang[b,**]

[a] *College of Electronic Information and Optical Engineering, Taiyuan University of Technology, Taiyuan Shanxi, 030024, China*

[b] *School of Computing, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK*

Abstract

Acoustic and articulatory signals are naturally coupled and complementary. The challenge of acquiring articulatory data and the nonlinear ill-posedness of acoustic–articulatory conversions have resulted in previous studies on speech emotion recognition (SER) primarily relying on uni-directional acoustic–articulatory conversions. However, these studies have ignored the potential benefits of bi-directional acoustic–articulatory conversion. Addressing the problem of nonlinear ill-posedness and effectively extracting and utilizing these two modal features in SER remain open research questions. To bridge this gap, this study proposes a Bi-A2CEmo framework that simultaneously addresses the bi-directional acoustic-articulatory conversion for SER. This framework comprises three components: a Bi-MGAN that addresses the nonlinear ill-posedness problem, KCLNet that enhances the emotional attributes of the mapped features, and ResTCN-FDA that fully exploits the emotional attributes of the features. Another challenge is the absence of a parallel acoustic-articulatory emotion database. To overcome this issue, this study utilizes electromagnetic articulography (EMA) to create a multi-modal acoustic-articulatory emotion database for Mandarin Chinese called STEM-E^{2}VA. A comparative analysis is then conducted between the proposed method and state-of-the-art models to evaluate the effectiveness of the framework. Bi-A2CEmo achieves an accuracy of 89.04% in SER, which is an improvement of 5.27% compared with the actual acoustic and articulatory features recorded by the EMA. The results for the STEM-E^{2}VA dataset show that Bi-MGAN achieves a higher accuracy in mapping and inversion than conventional conversion networks. Visualization of the mapped features before and after enhancement reveals that KCLNet reduces the intra-class spacing while increasing the inter-class spacing of the features. ResTCN-FDA demonstrates high recognition accuracy on three publicly available datasets. The experimental results show that the proposed bi-directional acoustic-articulatory conversion framework can significantly improve the SER performance.

Keywords: Speech emotion recognition, Acoustic and articulatory conversions, Cycle consistent generative adversarial networks, Temporal convolutional network, Contrastive Learning

[*] Corresponding author. E-mail address: zhangxy@tyut.edu.cn (X. Zhang).
[**] Corresponding author. E-mail address: Huizhi.Liang@newcastle.ac.u k (H. Liang).
E-mail address: duanshufei@tyut.edu.cn (S. Duan), 2023310098@link.tyut.edu.cn (H. Li).
The above authors contributed equally to this work.

1. Introduction

Speech emotion recognition (SER) is a crucial research area in contemporary human-computer interaction that strives to empower computers with the capacity to comprehend and identify emotional information conveyed in speech [1]. The outcomes of SER research have already exerted a significant influence across various domains, including autonomous driving, depression diagnosis, web call services, recommender systems, and healthcare science [2, 3, 4]. In its initial stages, most of this research focused on unimodal speech data for emotion analysis [5, 6, 7]. However, SER based solely on speech modality has several drawbacks, such as in the case of speakers with damaged vocal cords or low speech intelligibility caused by diseases. These limitations significantly restrict the recognition performance of the system. To address these drawbacks, an increasing number of studies have investigated emotions by combining speech with other modal data, such as behavioral signals like body movements and facial expressions. Furthermore, speech can be combined with physiological signals [8, 9], including motor signals from the articulatory organs. Articulatory features refer to the position and movement of the tongue, teeth, lips, and other articulatory organs. In general, articulatory signals encompass multiple features that can be utilized for emotion recognition, including the state (silence or movement), range, displacement, and velocity of the articulatory organs [10, 11, 12]. Nevertheless, the limited availability and high cost of articulation data have led these signals to be neglected in most of the existing research. Ensuring convenient access to parallel acoustic and articulatory data and leveraging the fused signals from both sources are crucial for advancing SER research.

Inverse mapping is currently considered the mainstream approach for acquiring articulatory data [13]. Inverse mapping utilizes acoustic features and converts them one-to-one into the corresponding articulatory features, thereby addressing the challenge of acquiring difficult-to-obtain articulatory data and reducing the data acquisition cost [14]. Forward mapping is the counterpart of inverse mapping and involves the conversion of the corresponding acoustic features using the kinematic features of the articulatory organs [15]. Both forward and inverse mapping are of great importance for SER research. However, acoustic and articulatory conversions are highly nonlinear, making both mapping and inversion challenging because of their ill-posed nature [16]. Consequently, scholars have faced numerous challenges in addressing this problem using traditional statistical and machine learning methods [17, 18].

Most SER algorithms based on acoustic and articulatory conversion utilize either forward or inverse mapping to extract acoustic or articulatory features and then employ a subsequent classification model for emotion classification. Previous studies have separated the forward and inverse mapping by considering them as separate tasks. However, parallel acoustic and articulatory signals are applicable for both forward and inverse mapping. We argue that the joint consideration of mapping and inversion will greatly improve the efficiency of the use of acoustic and articulatory data. The major challenge encountered in SER based on acoustic and articulatory data lies in the lower emotion recognition rate of the mapped features compared with real features, which directly hampers the practicality and widespread adoption of this research [10, 19]. Solving the problem of the low emotion recognition rate of mapped features is one of the most important issues in current research. Therefore, in this study, we design algorithms that can solve the problem of nonlinear ill-posedness and generate high-precision mapping features.

SER research comprises two parts: features [20] and recognition algorithms [21]. To enable SER networks to better recognize the emotions conveyed in speech, extracting salient features is indispensable. Some of the classical features include the mel frequency cepstral coefficient (MFCC), fundamental frequency, and pitch. As technology has advanced, one of the main research trends in feature extraction is capturing emotional features from multi-modal speech signals, e.g., the combination of speech with facial expressions or body movements [22, 23], or the integration of acoustic features, semantic primitives, and emotional dimensions (valence and arousal) [24]. Deep learning recognition algorithms including convolutional neural networks (CNNs) [25], transformers [6], recurrent neural networks (RNNs) [7], temporal convolutional networks (TCNs) [26], and deep belief networks (DBNs) have been successfully applied to SER [27]. In current research, TCNs are being rapidly deployed in SER [28, 25]. The TCN-based emotion recognition method is performed as follows: first, the emotion-dependent features are extracted using multilayer dilation convolution, and then a classifier is used to complete the emotion recognition. The methods mentioned above mainly focus on feature extraction and emotion recognition for the emotional information embedded in speech. However, articulatory features or the fusion of acoustic and articulatory features are not considered. Therefore, developing weighted adaptive emotion recognition algorithms that can incorporate both acoustic and articulatory fusion features is important. Attention mechanisms can offer a practical solution to the issue of redistributing weight coefficients. Therefore, in this study, we construct a TCN emotion recognition algorithm that can adaptively allocate weight coefficients using an attention mechanism.

To effectively leverage the complementarity and coupling of parallel acoustic and articulatory data in SER, this study proposes a **Bi**-directional **A**coustic–**A**rticulatory **C**onversion framework for speech **Emo**tion recognition (Bi-A2CEmo). Drawing inspiration from generative adversarial thinking, contrastive learning, and the attention mechanism in deep learning, this study integrates and applies these approaches to construct a Bi-A2CEmo framework. The framework consists of three components: a Bi-MGAN for bi-directional acoustic-articulatory conversion, a KCLNet for enhanced mapping of feature emotional attributes, and a ResTCN-FDA network that adaptively assigns weights to feature and dimension channels.

The contribution of this paper can be summarized as follows:

- This study introduces a unified and extensible Bi-A2CEmo framework. This framework can simultaneously perform the mapping and inversion tasks of acoustic and articulatory signals, enabling a better understanding of the distributions of real features and generating highly accurate mapped features. Moreover, the Bi-A2CEmo can enhance the emotional attributes of the mapped features, and the weight-adaptive operation of the recognizer can further enhance the recognition performance of the algorithm.

- We propose Bi-MGAN to address the nonlinear ill-posedness problem in acoustic and articulatory conversions. Bi-MGAN is based on a generative adversarial mechanism that learns the potential coupling between the mapping and inversion. In addition, we develop KCLNet based on contrastive learning to enhance the affective attributes of the mapped features.
- We propose utilizing feature dimension attention (FDA) for the adaptive assignment of weights to feature matrices; the FDA algorithm is then integrated with residual TCNs (ResTCNs) to build ResTCN-FDA.

The remaining sections of the paper are structured as follows. Section 2 provides a review of relevant underlying models in SER. Section 3 presents the specific components of the Bi-A2CEmo framework. Section 4 describes the database and features designed and recorded in this study. Section 5 outlines the experimental design and discusses the results. Section 6 discusses the interests and limitations of the proposed method. Section 7 summarizes our work and proposes future directions.

2. Related work

Over the past decade, forward and inverse mapping have been the dominant methods for studying acoustic-articulatory conversion. With the development of machine learning techniques, significant performance improvements have been achieved for both forward and inverse methods. In the field of forward mapping, most studies have focused on utilizing machine learning methods to model the potential coupling relationship between articulation and acoustic features. Iing et al. [29] proposed an improved hidden Markov model (HMM) to explore the joint articulatory-to-acoustic distribution relationship, which converted the acoustic features using articulatory features. That study found that known articulatory features could be converted into acoustic features. Acoustic data have wide utility in reality, and several studies have focused on deploying forward mapping in the field of acoustics. These studies have proposed practical methods for solving real-world problems [30]. The use of forward mapping to design articulatory synthesizers with native accents for non-native speakers is a classic example [31]. In research on inverse mapping, some studies have modeled and analyzed the coupling between articulation and acoustics, whereas others have attempted to apply inverse mapping to downstream tasks, such as emotion recognition and dysarthria [19, 32]. Compared with forward mapping, relatively few studies have focused on inverse mapping. This is because inverse mapping uses known acoustic signals to convert unknown articulatory signals, which are not as widely used in life as acoustic signals. These studies have demonstrated the potential coupling between acoustic and articulatory signals and the need to use mapping and inversion techniques to investigate the acoustic domain. However, these studies used split mapping and inversion, arguing that these processes could not be performed simultaneously in a single model. However, this reduces the efficiency of mining parallel acoustic and articulatory data and limits the analysis of the strong correlation between the two types of data.

Studies have demonstrated that machine learning-based mapping and inversion methods are often effective. Ren et al. [10] proposed a particle swarm optimization-based least-squares support vector machine (PSO-LSSVM) algorithm for exploring articulatory-to-acoustic conversion. However, the accuracy of these methods tends to deteriorated when predicting downstream acoustic tasks, leading to a significant disparity between the mapped and real features. To address the issue of nonlinear ill-posedness in acoustic-articulatory conversion, one approach is to utilize deep neural networks (DNN) for feature-level and score-level conversion of both modal features [19]. However, it should be noted that DNNs cannot effectively model long-term dependencies in features. Therefore, one proposed algorithm utilized a bi-directional long short-term memory recurrent (BiLSTM) [33] as a conversion model. This model efficiently captured acoustic-articulatory features over long distances. However, BiLSTM requires an externally specified context window. Thus, the deep recurrent mixture density network (DRMDN) algorithm was proposed, which can adaptively learn contextual information in the features [34]. In addition, with the rise of generative models, conversion models based on variational auto-encoders (VAE) have also been proposed. These models can be combined with regularization techniques to learn the kinematic trajectories of articulatory organs using the movement parameters of the jaw, tongue, and lips [32]. Our concept is to construct a bi-directional acoustic-articulatory conversion model based on the generative adversarial concept. This model will be able to perform mapping and inversion tasks simultaneously, allowing for a more comprehensive understanding of the potential conversion laws between the two modal features.

Acoustic signals are generated by the unique movements of articulatory organs [35]. Therefore, there is a natural coupling and complementarity between the emotional information embedded in speech and the movement trajectories of articulatory organs. Common approaches for exploring the emotional relevance of acoustic and articulatory signals include meta-analyses, multivariate discriminant analyses, and machine learning [19, 30, 32]. Among these, machine learning approaches have made great strides in the field. Lee et al. [11] used a machine learning approach to demonstrate that human articulatory joints are significantly advanced in recognizing emotions such as neutrality, anger, happiness, and excitement. Kim et al. [36] concluded that articulatory joint emotional information can be predicted using an inverse model. Based on this, Erickson et al. [37] used the XGBoost method to achieve a significant improvement in emotion recognition performance for bimodal features by fusing speech and articulation compared with the performance for single-modal features. These studies validated the advancement of articulatory features as well as the fusion of acoustic and articulatory features in emotion recognition. However, the problem of mapping features with a lower emotion recognition rate than the real features has been neglected [10, 19]. In recent years, contrastive learning has shown promising results in feature enhancement research. This method typically compares pairs of positive and negative samples to learn more discriminative feature representations. One study [38] found that incorporating comparative learning as a feature enhancement module into the overall recognition framework could significantly improve the generalization performance of the system. Based on this, we plan to build a feature enhancement module that can effectively enhance the emotional attributes of mapped features.

Multi-modal and multi-scale fusion of features is currently a mainstream research direction in SER [22, 23]. For example, Chen developed a network based on a connected attention mechanism to achieve the early fusion of multi-scale features [39]. Zhu used a global perceptual fusion module to learn multi-scale emotion representations [40]. Heqing used a multi-level acoustic information module to extract multi-scale features of MFCC, spectrograms, and acoustic information; these features were then fused using a collaborative attention mechanism [41]. Xingfeng proposed the usd of a three-layer model comprising acoustic features, semantic primitives, and affective dimensions to represent the subtle emotional information in speech [24]. These approaches have pushed the development of SER, but they have focused only on the expression of emotion from multiple perspectives, ignoring the variability in the internal parameters of the features themselves in portraying emotion. Acoustic and articulatory features are composed of many different fine-grained parameters, and some variability exists in the portrayal of emotions across different fine-grained parameters. However, many CNN-based SER systems have been developed in the field of classifier research [25, 28]. For example, Zhao et al. [42] constructed a multi-dimensional cascaded CNN-LSTM network to learn local and global emotion representations from speech and spectrograms. Anvarjon et al. [43] proposed a low-complexity model by improving the pooling strategy of the CNN convolutional layers. Zhang et al. [44] reported the existence of a gap between emotions and features and proposed deep convolutional neural networks (DCNNs) to bridge this gap. However, the fixed feeling field of the CNN and the same coefficients of the channel dimension weights limit the learning ability of the model, which leads to difficulty in fitting the model to the differences between channel dimensions and emotions. Moreover, the fusion of acoustic and articulatory features has a certain degree of variability in the portrayal of emotions in different dimensional channels after the features have been extracted by the deep network. Therefore, we combine a sensory field-scalable TCN with an attention mechanism to solve the problem of feature and dimension channel weight adaptation.

As mentioned above, traditional conversion models rely on a single mapping or inversion approach to investigate the correlation between two modalities. This limitation hinders the ability of these models to analyze the interplay between modalities in terms of reconstructing acoustic or articulatory signals, resulting in low predictive power and a low data mining rate. In SER studies that utilize acoustic-articulatory conversion, the issue of a lower recognition rate for mapped feature emotions compared with real features remains unresolved, directly impacting the applicability of the method. Conversely, the issue of the adaptive weighting of features and dimensional channels in the recognition network has been overlooked. To address these challenges, this study proposes a bi-directional acoustic-articulatory conversion-based framework for emotion recognition that incorporates enchantable mapping and inversion techniques into a weight-adaptive emotion recognition network. This framework not only synchronizes the mapping and inversion tasks through a generative adversarial mechanism, but also enhances the emotions of the mapped features using contrastive learning and attention mechanisms while also adapting the weights of the recognition network.

3. Proposed Bi-A2CEmo framework

The proposed Bi-A2CEmo framework can synchronize the generation of mapped acoustic and articulatory features, leading to improved recognition results and enhanced overall recognition of the SER system. Table 1 summarizes the notation used in this study.

Table 1: **Notation**.

Notation	Description	Notation	Description
$\mathbf{X}$	articulatory feature domain	E	expectations
x	articulation features	$\mathrm{G}_{\mathrm{X}\to\mathrm{Y}}$	forward generator
$\hat{x}$	mapped articulatory features	$\mathrm{G}_{\mathrm{Y}\to\mathrm{X}}$	inverse generator
$\widehat{x'}$	enhanced mapped articulatory features	D_{X}	articulation discriminator
$\tilde{x}$	cycle articulation features	D_{Y}	acoustic discriminator
$\mathbf{Y}$	acoustic MFCC feature domain	L_a	adversarial loss function
y	acoustic MFCC features	L_c	cycle consistency loss function
y_i	i-th order MFCC feature	L_1	L_1 regularization
$\hat{y}$	mapped acoustic MFCC features	L_{bce}	cross entropy loss function
$\widehat{y'}$	enhanced mapped acoustic features	L_g	generator loss function
$\tilde{y}$	cycle acoustic MFCC features	L_m	bounded mapping loss function
z	Acoustic-articulatory fusion features	$\otimes$	element-wise product
z'	features of TCN output	f_α	embedding layer
$\bar{z}$	features of ResTCN output	g_α	projection layer
$\bar{z}'$	features of F_f module outputs	q_α	prediction layer
$\bar{z}''$	features of F_f and F_d module outputs	$\oplus$	element-wise add
r_i	real features of the i th sample	L_k	KCLNet loss function
F	features of different types of parameters	$\leftarrow$	assign a value
C	network output dimension-channel	t	iteration
F_f	feature attention mechanisms	w	full connectivity layer mapping
F_d	dimensional attention mechanisms	Υ	network function of EN-branch
$\Vert\Vert_2^2$	L_2 cosine similarity loss function	$\Re$	field of real numbers
$\mathrm{E}_{\hat{x},y}$	sum of loss expectations for $\hat{x}$ and y	N_{test}	test set sample size
θ^t	trainable parameters of EN-branch at iteration t	rank_i	serial number of sample i
η^t	features for CN-branch clustering at iteration t	M	number of positive samples
F_{ave}	feature means under different dimensional channels	N	number of negative samples
η_y	optimal real representation clustered from y	$\ln$	logarithmic functions based on e
e_i	mapping features of the i th sample	θ	learnable parameters for EN-branch

Before model training, we conducted feature extraction on the bimodal emotion dataset captured by the EMA, which encompassed acoustic and articulatory data. The extracted features, comprising 28-dimensional kinematic features of articulatory organs, x, and 60-dimensional MFCC features, y, were then employed as inputs for the model. Fig. 1 shows the architecture of Bi-A2CEmo, which consists of three key components: Bi-MGAN, KCLNet, and ResTCN-FDA. During the training process, Bi-MGAN employs a generative adversarial approach to predict the mapped features. These mapped features are subsequently passed to KCLNet to enhance the emotional attributes. The enhanced mapped features are fused with real features, and the resulting fused features are utilized by ResTCN-FDA for the emotion recognition task. The following subsections provide a detailed discussion of these three components.

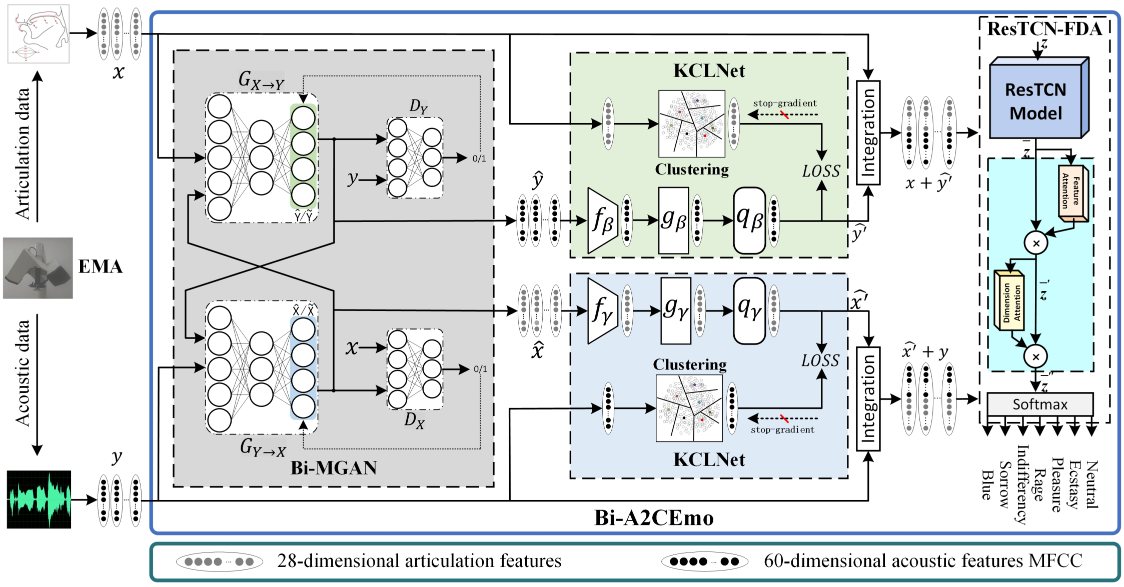


Figure 1: Proposed Bi-A2CEmo framework, which comprises three steps: (1) bi-directional acoustic-articulatory conversion is achieved by inputting real features (x or y) into Bi-MGAN to predict the corresponding mapped features ($\hat{y}$ or $\hat{x}$); (2) KCLNet enables the mapped features to learn the emotional attribute information of the optimal real features using a contrastive learning strategy; and (3) ResTCN-FDA performs adaptive emotion modeling on the fused features (real features and enhanced mapping features). Two cases of bi-directional acoustic-articulatory conversion are illustrated: (1) when the real articulatory feature, x, is known, Bi-A2CEmo sequentially predicts the mapped acoustic feature, $\hat{y}$, enhances the mapped acoustic feature, $\widehat{y'}$, and performs emotion recognition of the fused feature, $(x + \widehat{y'})$; (2) when the real acoustic feature, y, is known, Bi-A2CEmo sequentially predicts the mapped articulatory feature, $\hat{x}$, enhances the mapped articulatory feature, $\widehat{x'}$, and performs emotion recognition of the fused feature, $(y + \widehat{x'})$.

3.1. Bi-MGAN

The goal of the conversion network is to use real features to generate highly precise mapped features. The aim of this study is to investigate the impact of these mapped features on SER. The cycle consistent generative adversarial network (CycleGAN) does not require pairs of training data when applied to image-style conversion tasks [45], unlike for acoustic and articulatory feature conversion tasks. Most speech in the human body relies on the production of unique vocal tract shapes, which necessitates a parallel relationship between acoustic and articulatory data. To enhance the mapping capabilities of the conversion model, we propose Bi-MGAN for acoustic and articulatory conversion tasks. Our improvement focuses on the network structure and loss function of CycleGAN.

3.1.1. Bi-MGAN structure

Compared with image-style conversion tasks, acoustic-articulatory conversion is less computationally intensive. In this study, the generator and discriminator are optimized to reduce redundancy in the conversion network, prevent gradient vanishing, and improve the mapping accuracy. As shown in Fig. 2, the Bi-MGAN model consists of a forward generator ($\mathrm{G}_{\mathrm{X}\to\mathrm{Y}}$), inverse generator ($\mathrm{G}_{\mathrm{Y}\to\mathrm{X}}$), articulatory discriminator (D_{X}), and acoustic discriminator (D_{Y}). Fig. 2 shows the data flow when training the Bi-MGAN model with acoustic and articulatory features as inputs. The inverse generators in the upper and lower halves of the figure are the same modules. Similarly, the forward generators in the upper and lower halves of the figure are also the same modules.

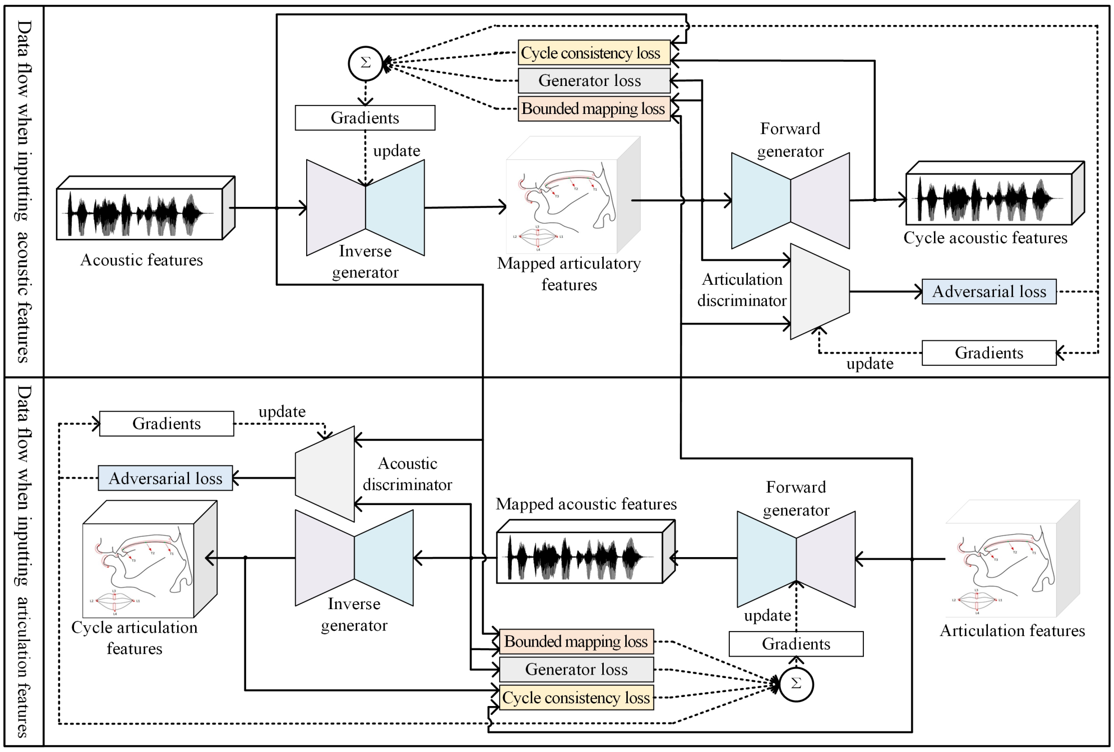


Figure 2: Loss function and data flow of Bi-MGAN during training.

- $G_{X \to Y}$. The forward generator utilizes articulatory features to establish a direct correspondence with acoustic features, aiming to prevent D_Y from accurately determining the mapped and real acoustic features. To minimize repetition, we implemented the up-sampling and down-sampling layers using a dense layer. The up-sampling layer increases the dimensionality of the input 28-dimensional articulatory features to 512 dimensions, while the down-sampling layer converts the high-dimensional articulatory features into 60-dimensional acoustic features. The structure of $G_{Y \to X}$ is designed to be the same as that of $G_{X \to Y}$, with the difference that $G_{Y \to X}$ utilizes MFCC features to represent the corresponding articulatory features. The objective of $G_{Y \to X}$ is to prevent D_X from accurately distinguishing between the mapped and real articulatory features.

- D_X. Articulation discriminators evaluate and calculate both real and mapped articulation characteristics. They utilize the weight parameters of the loss function callback $G_{Y \to X}$ to enhance the precision of the mapped features, effectively serving as supervisors and providing feedback for the mapped articulation features. D_X is essentially a binary recognizer that aims to accurately discriminate between mapped and real articulatory features. This is the exact opposite of what is expected from $G_{Y \to X}$. The conversion model determines the global optimal solution through an iterative optimization process that alternates between the two. D_Y is used to distinguish between real and mapped acoustic features, and the loss function is employed to adjust the weight parameters of $G_{X \to Y}$. This allows for supervision and feedback of the mapped acoustic features.

Fig. 3(a) shows a schematic diagram of the Bi-MGAN model, which converts real articulatory features, x, into mapped acoustic features, $\hat{y}$, and then converts $\hat{y}$ back into cyclic articulatory features, $\tilde{x}$. This process is described below.

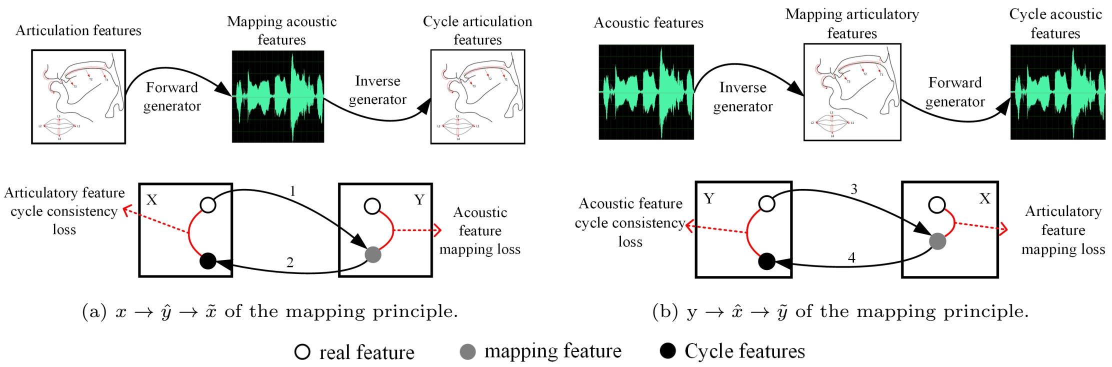
Step 1: The real acoustic features x are converted into their corresponding mapped articulatory features, $\hat{y}$ (1 in Fig. 3(a)).

Step 2: The acoustic feature mapping loss is calculated using the error between y and $\hat{y}$.

Step 3: The mapped articulatory features, $\hat{y}$, are converted into cyclic acoustic features, $\tilde{x}$ (2 in Fig. 3(a)).

Step 4: By calculating the error between x and $\tilde{x}$, the loss of cyclic consistency in the articulatory features can be determined.

Similarly, Fig. 3(b) shows a schematic diagram of the Bi-MGAN for converting real acoustic features, y, into mapped articulatory features, $\hat{x}$, and then converting $\hat{x}$ into cyclic acoustic features, $\tilde{y}$.



(a) $x \to \hat{y} \to \tilde{x}$ of the mapping principle.

(b) $y \to \hat{x} \to \tilde{y}$ of the mapping principle.

Figure 3: Schematic of the proposed Bi-MGAN network.

3.1.2. Bi-MGAN loss function

To address the issue of nonlinear ill-posedness in acoustic and articulatory conversion, we introduce a generator loss function and boundedness mapping loss function, which are derived from the loss function of CycleGAN. During training, Bi-MGAN incorporates four types of losses: generator loss, adversarial loss, cycle consistency loss, and bounded mapping loss. Fig. 2 illustrates the data flow relationship of the four loss functions in Bi-MGAN. The solid line represents forward propagation, whereas the dashed line represents backpropagation. In Bi-MGAN training, each epoch prioritizes the training of the discriminator. Once the discriminator can accurately identify the real and mapped features, it then proceeds to train $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}$ and $\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}$. This involves alternating iterative optimization of the generator and discriminator, which helps make the mapped features closer to the real features.

- Adversarial loss function [46]. The loss functions for $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}$ and D_{Y}, which are used to measure the discriminability of the mapped features from real features, are expressed as follows:

$$L_a\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}, \mathrm{D}_{\mathrm{Y}}\right) = E_{x\sim X}\left[\ln\left(1-\mathrm{D}_{\mathrm{Y}}\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\left(x\right)\right)\right)\right] + E_{y\sim Y}\left[\ln \mathrm{D}_{\mathrm{Y}}\left(y\right)\right] \tag{1}$$

  When D_{Y} discriminates against y, the loss value for the data is set to 1 if the discrimination is based on actual data. In contrast, when D_{Y} discriminates against $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\left(x\right)$, the loss value for the data is set to 0 if the discrimination is based on mapped data.

- Cycle consistency loss function [46]. The mapped features are converted to cycle features to enable the convergence of cycle features to real features. The equation is given as follows:

$$L_c\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}, \mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}\right) = E_{y-Y}\left[L_1\left(y, \mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\left(\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}(y)\right)\right)\right] + E_{x\rightarrow X}\left[L_1\left(x, \mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)\right)\right)\right] \tag{2}$$

  Where L_1 represents L_1 regularization.

- Generator loss function. We have added L_g as a base mapping function to enhance the conversion capabilities of the generator. The loss functions $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}$ and $\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}$ are thus defined as follows:

$$L_g\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\right) = E_{x-X}\left[L_{\mathrm{bce}}\left(\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\left(x\right)\right)\right] \tag{3}$$

$$L_g\left(\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}\right) = E_{y-Y}\left[L_{\mathrm{bce}}\left(\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}\left(y\right)\right)\right] \tag{4}$$

  Where L_{bce} denotes the cross-entropy loss function, and Bi-MGAN utilizes L_{bce} to determine $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)$. If the result of the judgment is true, it means that $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)$ has become indistinguishable from the real feature, y. If this judgment is false, it will result in errors. Eq. (4) is identical to Eq. (3); the only distinction is that Eq. (4) involves the manipulation of $\mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}(y)$.

- Bounded mapping loss function. To ensure the accuracy of the mapping features, relying solely on Eqs. (1)-(4) is not adequate to accomplish the acoustic and articulatory conversion tasks. In this study, the regularization of real and mapped features is applied to Bi-MGAN to constrain the range of the generated mapped features. This is achieved by reducing the number of mapped features with large errors generated by the model during training. The equations for the forward and inverse bounded mapping loss functions are as follows:

$$L_m\left(y, \mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}\right) = E_{x-X,y-Y}\left[L_1\left(y, \mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)\right)\right] \tag{5}$$

$$L_m\left(x, \mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}\right) = E_{x-X,y-Y}\left[L_1\left(x, \mathrm{G}_{\mathrm{Y}\rightarrow\mathrm{X}}(y)\right)\right] \tag{6}$$

  Eq. (5) states that $L_1\left(y, \mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)\right)$ is the L_1 difference between the real acoustic feature, y, and the mapped acoustic feature, $\mathrm{G}_{\mathrm{X}\rightarrow\mathrm{Y}}(x)$. Eq. (6) is equivalent to Eq. (5).

3.2. KCLNet

The purpose of KCLNet is to enhance the emotional information associated with the mapped features and address the issue of insufficient emotional information available for these features. KCLNet is a two-channel neural network comprising a clustered neural branch (CN-branch) and an enhanced neural branch (EN-branch). The CN-branch clusters real features using k-means clustering to extract the most emotionally expressive features from the sample. The EN-branch enhances the mapped features by incorporating emotional information through an encoder. Finally, KCLNet calculates the difference between the improved mapped features and the actual features using a cosine similarity (CosSim) function. This function enhances the emotional information of the mapped features.

3.2.1. KCLNet structure

KCLNet primarily aims to enhance the emotional information of the mapped features and address the issue of insufficient emotional information in the mapped features. This is achieved by continuously aligning the mapped features with the actual features, using an optimal emotional expression through a comparison prediction method. Fig. 4 illustrates KCLNet with real acoustic features y and mapped articulatory features, $\hat{x}$, as inputs. Its structure can be interpreted in terms of the CN-branch and EN-branch.

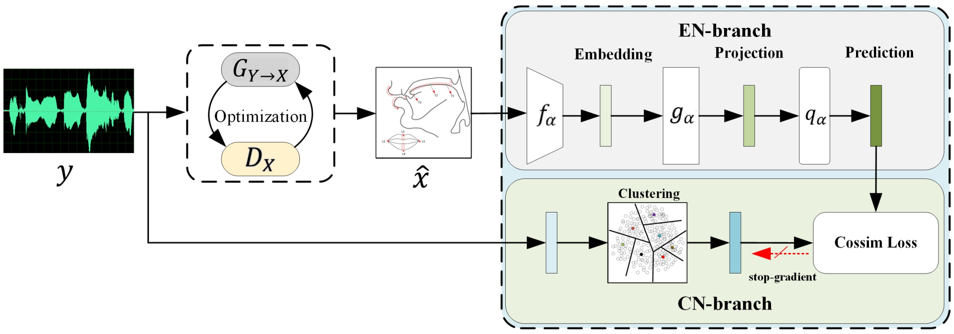


Figure 4: KCLNet network for enhanced mapping of articulatory features.

- CN-branch. First, we randomly set seven initial center-of-mass points. Then, we calculate the distance between each feature and the seven center-of-mass features. Based on this distance, each feature is assigned to an appropriate cluster. Finally, the center-of-mass points are updated based on the affiliation of the clusters. The most emotionally informative real acoustic features are obtained by iteratively performing operations until all feature points are closest to the center of mass.

- EN-branch. The EN-branch consists of embedding encoders, projection, and prediction. The embedding encoder (f_α in Fig. 4) deeply encodes the mapped articulatory features to generate emotionally expressive embedding features. The projection encoder (g_α in Fig. 4) projects the embedding features into a high-dimensional space. It consists of a fully connected (FC) layer, a normalization layer, and a ReLU. The FC layer that maps articulatory features has 60 dimensions, whereas that of the FC layer that maps acoustic features has 28 dimensions. The normalization layer and ReLU are connected after the FC layer. The predictive encoder (q_α in Fig. 4) predicts the projected features to generate enhanced mapped articulatory features.

We comprehend KCLNet using the expectation-maximization (EM) algorithm, which consists of four steps: optimizing real feature extraction, predicting mapped features, contrasting emotional similarities, and applying a stop-gradient. The specific operations for each epoch are as follows:

Step 1: The optimization of real features is performed by KCLNet, which clusters the real features based on different emotional states and identifies the optimal real features for each emotion from a pool of 2415 real features.

Step 2: Prior to predicting the mapping features, KCLNet initially conducts embedding and projection of the mapping features and subsequently generates enhanced mapping features using the prediction encoder.

Step 3: The cosine similarity comparison loss function is employed to evaluate the dissimilarity between the emotional information of the real and mapped features.

Step 4: The CN-branch utilizes a stop-gradient to prevent the back-propagation process and ensure optimal emotion representation of the real features obtained from clustering. Conversely, the EN-branch combines forward and back propagation to update the network parameters in alternating iterations.

3.2.2. KCLNet loss function

KCLNet optimizes the real features and enhances the emotion of mapped features through iterative optimization. Using the real acoustic features, y, and mapped articulatory features, $\hat{x}$, as inputs to KCLNet, the loss function, L_k, of the model is defined as follows:

$$L_k(\theta, \eta) = \mathrm{E}_{\hat{x},y}\left[\|\Upsilon_\theta(\hat{x}) - \eta_y\|_2^2\right] \tag{7}$$

where $\hat{x}$ represents the mapped articulatory features, y represents the real acoustic features, Υ represents the network function of the EN-branch, θ represents the learnable parameter of the EN-branch, η_y represents the optimal real representation under different emotional clusters from k-means clustering, $\mathrm{E}_{\hat{x},y}$ represents the sum of the loss expectations of $\hat{x}$ and y, and $\|\|_2^2$ represents the L_2 cosine similarity loss function. Eq. (8) shows the optimization objective of KCLNet.

$$\min_{\theta,\eta} L_k(\theta, \eta) \tag{8}$$

We interpret this optimization objective in terms of EM, thus splitting the above equation into two sub-objectives: Eq. (9) and Eq. (10).

$$\theta^t \leftarrow \arg\min_{\theta} p\left(\theta, \eta^{t-1}\right) \tag{9}$$

$$\eta^t \leftarrow \arg\min_{\eta} L_k\left(\theta^t, \eta\right) \tag{10}$$

where $\leftarrow$ represents the assignment and number of epoch iteration rounds, t is the iteration number, θ^t represents the learnable parameters of the EN-branch at t iterations, η^{t-1} denotes the optimal real features clustered by the CN-branch at $t-1$ iterations, and η^t denotes the optimal real features clustered by the CN-branch at t iterations. KCLNet solves for ot by fixing the η^{t-1} variable through the stop-gradient. When θ^t is known, it is substituted into Eq. (10) to find η^t. The above derivation uses only mapped articulatory features and real acoustic features as inputs, and it is necessary to swap x and y when mapped acoustic features and real articulatory features are used as inputs.

3.3. ResTCN-FDA

During the training of the recognition model, the same weights are assigned to different channels with different dimensional features. This can result in underutilization of emotional information. In this study, we propose a network for emotion recognition that combines ResTCN with an FDA attention mechanism. The FDA allows for weighting and adjustment of the features of the ResTCN output, thereby improving the utilization of acoustic and articulatory features that are significantly correlated with emotions.

Fig. 5 shows the overall ResTCN-FDA emotion recognition network. Here, z, which combines the real and mapped features, undergoes sequential dilation convolution, normalization, ReLU, and dropout operations in ResTCN. This process will generate the feature z', which contains the elemental dependencies. Then, z and z' are concatenated such that the features contain both overall emotion information and local element dependency information. Finally, the ResTCN output feature, $\bar{z}$, is input into the FDA to finalize the weight reallocation of the feature and dimension channels. During training, the convolution kernel for the dilated convolution of ResTCN is set to two. The module has a total of three ResTCN layers, so the overall dilation factor is $d = \{2^0, 2^1, 2^2\}$ in order. The variables $\boldsymbol{z} \in \Re^{F\times C}$, F, and C represent the number of features and output channel dimension of the feature map, respectively.

As shown in Fig. 5, the output signal, $\bar{z}$, of ResTCN sequentially passes through the feature attention mechanism, $F_f \in \Re^{F\times 1}$, and the dimension attention mechanism, $F_d \in \Re^{1\times C}$, to obtain the output signal, $\bar{z}'' \in \Re^{F\times C}$. The entire process is represented as follows:

$$\bar{z}' = F_f(\bar{z}) \otimes \bar{z} \tag{11}$$

$$\bar{\boldsymbol{z}}'' = F_d\left(\bar{z}'\right) \otimes \bar{\boldsymbol{z}}' \tag{12}$$

where $\bigotimes$ is the element-wise product. The details of F_f and F_d are given below.

- Feature attention. Different features respond differently into emotion recognition. To enhance the extraction of emotional information from multi-class features, this study calculates the weights of each feature class in $\bar{z}$. As shown in Fig. 5, the transposed feature vectors are

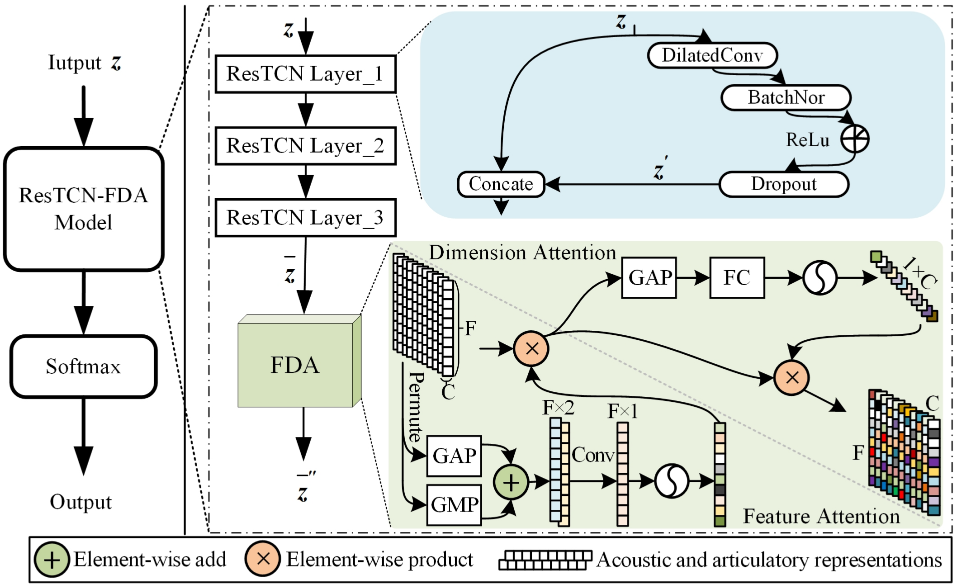


Figure 5: Overall structure of the ResTCN-FDA network.

first passed through global maximum pooling (GMP) and global average pooling (GAP). Then, the outputs of both are combined and passed through the convolutional and sigmoid layers to calculate the feature attention weight, $F_f \in \Re^{F\times 1}$.

- Dimensional attention. As shown in Fig. 5, the GAP operation is first performed on $\bar{z}'$ to obtain the mean feature, F_{avc}, for each dimension channel. Dimensional attention is implemented using an FC layer and a sigmoid function. Finally, the weight coefficients of the dimensional attention are applied to $\bar{z}'$, thereby assigning different weight coefficients to each dimension channel. The relevant equations for these calculations are as follows:

$$F_{ave,c} = \frac{1}{F}\sum_{f=1}^{F}\left(\overline{\boldsymbol{z}}'_c(f)\right) \tag{13}$$

$$F_d\left(\bar{z}'\right) = \mathrm{Sigmoid}\left(wF_{\mathrm{ave}}\right) \tag{14}$$

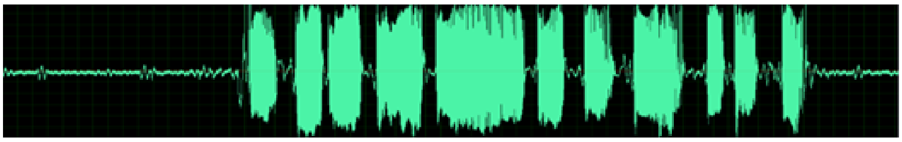
Eq. (13) represents the mean $F_{\mathrm{ave},c}$ of the features in channel c, and $\overline{\boldsymbol{z}}'_c \in \Re^{F\times 1}$ represents the $F\times 1$ features in channel c. In Eq. (14), w is the FC layer.

4. Emotion database and data representation

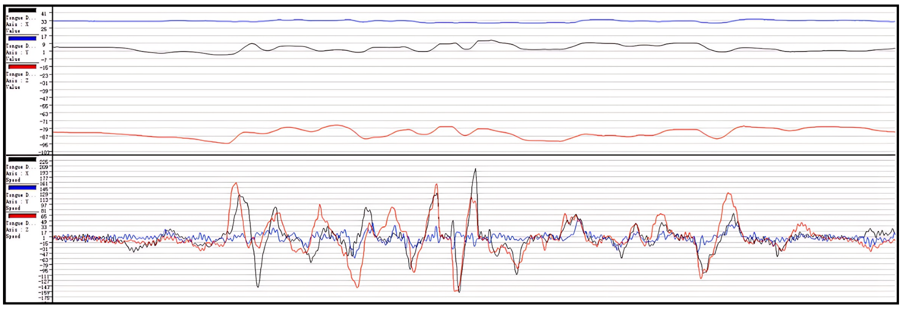
Owing to the lack of publicly available parallel acoustic and articulatory multi-modal emotional datasets, we recorded the Suzhou and Taiyuan emotional datasets in Mandarin with electromagnetic articulation, electroglottography, video, and audio (STEM-E^{2}VA) using EMA AG501 and extracted the features of both modalities based on this database.

4.1. Construction of the STEM-E^{2}VA acoustic-articulatory emotional database

Owing to the absence of parallel acoustic and articulatory emotion datasets, we recorded the STEM-E^{2}VA dataset and used it as the primary database for this study. It contains recordings and articulatory data from 22 native Mandarin-speaking individuals. Of the 22 participants, 62.5% had a bachelor's degree, and 37.5% had a master's degree. The average age of the participants was 25 years, and the male-to-female ratio was 1:1. Prior to data collection, all participants completed the Symptom Self-Rating Scale SCL-90. Only those who passed the scale were informed of the data collection process. The STEM-E^{2}VA database was completed based on the EMA AG501 collection. During recording, the EMA acquired the Cartesian coordinates of transducers fixed to the articulatory organs as articulatory data through electromagnetic coupling. The data was collected at a sampling rate of 250 Hz. In addition, the EMA synchronously recorded the acoustic data to form parallel acoustic and articulatory data. Fig. 6 shows the synchronized acoustic-articulatory signal waveforms acquired by the EMA AG501. Fig. 6(a) shows the acoustic signals. Fig. 6(b) shows the articulatory signals recorded by the hypoglossus sensor. The top three lines in the figure represent the positional parameters of the sensor along the X-axis (black), Y-axis (blue), and Z-axis (red). The bottom three lines of the graph represent the speed variation of the sensor on the X-axis (black), Y-axis (blue), and Z-axis (red).



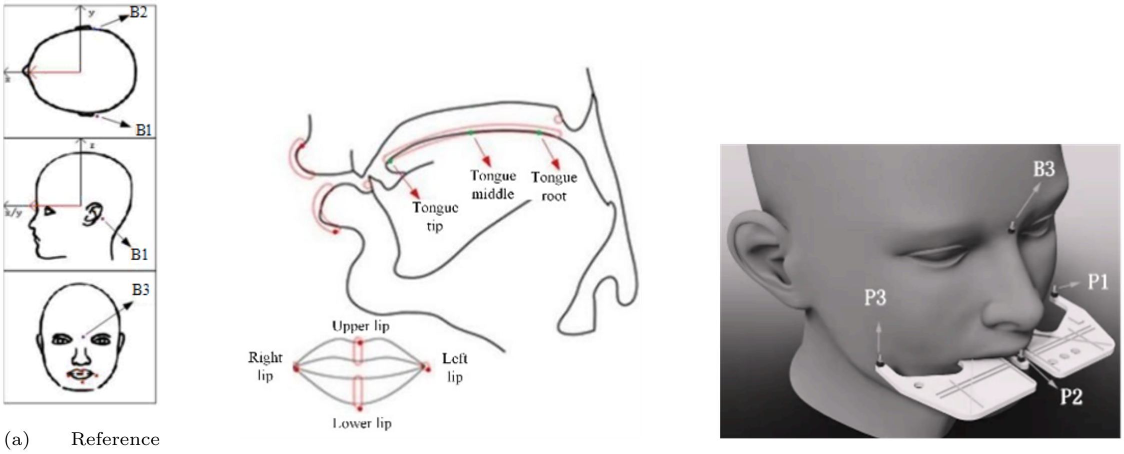
(a) Acoustic signals acquired by EMA AG501.



(b) Articulatory signals recorded by the hypoglossus sensor.

Figure 6: Synchronized acoustic-articulatory signals acquired by EMA AG501.

In the data acquisition of STEM-E^{2}VA, we installed 13 sensors. These sensors included three reference surface sensors, three bite-plate sensors, four lip sensors, and three tongue sensors. The detailed configurations of these sensors are shown in Fig. 7. The reference surface sensors were placed at positions B1, B2, and B3 on the participant to minimize errors caused by head movements during data collection. The bite plate sensors were arranged at positions P1, P2, and P3 on the surface of the bite plate. The reference surface sensor and bite plate sensor were intended to perform head and tooth calibration during the pre-processing of articulatory data and were not involved in articulatory data acquisition. The lip and tongue sensors were used to collect trajectory data of the articulatory organs. They were arranged as follows: left lip, right lip, upper lip, lower lip, tongue root, middle tongue, and tongue tip. After the sensors were able to transmit data consistently, the subjects were asked to articulate the contents of the corpus. We obtained 2415 parallel acoustic and articulatory data points. The amount of data was consistent for each emotion; therefore, we do not consider the error in recognition results caused by data imbalance in this study.



(a) Reference surface sensor

(b) Vocal organ sensor

(c) Bite plate sensor

Figure 7: Sensor settings for data acquisition by EMA AG501.

4.2. Acoustic and articulatory features

In this study, we extract MFCC and articulatory features from acoustic and articulatory data, respectively. The MFCC is widely used in acoustic and articulatory conversion tasks because of its robust representation of timbres in speech. This enables the articulatory features generated through MFCC mapping to exhibit good performance. In this study, MFCC is used as the acoustic feature, and the acoustic feature set is denoted by $\mathbf{Y}$. y_i is the i-th order MFCC; the skewness, kurtosis, mean, variance, and median parameters are extracted sequentially for y_i with $i = 12$. Therefore, $\mathbf{Y}$ is the MFCC feature with a dimension of 60.

Among the articulatory features, this study considers the motion features of the tongue and lips as the main features, and the displacement and velocity features of the articulatory organs when they move are extracted; $\mathbf{X}$ is used to denote the articulatory feature set. As shown in Fig. 7(b), $\mathbf{X}$ contains 21-dimensional displacement parameters for the left lip, right lip, upper lip, lower lip, tongue root, tongue middle, and tongue tip in a three-dimensional coordinate system, as well as seven-dimensional velocity parameters. Therefore, the $\mathbf{X}$ feature set is a 28-dimensional articulatory motion feature.

5. Experiments

To demonstrate the state-of-the-art of the proposed Bi-A2CEmo framework, this study presents an experimental evaluation of the overall framework and each of its three components. This section describes the experiments designed to answer the following research questions:

RQ1: How does the proposed Bi-A2CEmo framework perform in emotion recognition tasks compared with the baseline?

RQ2: How does the proposed Bi-MGAN algorithm perform in terms of forward and inverse mapping, compared with the baselines?

RQ3: Can the proposed KCLNet algorithm effectively solve the problem of low emotion recognition of mapped versus real features?

RQ4: Can the proposed ResTCN-FDA algorithm handle speech emotion recognition tasks better than the baselines?

In addition, a five-fold cross-validation scheme was used for all experiments during the training phase. The ADAM optimizer was used to update the step sizes. The neural networks implemented in this study were built using the TensorFlow library, Keras, and Scikit-learn.

5.1. Datasets

The experiments were evaluated using our self-constructed dataset and three publicly available datasets.

(1) **STEM-E^{2}VA:** Our constructed dataset includes both acoustic and articulatory data. We selected 2415 parallel acoustic-articulatory emotion data in this study, which included seven emotions: neutral, ecstatic, pleased, angry, indifferent, pained, and sad.

(2) **EMO-DB[1]:** This public speech database was organized by the University of Berlin, Germany, and recorded by 10 professional actors [47]. We selected 535 speech data samples from the dataset, which included seven emotions: anger, fear, boredom, disgust, joy, nertral, and sadness.

(3) **CASIA[2]:** This is a Chinese speech emotion dataset recorded by the Institute of Automation, Chinese Academy of Sciences [48]. For this study, we selected 1200 speech data points from this dataset, which included six emotional states: anger, fear, happiness, neutrality, sadness, and surprise.

(4) **RAVDESS[3]:** The Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS) [49] is a publicly available multi-modal emotional dataset. The dataset contains video and audio-only emotional data from 24 professional performers consisting of 12 females and 12 males. The RAVDESS comprises 7356 files. In this study, we selected only 1440 speech files that included eight emotional expressions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted.

5.2. Bi-directional acoustic-articulatory conversion for emotion recognition (RQ1)

In this study, we explored the effect of bi-directional acoustic-articulatory conversion on SER based on Bi-A2CEmo and compared it with mainstream recognition algorithms to validate the effectiveness of the Bi-A2CEmo framework. The experiments were dominated by parallel acoustic-articulatory data from STEM-E^{2}VA, and the EMO-DB [47], CASIA [48], and RAVDESS [49] datasets were used to validate the improvement provided by the ResTCN-FDA. The accuracy (ACC), F1-score (F1), area under the curve (AUC), and confusion matrix were used as evaluation

[1] http://emodb.bilderbar.info/docu/#emodb

[2] http://www.chineseldc.org/resource_info.php?rid=76

[3] https://zenodo.org/record/1188976

metrics. The ACC reflects the proportion of samples correctly classified and is expressed as $(TP+TN)/(TP+TN+FP+FN)$, where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. The closer the parameter values of these evaluation metrics are to 1, the better the classification performance of the model.

In experiments on the effect of bi-directional acoustic-articulatory conversion on emotions, the Bi-A2CEmo framework gave rise to two variants: Bi-A2CEmo^a and Bi-A2CEMo^b. Bi-A2CEmo^a utilized the ResTCN-FDA algorithm, whereas Bi-A2CEMo^b employed both the Bi-MGAN and ResTCN-FDA algorithms. The experimental paradigm for exploring the effect of bi-directional acoustic-articulatory conversion on SER based on Bi-A2CEmo was as follows: first, we extracted the acoustic and articulatory features of STEM-E^{2}VA, generated the mapping features using Bi-MGAN, enhanced the emotional attributes of the mapping features using KCLNet, and finally used ResTCN-FDA as the recognition network. The respective evaluation metrics were compared by inputting the features of different stages of the modality into ResTCN-FDA. Table 2 summarizes the evaluation metrics for the real, mapped, and enhanced mapped features of the acoustic and articulatory signals on ResTCN-FDA.

As indicated in Table 2, among the unimodal features, the enhanced mapped acoustic features had the highest ACC of 80.93%, and the mapped articulatory features had the lowest ACC of 53.02%. For both articulatory and acoustic features, the evaluation metrics of the enhanced mapped features were higher than those of the real features, which in turn were higher than those of the mapped features. This indicates that the mapped features contain less emotional information than the real features, i.e., forward and inverse mapping will reduce the amount of emotional information in the features. Meanwhile, (b) and (c) or (e) and (f) in Table 2 confirm that KCLNet can effectively enhance the emotional attributes of the mapped features.

As indicated in Table 2(k), among the bimodal fusion features, the feature that fused the enhanced mapped acoustics with real articulation exhibited the highest emotion recognition rate of 89.04%. As presented in Table 2(h), the features fusing mapped acoustics with real articulation had the lowest recognition rate of only 72.47%. The recognition rate of real acoustic features was improved by 3.88% and 12.19% after fusion with mapped articulatory features and enhanced mapped articulatory features, respectively; the recognition rate of real articulatory features was improved by 8.91% and 25.48% after fusion with mapped acoustic features and enhanced mapped acoustic features, respectively. This indicates that both mapped features and enhanced mapped features act as emotional complements to real features, and the emotional complementary effect of enhanced mapped features is more effective than that of mapped features. This also demonstrates that bi-directional acoustic-articulatory conversion can help the SER system learn potential emotional attributes in acoustic and articulatory signals that have been ignored previously, leading to a significant increase in the emotion recognition rate of the model.

To test the performance and robustness of the SER system, two comparative analyses were performed: the benchmarking of Bi-A2CEmo^a with previous research and the comparison of the model with the same input features. Table 3 summarizes the results of the comparison of

Table 2: Evaluating the emotion recognition performance (%) of the proposed method of the STEM-E^{2}VA dataset.

No.	Features	Framework	Dimension	ACC	F1	AUC
a	Acoustic(R)	Bi-A2CEmo^a	60	75.63	75.44	76.82
b	Acoustic(C)	Bi-A2CEmo^b	60	59.23	58.69	59.82
c	Acoustic(E)	Bi-A2CEmo	60	80.93	80.69	82.92
d	Articulatory(R)	Bi-A2CEmo^a	28	63.56	62.96	63.87
e	Articulatory(C)	Bi-A2CEmo^b	28	53.02	52.03	53.61
f	Articulatory(E)	Bi-A2CEmo	28	68.76	69.68	70.38
g	Acoustic(R)+Articulatory(C)	Bi-A2CEmo^b	88	79.51	79.69	79.97
h	Acoustic(C)+Articulatory(R)	Bi-A2CEmo^b	88	72.47	72.45	72.95
i	Acoustic(R)+Articulatory(R)	Bi-A2CEmo^a	88	83.77	83.64	83.97
j	Acoustic(R)+Articulatory(E)	Bi-A2CEmo	88	87.82	87.70	87.98
k	Acoustic(E)+Articulatory(R)	Bi-A2CEmo	88	**89.04**	**88.74**	**89.22**

[1] (R) represents the real features recorded by the EMA. (C) represents the mapped features converted by Bi-MGAN. (E) represents the enhanced mapped features jointly processed by Bi-MGAN and KCLNet.

[2] Bi-A2CEmo^a represents a variation of the framework that utilizes the ResTRCN-FDA algorithm. Bi-A2CEmo^b represents a variation of the framework that utilizes the Bi-MGAN and ResTCN-FDA algorithms. The Bi-A2CEmo framework utilizes the Bi-MGAN, KCLNet, and ResTCN-FDA algorithms.

Bi-A2CEmo^a with previous studies on the EMO-DB, RAVDESS, and CASIA databases. After a thorough and meticulous analysis of all of the studies presented in Table 3, the methodology introduced in this study achieved the highest ACC and F1 on the EMO-DB and CASIA datasets. However, upon evaluation of the RAVDESS database, the recognition efficacy of Bi-A2CEmo^a was found to be slightly lower than that of the benchmarks set forth in Ref. [50]. To elucidate this variation, we used the experimental methodology outlined in Ref. [50]. Our investigation revealed that the study employed a comprehensive LibROSA feature ensemble comprising 386 dimensions, including MFCCs, chroma vectors, mel-scaled spectrograms, spectral contrast features, and tonal centroid features. It is worth noting that the 60-dimensional MFCC features used in our study were a subset of the LibROSA feature set. In table 4, we compare Bi-A2CEmo^a with a conventional CNN and LSTM [50], as well as the latest HS-TCN [26] and DRN [51] algorithms, using 60-dimensional MFCC as the input features. As indicated in Table 4, the proposed network achieved accuracies of 80.41%, 75.63%, 80.16%, and 66.55% on the CASIA [48], STEM-E^{2}VA, EMO-DB [47], and RAVDESS [49] databases, respectively, which is a significant improvement in performance compared with the CNN, LSTM, HS-TCN, and DRN models. In addition, ResTCN-FDA achieved 2.01%-7.85% and 3.69%-7.19% improvements in the F1-score and 3.28%-6.07% and 2.96%-7.96% improvements in AUC compared with the HS-TCN and DRN networks, respectively, thus verifying the effectiveness of the improved Bi-A2CEmo model.

In the feature validity demonstration, the conventional algorithm uses acoustic features as inputs and ignores articulatory information. The Bi-A2CEmo model not only considers the

Table 3: Comparison of the current approach with previous studies based on the recognition rate for the EMO-DB, RAVDESS, and CASIA databases.

Datasets	References	Results(%)		Brief Description		
		ACC	F1	Feature	Classifier	Categories
EMO-DB	Li et al. [50]	N/A	90.93	MSF	LMT	Neutral, Happy, Angry, Sad
	This work	**91.88**	**91.45**	MFCC	Bi-A2CEmo[a]	
	Latif et al. [52]	70.98	N/A	GeMAPS	SVM	Angry, Fear, Boredom, Disgust, Joy, Neutral, Sad
	Liu et al. [53]	78.66	N/A	MFCC	SVM	
	Singh et al. [54]	79.86	N/A	MSF	DNN-SVM	
	This work	**80.16**	**80.78**	MFCC	Bi-A2CEmo[a]	
RAVDESS	Bagus et al. [24]	**78.50**	N/A	HSF	LSTM	Calm, Happy, Sad, Angry, Fear, Surprised, Disgust
	This work	75.19	74.97	MFCC	Bi-A2CEmo[a]	
	Singh et al. [54]	52.24	N/A	MSF	DNN-SVM	Calm, Happy, Sad, Angry, Fear, Surprised, Disgust, Neutral
	Liu et al. [53]	64.32	N/A	MFCC	SVM-RBF	
	Zeng et al. [55]	64.48	N/A	Spectrograms	GResNets	
	This work	**66.55**	**65.57**	MFCC	Bi-A2CEmo[a]	
CASIA	Li et al. [50]	N/A	78.51	MSF	LMT	Neutral, Happy, Angry, Sad
	This work	**88.75**	**88.67**	MFCC	Bi-A2CEmo[a]	
	Jiang et al. [56]	51.3	N/A	Spectrogram	CNN	Angry, Fear, Happy, Neutral, Sad, Surprised
	Zhang et al. [57]	79.67	N/A	HSF	HPCB	
	Mao et al. [58]	80.02	N/A	LLD	SVM	
	This work	**80.41**	**81.22**	MFCC	Bi-A2CEmo[a]	

[1] geneva minimalistic acoustic parameter set (GeMAPS).

[2] high-level statistical functions (HSF), modulation spectral features (MSF), low-level descriptor (LLD).

[3] logistic model trees (LMT), heterogeneous parallel conv-bilstm (HPCB), gated residual networks (GResNets).

emotion of acoustic features, but also the conversion and enhancement of acoustic features. In addition, the fusion features of the two modalities are modeled, unlike in conventional algorithms, which greatly improves the recognition accuracy of the system. For the STEM-E^{2}VA dataset, the recognition rate of ResTCN-FDA with only MFCC input was 75.63%, which was 5.30% lower than that with the enhanced mapped MFCC (Table 2(c)), thus demonstrating that Bi-MGAN and KCLNet can significantly improve the recognition rate of the system. When real acoustic MFCC features are used as the input, Bi-A2CEmo sequentially completes inversion and feature enhancement, as well as emotion recognition of the fused features. Table 2(g) and (j) present the recognition rates of MFCC with the fusion of mapped articulatory features and enhanced mapped articulatory features, respectively; their recognition rates are improved by 3.88% and 13.41% compared with those of ResTCN-FDA with MFCC as the input. Comparing Table 2 and Table 4, we can conclude that ResTCN-FDA in Bi-A2CEmo can significantly improve the emotion recognition accuracy of the system, and the Bi-A2CEmo framework can accomplish the functions of bi-directional acoustic-articulatory conversion and feature emotion enhancement, which significantly improves the emotion recognition rate of the system.

Fig. 8 shows the confusion matrix with a single acoustic or articulatory feature set as the

Table 4: Comparison of the recognition algorithm of Bi-A2CEmo with conventional recognition algorithms for the CASIA, STEM-E^{2}VA, EMO-DB, and RAVDESS datasets (%).

Database	**CASIA**			**STEM-E^{2}VA**			**EMO-DB**			**RAVDESS**		
Categories	6			7			7			8		
Metrics	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
CNN	63.00	62.43	63.19	56.77	56.84	57.77	69.72	69.09	69.86	57.29	55.67	57.85
LSTM	62.92	62.55	63.37	58.10	58.04	59.19	68.60	68.51	68.83	56.04	55.87	56.57
HS-TCN	76.25	76.64	76.91	72.81	72.59	72.92	74.76	73.60	75.51	63.29	63.56	63.58
DRN	76.91	76.67	76.94	68.15	68.25	68.86	76.64	74.72	76.96	63.46	61.88	63.90
Bi-A2CEmo^a	**80.41**	**81.22**	**81.43**	**75.63**	**75.44**	**76.82**	**80.16**	**80.78**	**81.58**	**66.55**	**65.57**	**66.86**

input. Fig. 9 shows the confusion matrix with acoustic-articulatory bimodal features as the input. From Fig.8(a), (b), and (c), it can be observed that the recognition rate of mapped acoustic features in the "Ecstatic" and "Painful" states is lower compared with the real features. However, the recognition rate of the enhanced mapped acoustic features in the "Pleased" and "Angered" states is significantly improved. This suggests that the quality of the features is influenced by the mapped acoustic features generated through forward mapping, which vary with different emotions. Additionally, it demonstrates the effectiveness of KCLNet in enhancing the mapped acoustic features. By comparing Fig. 8(d), (e), and (f), it is evident that the emotional quality of the mapped articulatory features varies significantly for different emotions. This also serves as evidence of the effectiveness of KCLNet in enhancing the mapped articulatory features. By comparing Fig. 8(a), (d), and Fig. 9(c), it can be seen that the fusion of real acoustic features with real articulatory features results in a significant improvement in the emotion recognition rate.

By comparing Fig. 8(a) and Fig. 9(a), (c), and (d), we observe that the mapped articulatory features, real articulatory features, and enhanced mapped articulatory features all provide additional emotional information to complement the real acoustic features. Specifically, the enhanced mapped articulatory features exhibit the strongest emotional complementation, whereas the mapped articulatory features demonstrate the weakest emotional complementation. By comparing Fig. 8(d) and Fig. 9(b), (c), we can also observe that the mapped acoustic features have the most negative impact on the emotion complementation of articulatory features, whereas the augmented mapped acoustic features have the most positive impact on emotion complementation.

5.3. Bi-MGAN performance comparison (RQ2)

To assess the effectiveness of the generator loss function and boundedness mapping loss function, we conducted ablation experiments on the conversion network based on STEM-E^{2}VA. This study establishes five sets of networks for validation: GAN [45], CycleGAN [46], Bi-MGAN(G) with the inclusion of a generator loss function, Bi-MGAN(M) with the inclusion of a boundedness mapping loss function, and Bi-MGAN(GM) with the inclusion of both generator and boundedness loss functions. This section evaluates the prediction performance using the mean absolute error

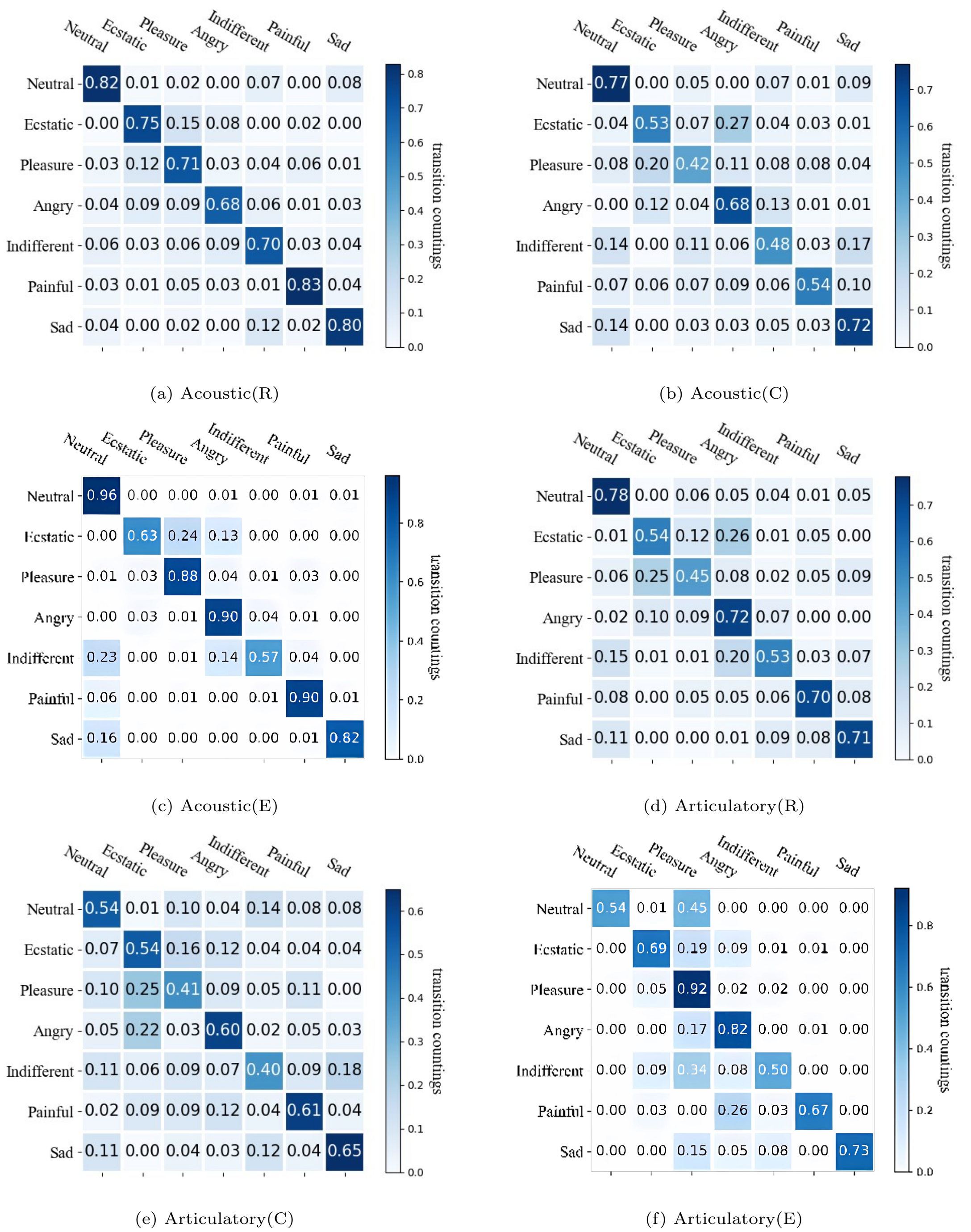


Figure 8: Confusion matrix of real, mapped, and enhanced mapped features for acoustic or articulatory features.

(MAE) and root mean square error (RMSE).

As indicated in Table 5, the MAE and RMSE of Bi-MGAN(G) were reduced by 0.010-0.093 mm and 0.011-0.087 mm, respectively, whereas those of Bi-MGAN(M) were reduced by 0.169-0.248 mm and 0.038-0.294 mm, respectively, compared with those of CycleGAN. This indicates that both the generator loss function and bounded mapping loss function are beneficial for both forward and inverse mapping. These functions help convert the model to generate highly accurate mapping features. In addition, the MAE and RMSE of Bi-MGAN(GM) are lower than those of

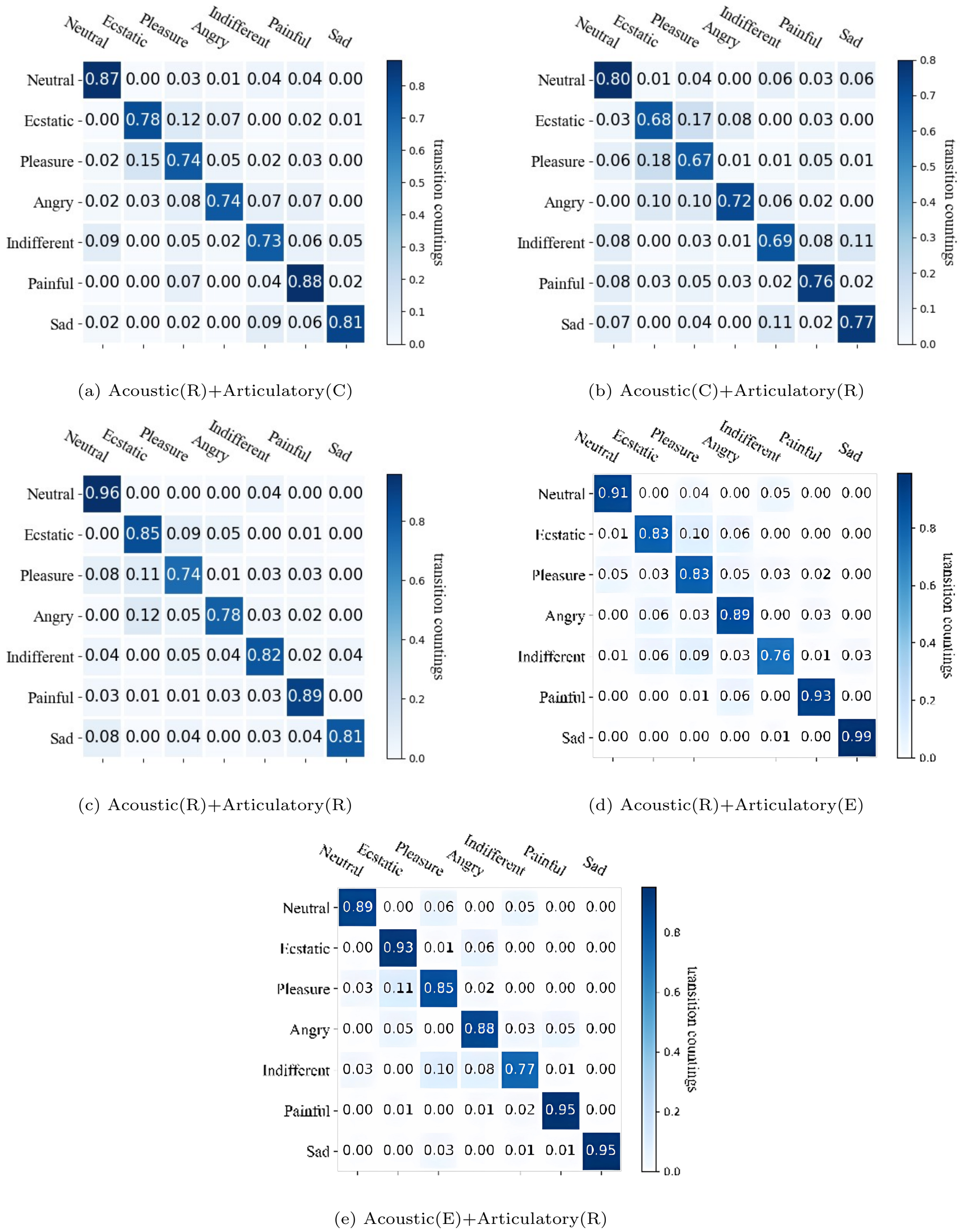


Figure 9: Confusion matrix for acoustic-articulatory bimodal features.

Bi-MGAN(M) and Bi-MGAN(G). This indicates that the combination of the two losses proposed in this study enhances the mapping ability of the conversion model and brings the mapped features closer to the real features.

Table 6 presents the performance of the proposed conversion network for both forward and inverse mapping compared with the baseline. The baseline methods include the PSO-LSSVM [10], DRMDN [34], BiLSTM [33], and DNN. DNN is implemented using neural networks with three hidden nonlinear layers, each consisting of 2048 nodes. BiLSTM has 500 units in the first

Table 5: Bi-MGAN network ablation experiment.

Method	**Forward Mapping**		**Inverse Mapping**	
	MAE	**RMSE**	**MAE**	**RMSE**
GAN[1]	1.217	1.642	–	–
GAN[2]	–	–	0.946	1.189
CycleGAN	1.127	1.428	0.811	0.919
Bi-MGAN(G)	1.034	1.341	0.801	0.908
Bi-MGAN(M)	0.879	1.134	0.642	0.881
Bi-MGAN(GM)	**0.703**	**0.920**	**0.501**	**0.683**

[1] GAN[1] is the Generative Adversarial Network with forward mapping.

[2] GAN[2] is the Generative Adversarial Network with inverse mapping.

two layers and 150 units in the last two layers. DRMDN uses a Gaussian mixture density output layer. The learning factors c_1 and c_2 of the PSO-LSSVM algorithm are both set to 1.5. From Table 6, it is evident that the MAE and RMSE values of Bi-MGAN are significantly lower than those of the baseline methods. This demonstrates that Bi-MGAN can significantly enhance the conversion accuracy of the network.

Table 6: Comparison of conversion networks.

Method	**Forward Mapping**		**Inverse Mapping**	
	MAE	**RMSE**	**MAE**	**RMSE**
DNN[1]	1.479	1.613	-	-
BiLSTM[1]	1.298	1.422	-	-
PSO-LSSVM[1]	1.185	1.252	-	-
DRMDN[1]	0.884	0.948	-	-
DNN[2]	-	-	1.143	1.259
BiLSTM[2]	-	-	1.003	1.217
PSO-LSSVM[2]	-	-	0.967	1.136
DRMDN[2]	-	-	0.831	0.939
Bi-MGAN	**0.703**	**0.908**	**0.501**	**0.683**

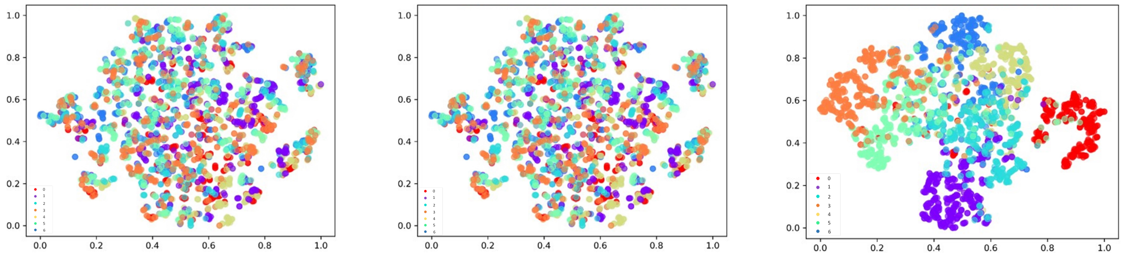
[1] DNN[1], BiLSTM[1], PSO-LSSVM[1], and DRMDN[1] represent forward mapping networks based on this baseline.

[2] DNN[2], BiLSTM[2], PSO-LSSVM[2], and DRMDN[2] represent inverse mapping networks based on this baseline.

5.4. Feature enhanced network performance analysis (RQ3)

Considering the complexity of acoustic and articulatory features, this study conducted a dimensionality reduction visualization and comparison analysis of real, mapped, and enhanced mapped features of the STEM-E^{2}VA dataset using t-SNE. The purpose of this analysis was to validate the emotion enhancement ability of KCLNet for mapped features.

Fig. 10(a), (b), and (c) show the distributional projections of t-SNE after dimensionality reduction for real, mapped, and enhanced mapped articulatory features, respectively. It is clear that the articulatory features recorded by the EMA and the mapped articulatory features generated by Bi-MGAN exhibit random distributions of articulatory data across similar emotions after dimensionalizing over t-SNE. Fig. 10(a) and (b) do not show a clear data distribution pattern. Instead, the enhanced mapping of articulatory features reveals a clear distribution rule under different emotional states, as shown in Fig. 10(c). This demonstrates that KCLNet results in a significant reduction in intra-class spacing and an increase in inter-class spacing for mapped articulatory features.

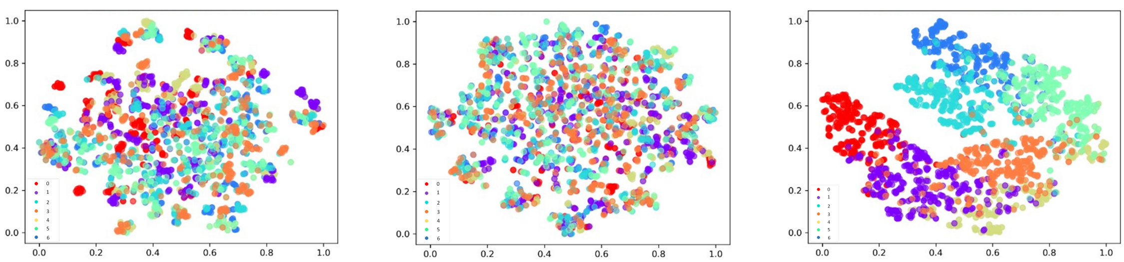


(a) Real articulatory features (b) Mapped articulatory features (c) Enhanced articulatory features

Figure 10: Visualization of articulatory features.

Fig. 11(a), (b), and (c) show the projection of the data distribution after t-SNE dimensionality reduction for real, mapped, and enhanced mapped acoustic features, respectively. Fig. 11 shows that the distribution of the acoustic data for both the real and mapped acoustic features in Fig. 11(a) and (b), respectively, is scattered and does not exhibit a clear data distribution pattern. The enhanced mapped acoustic features in Fig. 11(c) are clearly distinguishable from the articulatory features of the different emotions after t-SNE dimensionality reduction. Therefore, we can infer that KCLNet results in a significant reduction in intra-class spacing and an increase in inter-class spacing of the mapped acoustic features.

In summary, KCLNet can significantly enhance the emotional information of the mapped features and effectively address the issue of insufficient emotional information of the mapped features generated by the conversion model.



(a) Real acoustic features (b) Mapped acoustic features (c) Enhanced acoustic features

Figure 11: Visualization of acoustic features.

5.5. ResTCN-FDA network ablation experiment (RQ4)

To investigate the role of the FDA in emotion recognition, we extracted 60-dimensional MFCC features from the STEM-E^{2}VA, CASIA, RAVDESS, and EMO-DB databases as inputs for the ablation experiments. The models were evaluated in terms of accuracy, F1-score, and AUC.

Table 7: ResTCN-FDA network ablation experiments (%).

Database	**CASIA**			**STEM-E^{2}VA**			**EMO-DB**			**RAVDESS**		
Categories	6			7			7			8		
Metrics	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
TCN	70.69	69.18	70.79	64.71	64.52	66.69	71.26	68.67	71.71	59.07	59.02	59.66
ResTCN	72.25	72.16	72.67	68.31	67.68	71.14	73.71	74.28	74.83	62.41	61.15	61.77
ResTCN-FA	76.25	76.56	76.96	73.63	73.78	74.83	76.22	76.62	76.94	63.93	62.40	63.98
ResTCN-DA	73.75	73.71	73.97	72.85	72.61	73.36	77.29	75.81	77.91	64.07	63.82	64.90
Bi-A2CEmo[a]	**80.41**	**81.22**	**81.43**	**75.63**	**75.44**	**76.82**	**80.16**	**80.78**	**81.58**	**66.55**	**65.57**	**66.86**

As indicated in Table 7, when comparing the residual temporal convolution network of feature attention (ResTCN-FA) and the residual temporal convolution network of dimension attention (ResTCN-DA) to ResTCN, the accuracy of the feature attention pair is improved by 1.52%-5.32%, and the accuracy of the dimension attention pair is improved by 1.50%-4.54%. This demonstrates that assigning different weight parameters to various features and dimension channels can enhance the accuracy of emotion recognition. ResTCN-FDA shows a significant improvement in accuracy compared with TCN, ResTCN, ResTCN-FA, and ResTCN-DA. In addition, the F1-score and AUC of ResTCN-FDA are also improved to a certain extent compared with the other networks. This improvement suggests that the ResTCN-FDA network is more effective for handling emotional features.

6. Discussion

The specific effects of the potential coupling between speech and articulation on emotion recognition remain understudied. Speech, as a product of the synergistic interactions of human vocal organs, inherently involves interactions between multiple modalities during its production. After observing the coupling phenomenon between speech and articulatory waveforms (Fig. 6) inspired by the human articulation mechanism, this study proposes an emotion recognition method based on bi-directional acoustic-to-articulatory conversion. The following provides an in-depth discussion of the results of this study.

First, as illustrated in Figs. 8 and 9, the fusion of acoustic and articulatory features significantly enhances the emotion recognition rate compared with using a single modality, with varying degrees of improvement across different emotional states. Notably, the recognition rates for neutral, ecstatic, and painful emotions exhibit the most prominent increases. This finding not only underscores the pivotal role of integrating acoustic and articulatory features in emotion recognition tasks, but also demonstrates that by learning from the information present in both modalities, the system is able to capture subtleties in emotional expressions more comprehensively, thereby enhancing its recognition performance.

Furthermore, the emotion recognition rate of the bimodal fusion features generated by Bi-A2CEmo is higher than that of the fusion features recorded by EMA. This further validates our hypothesis that learning the potential bi-directional coupling relationship between the two modalities can yield higher-quality representations. Although EMA is an accurate measurement method for articulation, it is prone to interference from external noise when capturing the dynamic changes in articulatory organs. In contrast, Bi-A2CEmo can more precisely model and predict the interaction between speech and articulation by simulating the human articulatory mechanism, thereby extracting more representative emotional features.

Table 3 presents a detailed comparison of the performance of our proposed method with that of previous studies in various experimental settings. Given the diversity of experimental configurations, a direct comparison of the effectiveness of the different methods is challenging. For instance, Ref. [50] employed MSF features for four-class emotion recognition on the EMO-DB and CASIA datasets, Ref [24] utilized HSF features to achieve seven-class emotion recognition on the RAVDESS dataset, and Ref. [53] extracted MFCC features from the EMO-DB and fed them into an SVM for seven-class classification. As is evident from the data in Table 3, although our method lags slightly behind the results of Ref. [24], considering that the MFCC features we used are merely a subset of those in Ref. [24], this sufficiently demonstrates the remarkable superiority of our proposed method for emotional feature extraction.

Table 4 compares the emotion recognition effectiveness of our proposed method with that of mainstream methods in the same experimental settings. The results indicate that our method achieves the best classification performance among all listed methods. Notably, that the Bi-A2CEmo model achieves high recognition rates on multiple public datasets, which not only verifies the effectiveness of our method but also demonstrates its strong generalization ability, providing robust support for excellent performance across different datasets. This finding lays a solid foundation for the future application of the proposed method in practical scenarios.

7. Conclusion

In this study, we propose a bi-directional acoustic-articulatory conversion framework for emotion recognition. It leverages the coupling and complementarity between acoustic and articulatory signals to improve the overall performance of the SER system. Specifically, we incorporate a generative adversarial mechanism and contrast enhancement strategy into Bi-A2CEmo. Building on the generative adversarial mechanism, we propose Bi-MGAN for acoustic and articulatory conversion, effectively addressing the nonlinear ill-posedness problem in feature conversion to generate highly precise mapped acoustic and articulatory features. To address the low emotion recognition rate of the mapped features, we introduce KCLNet, which significantly enhances the mapped features by comparing emotions within and across the mapped and real features. In the emotion recognition network of Bi-A2CEmo, we introduce the FDA attention module and integrate it with ResTCN, enabling the recognition model to dynamically allocate weight coefficients to features and maximize the extraction of emotional elements. Additionally, we design and collect a database of STEM-E^{2}VA emotional speech and articulatory Mandarin to address the data gap in this research area.

The Bi-A2CEmo is based on two metrics, MAE and RMSE, for the acoustic and articulatory conversion task. Bi-MGAN conducted ablation experiments and comparison tests on the STEM-E^{2}VA dataset. The results demonstrate that the generator loss function and the bounded mapping loss function can significantly reduce the dispersion of the mapped features. The Bi-MGAN algorithm can generate mapping features with higher accuracy compared to BiLSTM, PSO-LSSVM, and DRMDN. To enhance mapping features, we propose a contrast enhancement strategy and conduct an interpretive analysis of the features before and after enhancement using a visualization algorithm. The results demonstrate that KCLNet can effectively decrease the intra-class spacing of mapped features and increase the inter-class spacing of mapped features. Finally, we have conducted extensive experiments on the proposed ResTCN-FDA using CASIA, STEM-E^{2}VA, EMO-DB, and RAVDESS datasets. This paper reveals that in acoustic and articulatory bimodal signals, the mapping feature, the true feature, and the enhanced mapping feature of each modality serve as emotional complements to the true feature of the other modality, and this complementary effect is enhanced in turn. The Bi-A2CEmo framework can not only effectively recognize the emotions embedded in articulatory signals and those embedded in parallel acoustic-articulatory signals, but also improve the recognition performance of the SER system by exploiting the coupling and complementarity of the two signals. The experimental results show that the proposed bi-directional acoustic-articulatory conversion is very effective for the study of SER.

The current study has several limitations that offer potential avenues for future exploration. Our approach relies primarily on EMA-captured acoustic-articulatory signals as the primary data source. However, the bulkiness of these devices, their high cost, and the constraints of wired sensors pose considerable challenges to the development of real-time, portable emotional recognition systems. Nonetheless, with ongoing advancements in sensor design technology, the process of collecting articulatory data will become increasingly streamlined, offering more convenient conditions for research.

CRediT authorship contribution statement

Haifeng Li: Conceptualization, Methodology, Software, Writing- Original draft preparation, Data curation, Revised paper, Writing - Review & Editing. **Xueying Zhang:** Funding Acquisition, Supervision, Data curation, Project Administration. **Shufei Duan:** Funding Acquisition, Formal Analysis, Investigation, Supervision, Project Administration, Revised paper, Writing - Review & Editing. **Huizhi Liang:** Validation, Supervision, Methodology, Revised paper, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (No.62271342), and in part by the Youth Fund of the National Nature Science Foundation of China (No.12004275).

References

[1] J. Hu, Y. Huang, X. Hu, Y. Xu, The acoustically emotion-aware conversational agent with speech emotion recognition and empathetic responses, IEEE Transactions on Affective Computing 14 (1) (2022) 17–30.

[2] W. Li, J. Xue, R. Tan, C. Wang, Z. Deng, S. Li, G. Guo, D. Cao, Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving, IEEE Transactions on Intelligent Vehicles (2023).

[3] L. Hansen, Y.-P. Zhang, D. Wolf, K. Sechidis, N. Ladegaard, R. Fusaroli, A generalizable speech emotion recognition model reveals depression and remission, Acta Psychiatrica Scandinavica 145 (2) (2022) 186–199.

[4] A. Vij, J. Pruthi, An automated psychometric analyzer based on sentiment analysis and emotion recognition for healthcare, Procedia computer science 132 (2018) 1184–1191.

[5] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern recognition 44 (3) (2011) 572–587.

[6] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[7] D. Hu, X. Hu, X. Xu, Multiple enhancements to lstm for learning emotion-salient features in speech emotion recognition, Proc. Interspeech 2022 (2022) 4720–4724.

[8] T. Song, W. Zheng, P. Song, Z. Cui, Eeg emotion recognition using dynamical graph convolutional neural networks, IEEE Transactions on Affective Computing 11 (3) (2018) 532–541.

[9] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, C.-H. Lee, Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 2617–2629.

[10] G. Ren, J. Fu, G. Shao, Y. Xun, Articulatory-to-acoustic conversion of mandarin emotional speech based on pso-lssvm, Complexity 2021 (2021) 1–10.

[11] S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, An articulatory study of emotional speech production, in: Ninth European Conference on Speech Communication and Technology, 2005.

[12] Z. Zhang, M. Huang, Z. Xiao, A study of correlation between physiological process of articulation and emotions on mandarin chinese, Speech Communication 147 (2023) 82–92.

[13] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, E. Saltzman, Accurate recovery of articulator positions from acoustics: New conclusions based on human data, The Journal of the Acoustical Society of America 100 (3) (1996) 1819–1834.

[14] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, L. Cai, A deep recurrent approach for acoustic-to-articulatory inversion, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4450–4454.

[15] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Doñas, J. L. Pérez-Córdoba, A. M. Gomez, Silent speech interfaces for speech restoration: A review, IEEE access 8 (2020) 177995–178021.

[16] S. Aryal, R. Gutierrez-Osuna, Reduction of non-native accents through statistical parametric articulatory synthesis, The Journal of the Acoustical Society of America 137 (1) (2015) 433–446.

[17] J. Kim, S. Lee, S. Narayanan, An exploratory study of the relations between perceived emotion strength and articulatory kinematics, in: Twelfth Annual Conference of the International Speech Communication Association, 2011.

[18] A. S. Shahrebabaki, G. Salvi, T. Svendsen, S. M. Siniscalchi, Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021) 135–147.

[19] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, S. Narayanan, Speaker verification based on the fusion of speech acoustics and inverted articulatory signals, Computer speech & language 36 (2016) 196–211.

[20] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, C. Di Natale, Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure, Knowledge-Based Systems 63 (2014) 68–81.

[21] A. Bhavan, P. Chauhan, R. R. Shah, et al., Bagged support vector machines for emotion recognition from speech, Knowledge-Based Systems 184 (2019) 104886.

[22] A. I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowledge-Based Systems 244 (2022) 108580.

[23] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, C. Di Natale, Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure, Knowledge-Based Systems 63 (2014) 68–81.

[24] B. T. Atmaja, M. Akagi, On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers, in: 2020 ieee region 10 conference (TENCON), IEEE, 2020, pp. 968–972.

[25] S. Kwon, et al., Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach, Expert Systems with Applications 167 (2021) 114177.

[26] Y. Cheng, Y. Xu, H. Zhong, Y. Liu, Hs-tcn: A semi-supervised hierarchical stacking temporal convolutional network for anomaly detection in iot, in: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), IEEE, 2019, pp. 1–7.

[27] D. Le, E. M. Provost, Emotion recognition from spontaneous speech using hidden markov models with deep belief networks, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, 2013, pp. 216–221.

[28] L.-Y. Liu, W.-Z. Liu, J. Zhou, H.-Y. Deng, L. Feng, Atda: Attentional temporal dynamic activation for speech emotion recognition, Knowledge-Based Systems 243 (2022) 108472.

[29] Z.-H. Ling, K. Richmond, J. Yamagishi, R.-H. Wang, Integrating articulatory features into hmm-based parametric speech synthesis, IEEE Transactions on Audio, Speech, and Language Processing 17 (6) (2009) 1171–1185.

[30] Z.-C. Liu, Z.-H. Ling, L.-R. Dai, Statistical parametric speech synthesis using generalized distillation framework, IEEE Signal Processing Letters 25 (5) (2018) 695–699.

[31] S. Aryal, Statistical parametric methods for articulatory-based foreign accent conversion, Ph.D. thesis (2015).

[32] M.-A. Georges, L. Girin, J.-L. Schwartz, T. Hueber, Learning robust speech representation with an articulatory-regularized variational autoencoder, arXiv preprint arXiv:2104.03204 (2021).

[33] S. Zhang, X. Zhao, Q. Tian, Spontaneous speech emotion recognition using multiscale deep convolutional lstm, IEEE Transactions on Affective Computing 13 (2) (2019) 680–688.

[34] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, L. Cai, A deep recurrent approach for acoustic-to-articulatory inversion, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4450–4454.

[35] C. Qin, M. Á. Carreira-Perpiñán, An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping, in: Eighth Annual Conference of the International Speech Communication Association, 2007.

[36] J. Kim, P. Ghosh, S. Lee, S. S. Narayanan, A study of emotional information present in articulatory movements estimated using acoustic-to-articulatory inversion, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, 2012, pp. 1–4.

[37] D. Erickson, C. Zhu, S. Kawahara, A. Suemitsu, Articulation, acoustics and perception of mandarin chinese emotional speech, Open Linguistics 2 (1) (2016).

[38] K. Li, L. Wu, Q. Qi, W. Liu, X. Gao, L. Zhou, D. Song, Beyond single reference for training: underwater image enhancement via comparative learning, IEEE Transactions on Circuits and Systems for Video Technology (2022).

[39] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, Q. Zheng, Learning multi-scale features for speech emotion recognition with connection attention mechanism, Expert Systems with Applications 214 (2023) 118943.

[40] W. Zhu, X. Li, Speech emotion recognition with global-aware fusion on multi-scale feature representation, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6437–6441.

[41] H. Zou, Y. Si, C. Chen, D. Rajan, E. S. Chng, Speech emotion recognition with co-attention based multi-level acoustic information, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 7367–7371.

[42] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1d & 2d cnn lstm networks, Biomedical signal processing and control 47 (2019) 312–323.

[43] T. Anvarjon, Mustaqeem, S. Kwon, Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features, Sensors 20 (18) (2020) 5212.

[44] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia 20 (6) (2017) 1576–1590.

[45] J. Yuan, C. Bao, Cyclegan-based speech enhancement for the unpaired training data, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2019, pp. 878–883.

[46] B.-H. Su, C.-C. Lee, Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan, IEEE Transactions on Affective Computing (2022).

[47] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al., A database of german emotional speech., in: Interspeech, Vol. 5, 2005, pp. 1517–1520.

[48] C. A. o. S. Institute of Automation, Caisa mandarin emotional speech corpus (2005).

[49] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (5) (2018) e0196391.

[50] X. Li, M. Akagi, Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model, Speech Communication 110 (2019) 1–12.

[51] M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images, IEEE Transactions on Information Forensics and Security 14 (5) (2018) 1181–1193.

[52] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Transfer learning for improving speech emotion classification accuracy, arXiv preprint arXiv:1801.06353 (2018).

[53] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, M. Hao, Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence, Information Sciences 563 (2021) 309–325.

[54] P. Singh, M. Sahidullah, G. Saha, Modulation spectral features for speech emotion recognition using deep neural networks, Speech Communication 146 (2023) 53–69.

[55] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools and Applications 78 (2019) 3705–3722.

[56] P. Jiang, H. Fu, H. Tao, Speech emotion recognition using deep convolutional neural network and simple recurrent unit., Engineering Letters 27 (4) (2019).

[57] H. Zhang, H. Huang, H. Han, A novel heterogeneous parallel convolution bi-lstm for speech emotion recognition, Applied Sciences 11 (21) (2021) 9897.

[58] S. Mao, P. Ching, T. Lee, Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition., in: Interspeech, 2019, pp. 1686–1690.