Journal Pre-proof

Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network

Zijun Yang, Shi Zhou, Lifeng Zhang, Seiichi Serikawa

PII: S2667-2413(23)00041-1
DOI: https://doi.org/10.1016/j.cogr.2023.12.001
Reference: COGR 75

To appear in: *Cognitive Robotics*

Received date: 28 October 2023
Revised date: 3 December 2023
Accepted date: 3 December 2023

Please cite this article as: Zijun Yang, Shi Zhou, Lifeng Zhang, Seiichi Serikawa, Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network, *Cognitive Robotics* (2023), doi: https://doi.org/10.1016/j.cogr.2023.12.001

[COR: Number]

Highlights

Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network

Zijun Yang,Shi Zhou,Lifeng Zhang,Seiichi Serikawa

- Innovative approach: Hilbert curves for efficient data conversion in speech emotion recognition
- Efficient feature extraction: Hilbert curves, neural networks, and reduced costs for deep learning
- Cost savings: New data storage method lowers computational and storage expenses
- Transformative technique: Hilbert curves enhance speech emotion recognition efficiency
- Effective support: Hilbert curve approach improves emotion classification accuracy

Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network

Zijun Yang[a], Shi Zhou[b], Lifeng Zhang[a] and Seiichi Serikawa[a]

[a]*organization=Kyushu Institute of Technology, addressline=1-1 Sensuicho, Tobata Ward, city=Kitakyushu, postcode=8040011, state=Fukuoka, country=Japan*

[b]*organization=Huzhou College, addressline=No.1, Bachelor Road, Wuxing District, city=Huzhou, postcode=8040011, state=Zhejiang, country=China*

ARTICLE INFO

Keywords:
speech emotion recognition
hilbert curvert
deep learninng
speech features

ABSTRACT

In the realm of speech emotion recognition, researchers strive to refine representation methods for improved emotional information capture. Traditional one-dimensional time series classification falls short in expressing intricate emotional patterns present in speech signals, posing challenges in accuracy and robustness. This study introduces an innovative algorithm leveraging Hilbert curves to transform one-dimensional speech data into two-dimensional form, enhancing feature extraction accuracy. A tiling module based on Hilbert curve maximizes Hilbert curve arrangements for improved emotional information capture. Results reveal spatial efficiency gains up to 23,195 times pixel units, enhancing data storage. With an exceptional 98.73% accuracy, the proposed approach traditional methods, affirming its superior emotion classification performance on the same dataset. These empirical findings underscore the effectiveness of our proposed method in advancing speech emotion recognition.

1. Introduction

Emotions play a crucial role in human life [1] and are integral to psychological survival [2]. Emotion detection research is evolving through collaborative research in the fields of psychology, cognitive science, machine learning, and natural language processing (NLP) [3]. In early studies, scientists such as Darwin [4] considered emotional expression to be the last behavioral pattern preserved in human evolution. Speech signaling is an important mode of emotional expression, accounting for 38% of emotional communication [5]. This function plays a crucial role in emotion recognition and communication [6]. Speech emotion recognition (SER) is a branch of emotion detection [7]. It has more than two decades of research history and has accumulated numerous results [8]. This subject focuses on the recognition of emotions in speech without considering the semantic content [9, 10]. Speech signals contain many acoustic features that can reflect the emotional state of the speaker as well as information related to the speaker and speech. Therefore, the central concept of emotion detection is to study the acoustic differences produced when speech is vocalized in different emotional contexts [11].

SER has a wide range of applications in emotion-related information. For example, speech emotion recognition provides important assistance in areas such as healthcare, education, and social interaction [12–14]. It helps to monitor stress-induced changes in mental health [15], conduct mental health assessments [16], and quickly predict depression severity [17]. Speech emotion recognition plays an active role in improving decision–making in the tourism industry [18, 19]. In addition, it helps to collect and analyze people's opinions and impressions of various topics, products, themes, and services, improving their sense of well-being and satisfaction, thus contributing to social stability and harmony [19] [20].

The convergence of SER with real–life applications enhances the close connection between human and computer interactions. By enabling computer systems to recognize and respond to human emotions, SER offers the possibility of more intuitive and emotionally rich human-computer interactions in various domains

Corresponding author: Zijun Yang
zijun.yang529@mail.kyutech.jp (Z. Yang)
ORCID(s):

[13]. For example, in 2017, [21] proposed a flexible emotion recognition system that utilizes visual and auditory input analyses. In 2020 [22], three models or experts were integrated using integrated learning, focusing on different feature extraction and classification strategies. Other studies [23] have employed rectangular filters and improved pooling strategies to create lightweight SER models. They used a CNN approach to learn deep frequency features and trained CNN models for sentiment assessment using the frequency features extracted from speech data. Additionally, a new SER approach is based on the bidirectional attentional long short-term memory (BLSTMwA) model and deep convolutional neural networks (DCNN) [24]. Integrated classifiers have also been created using deep convolutional recurrent neural networks, specifically for SER [25]. These studies enrich knowledge in the field of SER and provide more options for practical applications.

However, one of the main challenges in the field of SER is the extraction of relevant features from speech signals to recognize emotional states [26, 27] and to develop appropriate classifiers [10, 28]. Speech feature extraction is considered to be a key issue in speech emotion recognition systems. Many studies have proposed a variety of speech features such as pitch, energy, frequency, linear predictive resonance frequency coefficients (LPCC), MFCC, and modulation spectrum features that reflect the speaker's emotional information [29–31]. Therefore, many studies have combined multiple types of affective features to characterize available speech signals adequately. However, combining multiple affective features increases the dimensionality and redundancy of speech data, thereby increasing the learning difficulty of most machine-learning algorithms and the risk of overfitting [1]. Against this backdrop, two studies explored new approaches to combat this problem. Wang et al. proposed a compelling approach for transforming a one-dimensional speech time series into more informative two-dimensional image data [32]. This study utilizes the comprehensive and versatile nature of computer vision to improve the efficiency and accuracy of speech emotion recognition. Its innovation lies in presenting speech signals in a novel manner, thereby providing more possibilities for sentiment analysis. Another study utilized the quasiperiodic nature of speech signals to transform one-dimensional speech data into more information-rich two-dimensional data by periodically segmenting speech signals [33]. The uniqueness of this approach is that it exploits the periodic nature of speech signals to improve the emotion recognition performance. These two results introduce new ideas and methods in the field of speech sentiment recognition that are expected to improve the accuracy and practicality of sentiment analysis. However, Wang et al. and Bakhshi et al. proposed two methods that present unique challenges and complexities in the SER field. In Wang et al.'s approach, utilizing the Gram angle field for converting one-dimensional speech signals into two-dimensional images introduces challenges in terms of the accuracy and robustness of the transformation. On the other hand, Bakhshi et al.'s method, which is based on the periodicity of the speech signal, faces challenges related to the generalizability of the conversion technique.

Building on these foundations, our study proposes a new method that addresses the challenges associated with standardization, calibration, interpretability, and generalizability of emotion recognition from speech signals. This method converts one-dimensional time series data into more informative two-dimensional images. This was inspired by converting one-dimensional time-series data into a two-dimensional image with the structure and shape of a Hilbert curve. Subsequently, a convolutional neural network was utilized to extract the features of the image, and the extracted feature values were passed through a fully connected network for sentiment classification. It is worth mentioning that we compared the performance of this method with that of previous studies ([32] and [33]). This comparison validates the feasibility and effectiveness of the proposed algorithm in the field of emotion recognition. The results of this study show that the proposed algorithm has potential advantages in emotion-recognition tasks. The innovative aspect of this study is that it not only provides a new method for processing one-dimensional time-series data, but also demonstrates the potential application of this method in emotion recognition. By presenting time-series data as two-dimensional images, we open up more possibilities in the field of sentiment analysis, with promising improved accuracy and performance. The success of this approach demonstrates the potential value of deep learning and computer vision in sentiment recognition.

The remainder of this paper is organized as follows. Section 2 briefly describes the algorithms used in the study. Section 3 outlines the methodology and provides details of the specific experimental procedures. Section 4 presents and discusses the results obtained from the experiments. Finally, Section 5 concludes the paper.

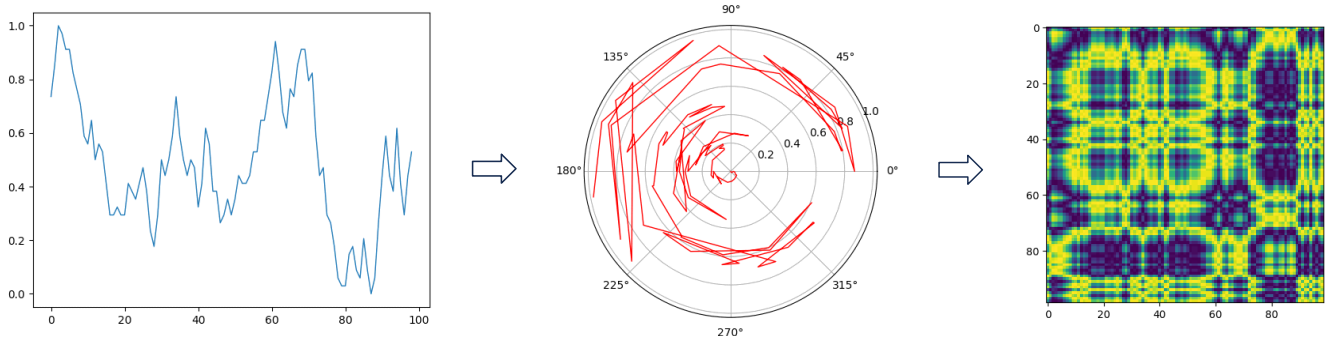


Fig. 1: Schematic diagram of the GAF algorithm. The time series are converted from a right-angle coordinate system to a polar coordinate system by transformation, and the GAF image is then generated by Eq 1.

2. Related work

Comprehending the current landscape of methodologies for emotion recognition is pivotal for advancing innovative solutions. In this section, we delve into the work of Wang et al. [32] and Bakhshi et al. [33], shedding light on their respective contributions and the challenges encountered in their proposed approaches.

2.1. GAF

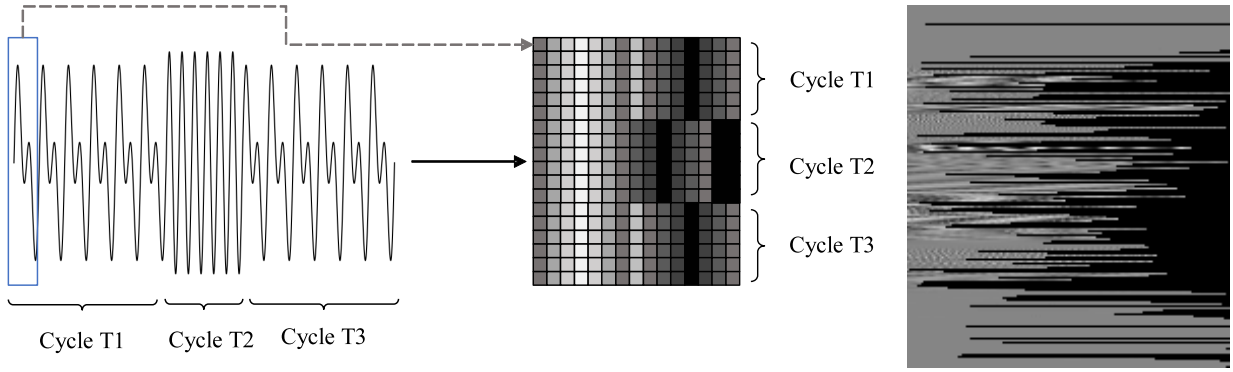
Wang et al. [32] introduced a novel methodology for converting one-dimensional speech signals into two-dimensional images, leveraging Gram Angle Fields (GAF). The GAF-based approach offers a unique perspective for visualizing and analyzing speech patterns, providing valuable insights into emotional expressions.

The GAF algorithm is a time–series data coding method that extends the framework proposed by Campanharo et al. [34] and aims to preserve the time-domain information [32]. For a time series X containing real–valued observations $(x_1, x_2, ..., x_n)$, the algorithm uses the rescaling method to normalize the time series X to the interval [0,1] or [-1,1]. Subsequently, the algorithm uses polar coordinates to represent the rescaled time series $\widetilde{X}$, mapping values to cosine angles and timestamps to radii. This innovative representation provides a new perspective for understanding time series. As time progressed, the values between the different angular points on the polar coordinates changed. Unlike the traditional Cartesian coordinate system, the polar coordinates retain absolute temporal relationships. Different rescaled data points corresponded to different angular boundaries. For example, [0, 1] corresponds to the cosine function of $[0, \pi/2]$, whereas the cosine value of [-1, 1] lies on the $[0, \pi]$ angular boundary.

After converting the time series to polar coordinates, the algorithm can easily recognize the temporal correlation between different time intervals and calculate the triangular sum/difference between points using Eq (1). One is to generate a two–dimensional image from a one-dimensional time series, as shown in Fig. 1. The generalized autocorrelation matrix (GAF) of the GAF algorithm has several advantages: firstly, it preserves temporal dependence; secondly, it preserves temporal correlation; and finally, it provides advanced time-series features for deep neural network learning. This approach has a wide range of applications in time-series analyses.

$$G(x_1, \cdots, x_n) = \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \tag{1}$$

Despite its innovation, Wang et al.'s approach faces challenges in terms of emotion recognition. The accuracy and robustness of the conversion process, particularly the dependence on GAF for capturing essential emotional features, present hurdles. The challenge lies in ensuring that the two-dimensional representation retains relevant emotional cues while minimizing information loss.



(a) CyTex Basic Principle. (b) CyTex imaging result.

Fig. 2: Schematic diagram of the CyTex algorithm. In Fig. 2(b) Time series with three cycles: T1, T2, and T3. Each cycle is represented as a row in the image, as shown in Fig. 2(a). The blue box in 2(a) represents the first cycle of T1, which occupies a row when converted to an image. This process was repeated six times for cycles T1, T2, and T3 so that the image texture remained consistent in columns 1-6, 7-12, and 13-18 of the image. For cycle T2, the number of data points is less than that of cycles T1 and T3; hence, there is a 0-fill in the image. Figure 2(b) shows the CyTex image corresponding to a segment of speech data from the experimental dataset.

Bakhshi et al. proposed the CyTex method, which focuses on leveraging the periodicity of speech signals for emotion recognition. This approach introduces a different perspective by utilizing the inherent speech signal patterns to extract emotional features.

2.2. CyTex

The CyTex algorithm is designed to transform speech signals into structured images [33]. Speech signals usually exhibit a quasi-periodic character, where neighboring speech cycles vary slightly in shape and duration. These subtleties, differences, and similarities reflect the amplitude and frequency variations of the speech signal, which are crucial for expressing features such as emotions. In the CyTex transform, each cycle in the speech signal is represented as a row in the image such that successive speech cycles form successive rows of the output image, as shown in Fig. 2. This approach analyzes a speech signal based on its fundamental cycles, each of which forms a row in the resulting image. The goal of the CyTex algorithm is to map the speech signal to an image in this way to better represent and analyze its speech characteristics.

Bakhshi et al.'s CyTex method confronts challenges related to the generalizability of the conversion technique. The periodicity of speech signals may not universally capture the diverse emotional patterns present across individuals and cultural contexts. Adapting this method to accommodate variations in speech characteristics and emotional expressions poses a significant challenge.

2.3. Proposed Approach

This study aimed to address the challenges identified in the methodologies of Wang et al. and Bakhshi et al. by introducing an innovative and targeted approach to emotion recognition.

This paper proposes an innovative method for converting a one-dimensional time series into two-dimensional images. Specifically, the method involves mapping the speech signal onto an image following the trajectory of the Hilbert curve by utilizing the arrangement of the Hilbert curve path. The feature values of the resulting images were extracted through a convolution operation. These eigenvalues were then flattened based on the Hilbert curve arrangement method, yielding a one-dimensional vector. Subsequently, a classifier was employed to categorize the vector, effectively achieving the objective of speech emotion recognition. This methodology capitalizes on the structural properties of the Hilbert curve, providing a unique and efficient means of representing and analyzing speech signals for emotion recognition.

Methods using Hilbert curves have certain advantages when dealing with time-series classification problems and can solve or avoid some of the challenges and difficulties mentioned previously. The proposed methodology refrains from downsampling, thus ensuring the preservation of crucial information. Addition-

ally, the challenges associated with the periodicity of speech signals were overcome by introducing a novel flattening method along the path of the Hilbert curve. This preserves the temporal correlations in the feature values, providing a unique solution to the challenges faced by prior studies. By avoiding downsampling and introducing a novel flattening technique, this approach enhances the conversion accuracy and generalizability. These innovations marked significant improvements in addressing challenges within the emotion recognition domain.

3. Method and experiment

In this section, we present the methodology and experiments, focusing on a comprehensive analysis of a one–dimensional time series using computer vision techniques.

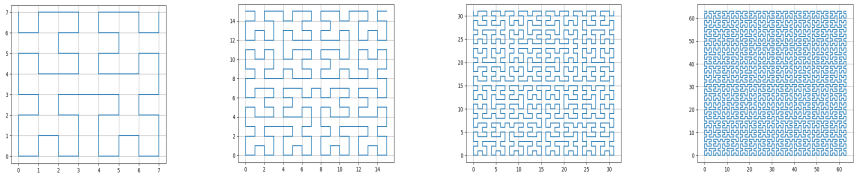
3.1. Approach

This study introduced an innovative method for transforming a one-dimensional time series into two-dimensional images. Specifically, this approach involves arranging the signal along the path of the Hilbert curve and mapping the speech signal onto an image that follows the trajectory of the Hilbert curve. The purpose of using Hilbert curves are used to reduce information loss and thus improve the accuracy of the conversion from 1D time series to 2D images. By mapping the speech signal onto the trajectory of the Hilbert curve, the temporal correlation was preserved, making the new representation more capable of capturing the temporal correlation and dynamic features in the speech signal.

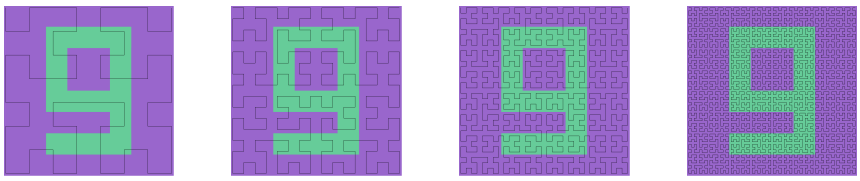
The Hilbert curve, proposed by German mathematician David Hilbert in 1891 [35], is a special type of space-filling curve that has a wide range of applications. This curve can fill any bounded two-dimensional space and is unique in that it maintains the locality of neighboring points, making neighboring points also neighboring the curve, which facilitates localized access to spatial data. As shown in Fig. 3, the image represents the morphology of Hilbert curves of different dimensions in two dimensions. In addition, Hilbert curves can map multidimensional data onto one-dimensional curves, improving the efficiency of indexing and searching the data, and having a nested structure that allows for resolution adjustments as needed. The image representation utilizes Hilbert curves to achieve a mapping of a two-dimensional image onto a one-dimensional curve as shown in Fig. 3(b). Such curves are also used in image compression and coding, in which pixels are arranged in the order of the curve to achieve a compact representation of the data. Thus, Hilbert curves play a key role in the fields of geographic information systems, image processing, database query optimization, and distributed computing and are particularly suitable for processing spatial and multidimensional data.

In this study, two innovative proposals based on the mapping and inverse mapping of Hilbert curves are presented to achieve an efficient conversion of 1D time series to 2D images.

1. **Conversion from one to two dimensions**: In this study, an innovative one-dimensional to two-dimensional upscaling method is proposed through the mapping of Hilbert curves. The overall conversion pipeline from one to two dimensions is illustrated in Fig. 5. This experiment involved treating speech data as grayscale pixel values in an image and plotting them according to the order of the Hilbert curve, creating an ordered image, as shown in Fig. 6. The image dimensions were set to (512, 512), determined by the minimum size required to encompass all the speech data with a dimension value of 9. For cases in which the data are insufficient to fill the entire image, two padding methods are employed: zero-padding and repetitive data read for padding, with excess portions discarded. Ultimately, this process provides an innovative approach to representing speech information in the form of an image, leveraging the ordered nature of the Hilbert curve. The flexible handling of padding accommodates different image sizes without compromising the integrity of the information.
2. **Conversion from two to one dimension**: In order to restore the information effectively, this study also proposes an innovative 2D to 1D dimensionality reduction method based on the Hilbert curve. The dimensionality reduction process from 2D to 1D is realized by inverse mapping, that is, rearranging the 2D image onto the trajectory of the Hilbert curve. The images processed by the convolutional neural network consist of a set of feature maps with multiple channels, exemplified here by three channels, as shown in Fig. 7(a). These feature maps are unfolded along the path of the Hilbert curve and concatenated in channel order, forming a one-dimensional array, as illustrated in Fig. 7(b). This



(a) Original shapes of Hilbert curves of dimensions 3, 4, 5, and 6.



(b) Hilbert curves represent the localization of neighboring points on a two-dimensional image 1.

(c) Hilbert curves represent the localization of neighboring points on a two-dimensional image 2.

Fig. 3: The two-dimensional shape of the Hilbert curve. Fig. 3(a) shows the original shape of the Hilbert curve in dimensions 3, 4, 5, 6. Fig. 3(b) and Fig. 3(c) shows how the Hilbert curve represents the localization of neighboring points on a two-dimensional image. It can be observed from the figure that the higher the dimension of the dimension of the Hilbert curve, the richer the data between neighboring points, and the better the curve can describe the image.

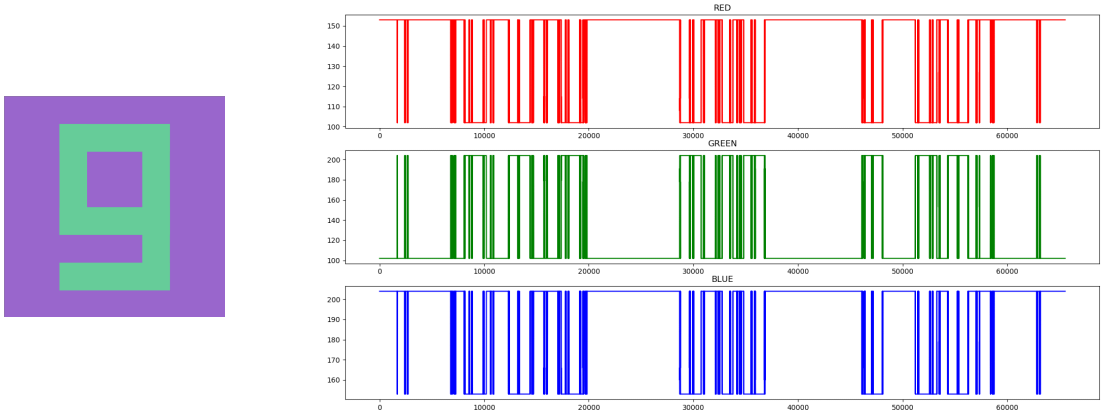
unfolding method better preserves the temporal correlation of speech data, ensuring that the order of the features in the one-dimensional array aligns with their temporal positions in the original image. This Hilbert curve-based transformation from two-dimensional to one-dimensional effectively maintains the temporal structure while considering the temporal correlation of the multichannel feature maps produced by the convolutional neural network. It provides a more accurate and ordered representation for the further analysis of speech data.

These two proposals aim to optimize the data representation and reduction process to meet the high demands of accuracy and effectiveness in the field of emotion recognition. By incorporating the unique mathematical properties of Hilbert curves, these proposals aim to provide a more informative representation of speech signals, thus providing a more reliable and powerful tool for emotion-recognition tasks.

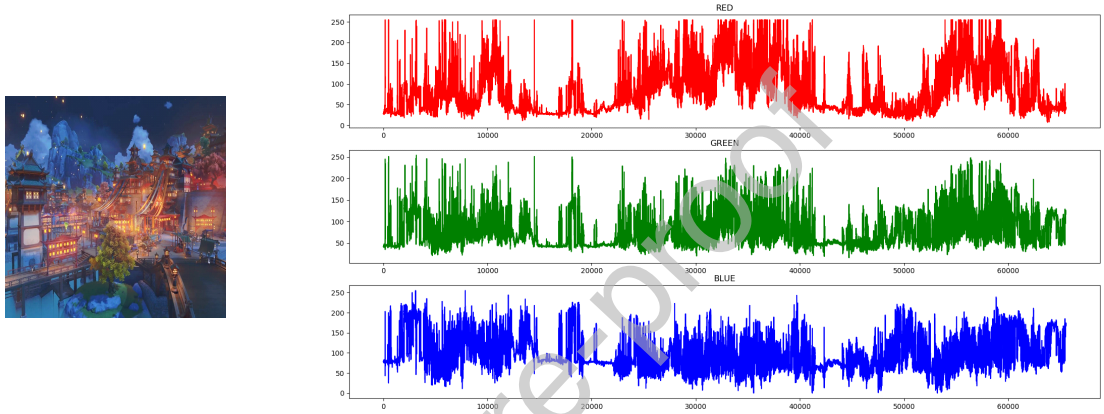
3.2. Experiments

Initiating with the preprocessing of the speech signal, the experimental workflow transformed the preprocessed data into Hilbert curves. Subsequent phases involved in-depth analysis employing a CNN, ultimately yielding conclusive classification results.

The study utilized the CASIA Emotion Corpus, which features recordings from four professional speakers reciting 50 unique sentences. Each sentence was associated with one of the six emotions: anger, happiness, fear, sadness, surprise, and neutrality. The corpus contains 1200 sentences, offering a valuable resource for analyzing the acoustic and prosodic attributes associated with various emotional states. These sentences vary in length, spanning five, six, and eight characters, respectively. To facilitate this study, the corpus was



(a) Waveforms of RGB image 1 of Hilbert curve projections.



(b) Waveforms of RGB image 2 of Hilbert curve projections.

Fig. 4: The Hilbert curve represents multidimensional data on a one-dimensional curve. Fig. 4(a) corresponds to Fig. 3(b), showing how the Hilbert curves can be transformed into each other in one or two dimensions. For ease of understanding, the experiment split the two-dimensional image according to the RGB three-color channel. Using the Hilbert curve, each channel is separately mapped from a higher-dimensional space to a one-dimensional space. The waveform in the projection shows the R, G, and B colors of the two-dimensional image according to the colors. The horizontal coordinate represents the number of sampling points and the vertical coordinate represents the size of the color pixel value. The number of sampling points determined the dimensions of the Hilbert curve. In this figure, every 20 pixel points was sampled to obtain the final waveform image. Fig. 4(b) corresponds to Fig. 3(c) shows a Hilbert waveform graph corresponding to a normal image.

randomly divided into a training set (80%) and test set (20%), as outlined in Table 1. Speech recordings were conducted within a controlled studio environment, free from external noise, with a sampling frequency of 16000 Hz and PCM storage format of PCM, 16-bit. The emotion labels were provided by a recording unit [36].

Table 2 lists the dataset and device information used in the experiment. The experimental procedure is illustrated in Fig. 8, which explained the method and steps of the experiment. This method fully utilizes the potential of computer vision and deep learning techniques for one-dimensional time-series data analysis.

In the preprocessing stage, the experiment first normalizes the speech signal between [0-1], and then regularizes it to make its value range between [0-255] to meet the requirement of an 8-bit image. Next, the one–dimensional data are converted to two–dimensional data according to the properties of the Hilbert curve. The relationship between the dimensions of the Hilbert curve and length of the time series X is presented in Table 3. For speech data of length 2^{2n}, the dimension of the Hilbert curve is typically $(2^n, 2^n)$, which is converted into a Hilbert curve $2^n \quad 2^n$. For speech data with lengths between $2^{2(n*1)}$ and 2^{2n}, two

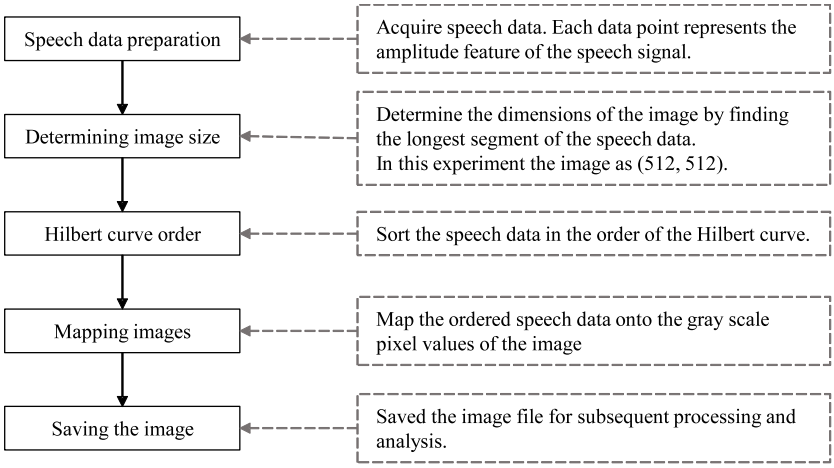
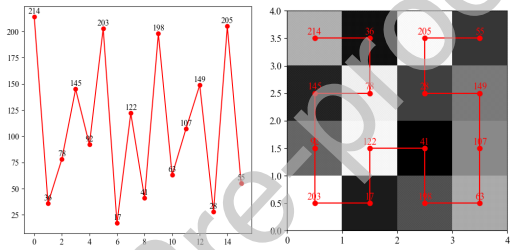


Fig. 5: Illustration of conversion from one to two dimensions. This experiment transforms speech data into an ordered image by mapping grayscale pixel values according to the Hilbert curve, with a fixed image size of (512, 512) determined by the longest segment of speech data, utilizing different padding techniques to handle incomplete regions.



(a) One-dimensional time series before conversion.

(b) 2-dimensional image after conversion.

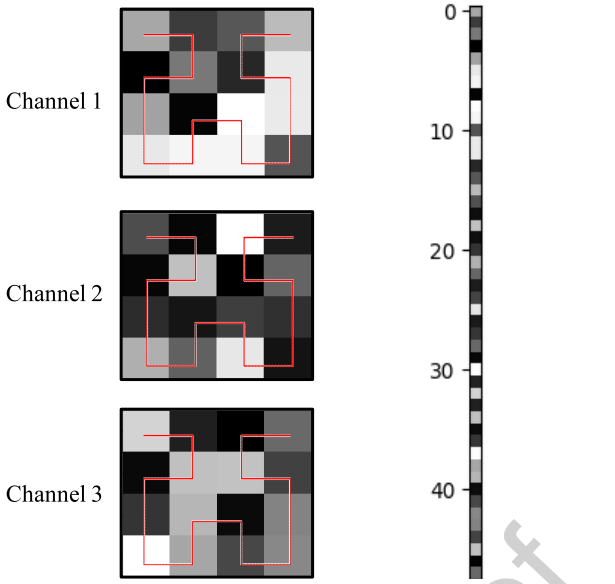
Fig. 6: illustrates the conversion of one-dimensional speech data into two-dimensional images. This example illustrates a time series of length 16, as illustrated in Fig. 6(a), where the sequence values are arranged as grayscale pixel values in accordance with the Hilbert curve's order, resulting in a two-dimensional image as shown in Fig. 6(b). To provide a clearer representation of the image, we annotated the pixel values at the corresponding positions and depicted the Hilbert curve path. This process aims to articulate the mapping of speech data to an image in a more academically nuanced and fluent manner.

Table 1
Composition of the corpus.

Number of words	5 words	6 words	8 words	Total
Sentences numbers	192	960	48	1200
Train set numbers	144	768	24	960
Test set numbers	48	192	24	240

different methods of data supplementation are used to ensure that the Hilbert curve dimension requirement is satisfied. Method one was to use for 0–padding, while method two was to iterate the data repeatedly to satisfy the dimensionality requirement. The method of complementing zeros was selected primarily because of the translation invariance characteristic of the convolution operation [37].

The experimental results are presented in Fig. 9. Fig. 9(a) represents the portion filled using 0, whereas Fig. 9(b) represents the speech signal replicated iteratively to satisfy the Hilbert curve dimensionality



(a) One-dimensional time series before conversion. (b) 2-dimensional image after conversion.

Fig. 7: Illustrate the conversion of two-dimensional speech data into one-dimensional images.

Table 2
Experimental environment and parameter configuration.

configuration item	**Numeric/Descriptive**
Programming Language and Version	Python 3.11
Deep Learning Framework	PyTorch
Libraries *	cv2, hilbertcurve, matplotlib, numpy, Scikit-image, torch, wave
Epochs	100
Training Rate	0.8
Optimizer	Adam

* The libraries listed in the table follow Python's naming conventions, are sorted alphabetically, and can be installed using the *-pip* command.

requirement. These generated images were feature-extracted using a convolutional neural network and then classified using a fully connected layer.

For comparison, different neural networks were used in the experiments to extract the feature values. One is the ResNet network model and the other is the experimentally proposed Hilbert-CNN network model, as shown in Fig. 10. Fig. 10(a) shows a flowchart of the network, using the Hilbert curve as the unfolding layer. The network was first convolved thrice to obtain a feature map. The feature map is then unfolded using a Hilbert curve. The flattened feature map passes through three fully connected layers to extract the feature values. Finally, the classification results were obtained. The data related to convolution are shown in Fig. 10(b). The size of the convolution kernel was (4, 4) and the step size was 4. The size of the input image was (1, 512, 512) and the sizes of the convolved images was (256, 8, and 8). Classification was performed by unfolding the Hilbert curve and inputting the fully connected layer. Because of the specificity of Hilbert CNN, it is only used as a partial parameter for Hilbert imaging.

The entire experimental flow is shown in Fig. 8, which helps us better understand the methods and steps

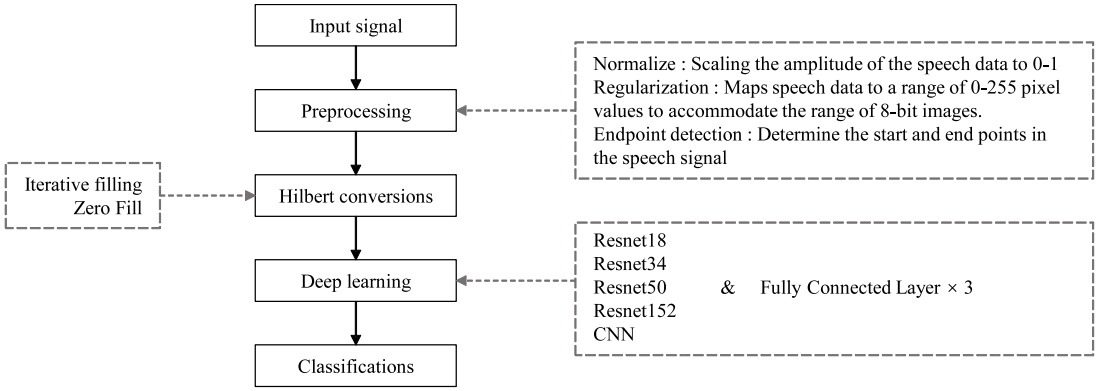
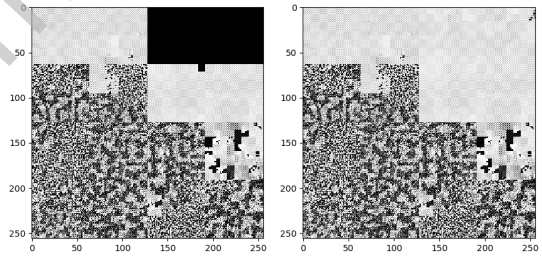


Fig. 8: The flowchart of the experiment.

Table 3
Hilbert Curve Dimension and Time Series Length Relationship.

Dimension	Length of Time Series[1]	Image Size ((h, w))[2]	Speech Duration (sec)[3]
5	1024	(32, 32)	0.064
6	4096	(64, 64)	0.256
7	16384	(128, 128)	1.024
8	65536	(256, 256)	4.096
9	262144	(512, 512)	16.384

[1] Speech Duration (sec) $= \frac{Number\ of\ Data\ Points}{Sampling\ Frequency}$, the dataset was sampled at 16,000Hz.
[2] Image size $=(2^n, 2^n)$,n is the dimension. h denotes the height of the image, w denotes the width of the image
[3] Legnth $= H \quad W = (2^n)^2$ Length of time series indicates the maximum number of time series points that can be included.

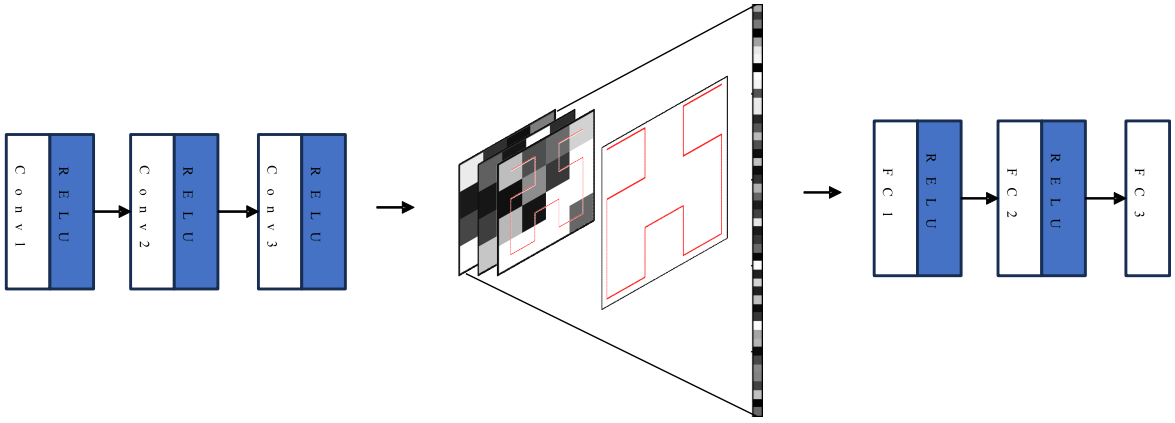


(a) Hilbert imaging algorithm with 0-filling.

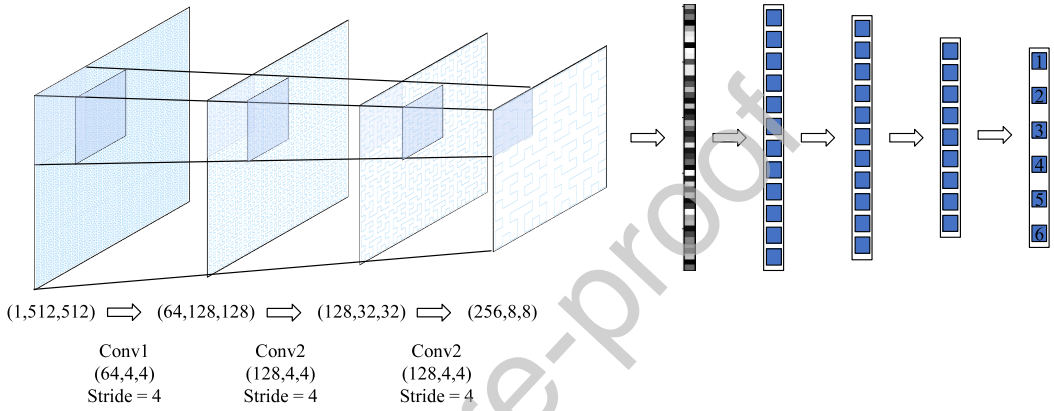
(b) Hilbert imaging algorithm with iterative padding.

Fig. 9: The result of Hilbert imaging algorithm.

of the experiment. This method fully utilizes the potential of computer vision and deep learning techniques for one-dimensional time-series data analysis.



(a) Hilbert-CNN network flow chart.



(b) Hilbert-CNN network flow chart schematic.

Fig. 10: Hilbert-CNN network model structure.

4. Result and discussion

In this section, we discuss different methods for converting one-dimensional data into two-dimensional images, and the performance of these methods is compared and analyzed in detail. The method used in this study is primarily based on the properties of the Hilbert curve, which can convert one-dimensional time series data into a two-dimensional image form with more information. By augmenting the training dataset with 1.5 × acceleration and 0.5 × deceleration, the aim is to enhance the model's robustness and generalization to diverse speech features, encompassing varying speeds and speech rates. This expansion seeks to capture and process the nuanced characteristics of real-world speech more comprehensively. Accuracy, which is the ratio of the number of samples correctly predicted by the model to the total number of samples in the classification problem, was used as the evaluation index to assess the effectiveness of this method, as shown in Eq. (2). Usually expressed as a percentage, it indicates the proportion of all samples that the model correctly classifies. We conducted a comparative study using related methods. In Table 4, bold indicates the data with the highest accuracy for each method. The tabulated results unmistakably revealed a significant enhancement in accuracy attributed to the utilization of the images generated by the Hilbert curve.

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}} \quad 100\sim \tag{2}$$

As shown in Table 5, compared with emotion speech recognition methods on the same dataset, the approach proposed in this study exhibits significant advantages. The bold font indicates the most accurate data. Based on the comparative analysis in this study, the proposed emotional speech recognition method

Table 4
Accuracy comparison of different 1D data to 2D image conversion methods(%).

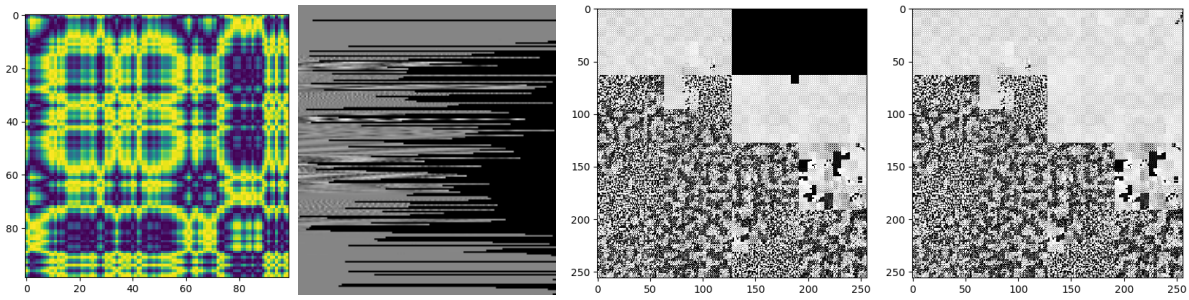
Algorithms	Network					
	Lr	Resnet18	Resnet34	Resnet50	Resnet153	Hilbert-CNN
GAF	1e-03	67.47	73.02	63.25	73.10	-
	1e-04	**90.32**	40.83	80.01	87.51	
	1e-05	90.00	34.58	81.32	90.01	
	1e-06	83.30	87.31	81.03	87.31	
	1e-07	74.37	92.38	81.73	82.13	
CyTex	1e-03	64.11	65.74	67.37	68.37	-
	1e-04	80.31	82.31	67.66	81.76	
	1e-05	82.94	84.53	85.34	86.31	
	1e-06	87.03	87.24	**87.42**	79.82	
	1e-07	75.22	73.01	70.28	73.50	
Hilbert	1e-03	73.58	76.29	80.14	80.01	81.91
	1e-04	80.09	81.14	81.79	80.13	84.08
	1e-05	86.17	88.06	89.37	87.16	90.13
	1e-06	94.16	95.21	95.19	92.13	**95.95**
	1e-07	82.93	84.19	88.16	87.97	89.01
Hilbert(0-filled)	1e-03	76.35	82.43	83.39	73.12	88.01
	1e-04	73.21	81.52	82.09	73.14	88.79
	1e-05	85.12	92.31	93.35	94.04	**98.73**
	1e-06	92.15	94.46	95.09	96.09	97.50
	1e-07	83.24	82.76	85.31	83.19	83.27

Table 5
Accuracy comparison of same dataset methods(%).

Method	Description	Accuracy
DBN & SVM [38]	Deep Belief Network combined with a Support Vector Machine for feature extraction and classification.	95.80%
ELM decision tree [39]	Extreme Learning Machine decision tree approach for emotion recognition in speech signals.	89.60%
GAF	The GAF methods transform one-dimensional speech signals into two-dimensional images.	95.95%
CyTex	The CyTex methods utilize the inherent periodicity of speech signals for conversion to two-dimensional images.	87.42%
Hilbert Curve	Innovative Hilbert Curve-based method mapping one-dimensional time series to two-dimensional images for emotion recognition.	**98.73%**

performed well on the same dataset and showed significant advantages. The higher accuracy reflects the robustness and effectiveness of our method in capturing complex emotional nuances in speech signals.

First, this study introduced methods similar to those used in this study, including GAF and CyTex, which also transform one-dimensional data into two-dimensional images. The 2D images generated by these methods exhibit different characteristics, as shown in Fig. 11(a), which represents a part of the image of the GAF method. Fig. 11(b) shows the image generated by the CyTex algorithm. Fig. 11(c), represents the image obtained using the Hilbert curve method, and the image is zero-filled, while Fig. 11(d) is transformed using the Hilbert curve but not zero-filled.The advantages and disadvantages of the three algorithms are shown in the Table 6. The GAF algorithm requires a significant amount of memory resources. For a time



(a) GAF algorithm imaging. (b) CytTex algorithm imaging. (c) Hilbert imaging algorithm with 0 padding. (d) Hilbert imaging algorithm with iterative padding.

Fig. 11: The result of Hilbert imaging algorithm.

Table 6
Comparison of GAF, Cytex, and Hilbert.

Method	Size[1]	Temporal Relationship	Inverse Mapping	Accuracy[2]
GAF	***	Preserve	Feasible	2
Cytex	**	Partially Preserve	Partially Feasible	3
Hilbert	*	Preserve	Feasible	1

[1] Size: Use * to indicate the size, the more * the larger the size.
[2] Accuracy: Positive ranking of accuracy expressed as a number, with smaller numbers indicating higher accuracy.

series of length L, the GAF algorithm generates an image of size (L, L), whereas CyTex generates an image of size greater than *(C, L/C)*, where C is the period. Hilbert generates an image of size $(2^n, 2^n)$, where n is the smallest integer that satisfies the condition that 2^{2n} is greater than or equal to L. The following is an example of an integer that satisfies this requirement: The image generated using the GAF algorithm retains the temporal correlation; hence, inverse mapping can be realized. Whereas the image generated by the Cytex algorithm retains a portion of the temporal correlation within each cycle, the temporal correlation between cycles is lost owing to 0-filling; hence, inverse mapping cannot be fully realized either. The image generated by the Hilbert algorithm also retains temporal correlation and performs better in terms of accuracy than the other two algorithms. Thus, it can be concluded that the Hilbert curve is theoretically feasible.

By comparing these methods, some important conclusions were drawn. In this study, accuracy was used as the main evaluation metric and was compared with the results of other studies. The results show that the proposed algorithm performs better in the same environment, particularly when the Hilbert method is used. In addition, from the perspective of computational cost, this study is more efficient than the two methods mentioned above, which provides more advantages for practical application.

Overall, the method proposed in this study is more promising for processing one-dimensional time-series data. By presenting these data as 2D images, we not only increased the richness of the information but also improved the accuracy and performance. The successful application of this method highlights the potential value of deep learning and computer vision in emotion recognition and other fields, thereby providing new directions for future research and application.

5. Conclusions

An innovative approach to transform one-dimensional time-series data into two-dimensional images was proposed and temporal correlation by employing the path of Hilbert curve scheduling. Our method utilizes a

convolutional neural network to extract image features and performs sentiment classification by Hilbert curve spreading. Experiments demonstrate that the proposed method significantly improves data storage efficiency and sentiment recognition accuracy. In terms of space utilization efficiency, our method saves up to 23,195 pixel units of space, which effectively improves the data storage efficiency. In terms of sentiment recognition accuracy, our method significantly outperforms other methods on the same dataset, especially when using the Hilbert unfolding data representation method, which achieves an accuracy of 98.73%. However, the performance of this method relies heavily on large-scale high-quality speech datasets. Acquiring sufficient amounts and diversity of data may be challenging in practical applications. Future research will focus on extending our method to a wider range of application areas, such as brain-based fundamental brain detection based on brainwave signals. Through applications in these fields, we expect to provide new perspectives and opportunities for interdisciplinary research and applications to promote scientific and technological innovations.

6. Acknowledgements

Funding: This work was supported by JST SPRING [grant number Grant Number is JPMJSP2154].

References

[1] C. Hema, F. P. G. Marquez, Emotional speech recognition using cnn and deep learning techniques, Applied Acoustics 211 (2023) 109492.
[2] W. Alsabhan, Human–computer interaction with a real-time speech emotion recognition with ensembling techniques 1d convolution neural network and attention, Sensors 23 (3) (2023) 1386.
[3] R. W. Picard, Affective computing: from laughter to ieee, IEEE transactions on affective computing 1 (1) (2010) 11–17.
[4] C. Darwin, P. Prodger, The expression of the emotions in man and animals, Oxford University Press, USA, 1998.
[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] R. P. Gadhe, D. Babasaheb, R. Deshmukh, D. Babasaheb, Emotion recognition from isolated marathi speech using energy and formants, International Journal of Computer Applications 975 (2015) 8887.
[7] R. W. Picard, Affective computing, MIT press, 2000.
[8] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76.
[9] M. Lech, M. Stolar, C. Best, R. Bolia, Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding, Frontiers in Computer Science 2 (2020) 14.
[10] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, S. L. Schneider, Speech emotion recognition using machine learning – a systematic review, Intelligent Systems with Applications (2023) 200266.
[11] A. Davletcharova, S. Sugathan, B. Abraham, A. P. James, Detection and analysis of emotion from speech signals, Procedia Computer Science 58 (2015) 91–96.
[12] S. Chamishka, I. Madhavi, R. Nawaratne, D. Alahakoon, D. De Silva, N. Chilamkurti, V. Nanayakkara, A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling, Multimedia Tools and Applications 81 (24) (2022) 35173–35194.
[13] A. Hashem, M. Arif, M. Alghamdi, Speech emotion recognition approaches: A systematic review, Speech Communication (2023) 102974.
[14] A. Jain, H. R. Sah, Student's feedback by emotion and speech recognition through deep learning, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, 2021, pp. 442–447.
[15] M. Higuchi, M. Nakamura, S. Shinohara, Y. Omiya, T. Takano, S. Mitsuyoshi, S. Tokuno, et al., Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study, JMIR formative research 4 (7) (2020) e16455.
[16] D. M. Low, K. H. Bentley, S. S. Ghosh, Automated assessment of psychiatric disorders using speech: A systematic review, Laryngoscope investigative otolaryngology 5 (1) (2020) 96–116.
[17] Y. Wang, L. Liang, Z. Zhang, X. Xu, R. Liu, H. Fang, R. Zhang, Y. Wei, Z. Liu, R. Zhu, et al., Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study, Frontiers in Psychiatry 14 (2023) 1195276.
[18] J. F. Sánchez-Rada, C. A. Iglesias, Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison, Information Fusion 52 (2019) 344–356.
[19] S. Abeysinghe, I. Manchanayake, C. Samarajeewa, P. Rathnayaka, M. J. Walpola, R. Nawaratne, T. Bandaragoda, D. Alahakoon, Enhancing decision making capacity in tourism domain using social media analytics, in: 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), IEEE, 2018, pp. 369–375.
[20] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges, Artificial Intelligence Review 55 (7) (2022) 5731–5780.

[21] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Transactions on Affective Computing 10 (1) (2017) 60–75.
[22] C. Zheng, C. Wang, N. Jia, An ensemble model for multi-level speech emotion recognition, Applied Sciences 10 (1) (2019) 205.
[23] T. Anvarjon, Mustaqeem, S. Kwon, Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features, Sensors 20 (18) (2020) 5212.
[24] S. Zhang, R. Liu, X. Tao, X. Zhao, Deep cross-corpus speech emotion recognition: Recent advances and perspectives, Frontiers in neurorobotics 15 (2021) 784514.
[25] M. Swain, B. Maji, P. Kabisatpathy, A. Routray, A dcrnn-based ensemble classifier for speech emotion recognition in odia language, Complex & Intelligent Systems 8 (5) (2022) 4237–4249.
[26] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, N. Amir, The automatic recognition of emotions in speech, Springer, 2011.
[27] C.-N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artificial Intelligence Review 43 (2015) 155–177.
[28] M. Ayadi, M. Kamel, F. Karray, Survey on speech recognition: Resources, features and methods, Pattern Recognition 44 (3) (2011) 572–587.
[29] Y. Yang, C. Fairbairn, J. F. Cohn, Detecting depression severity from vocal prosody, IEEE transactions on affective computing 4 (2) (2012) 142–150.
[30] J. C. Mundt, A. P. Vogel, D. E. Feltner, W. R. Lenderking, Vocal acoustic biomarkers of depression severity and treatment response, Biological psychiatry 72 (7) (2012) 580–587.
[31] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, D. S. Geralts, Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology, Journal of neurolinguistics 20 (1) (2007) 50–64.
[32] Z. Wang, T. Oates, Imaging time-series to improve classification and imputation, arXiv preprint arXiv:1506.00327 (2015).
[33] A. Bakhshi, A. Harimi, S. Chalup, Cytex: Transforming speech to textured images for speech emotion recognition, Speech Communication 139 (2022) 62–75.
[34] A. S. Campanharo, M. I. Sirer, R. D. Malmgren, F. M. Ramos, L. A. N. Amaral, Duality between time series and networks, PloS one 6 (8) (2011) e23378.
[35] D. Hilbert, D. Hilbert, Über die stetige abbildung einer linie auf ein flächenstück, Dritter Band: Analysis · Grundlagen der Mathematik · Physik Verschiedenes: Nebst Einer Lebensgeschichte (1935) 1–2.
[36] ChineseLDC, Chinese academy of sciences emotional speech database, https://www.ChineseLDC.Org, Accessed April 4, 2021.
[37] O. S. Kayhan, J. C. v. Gemert, On translation invariance in cnns: Convolutional layers can exploit absolute spatial location, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14274–14285.
[38] L. Zhu, L. Chen, D. Zhao, J. Zhou, W. Zhang, Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn, Sensors 17 (7) (2017) 1694.
[39] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: