Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 4127401, 12 pages
https://doi.org/10.1155/2017/4127401

Research Article

A New Fuzzy Cognitive Map Learning Algorithm for Speech Emotion Recognition

Wei Zhang, Xueying Zhang, and Ying Sun

College of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China

Correspondence should be addressed to Xueying Zhang; tyzhangxy@163.com

Received 8 February 2017; Revised 1 May 2017; Accepted 28 May 2017; Published 4 July 2017

Academic Editor: Paolo Crippa

Selecting an appropriate recognition method is crucial in speech emotion recognition applications. However, the current methods do not consider the relationship between emotions. Thus, in this study, a speech emotion recognition system based on the fuzzy cognitive map (FCM) approach is constructed. Moreover, a new FCM learning algorithm for speech emotion recognition is proposed. This algorithm includes the use of the pleasure-arousal-dominance emotion scale to calculate the weights between emotions and certain mathematical derivations to determine the network structure. The proposed algorithm can handle a large number of concepts, whereas a typical FCM can handle only relatively simple networks (maps). Different acoustic features, including fundamental speech features and a new spectral feature, are extracted to evaluate the performance of the proposed method. Three experiments are conducted in this paper, namely, single feature experiment, feature combination experiment, and comparison between the proposed algorithm and typical networks. All experiments are performed on TYUT2.0 and EMO-DB databases. Results of the feature combination experiments show that the recognition rates of the combination features are 10%–20% better than those of single features. The proposed FCM learning algorithm generates 5%–20% performance improvement compared with traditional classification networks.

1. Introduction

Speech emotion recognition is a method of recognizing emotions from human speech signals. As an important link in the human-computer interaction system, this method has received increasing attention in recent years. Speech emotion recognition has been widely applied to medical diagnosis [1], drowsy driving detection [2], and emotional state monitoring of students during the e-learning process [3].

Different studies on speech emotion recognition have been conducted, including studies on searching for available acoustic features. Numerous acoustic features, such as prosodic features [4–6], spectral features [7–9], and voice quality [10, 11], are applied to emotion recognition. Some emotions present similarities; thus, using only one type of acoustic feature to recognize emotions is inadequate. Other studies have focused on classification approaches. However, only a limited number of classifiers specifically for speech emotion recognition have been reported. Most classifiers, such as support vector machines (SVMs) [12], k-nearest neighbor (KNN) classifiers [13], and neural networks (NNs) [14], have a wide range of applications, including emotion recognition. However, their recognition rates are low, particularly when two or more similar emotions should be simultaneously identified.

The current study on speech emotion recognition has focused on the different combinations of acoustic features to improve the recognition performance [15]. New speech features should be identified, and innovative classification methods should be developed to improve classification accuracy. The low recognition rate reported in the literature could be attributed to the use of a discrete emotional model in the speech emotion recognition methods. This model classifies emotions into happiness, sadness, anger, surprise, and other primary emotions and regards these emotions as separate and distinct categories. However, this taxonomy is neither aligned with the opinions of cognitive psychology nor approximate to the mood in real life. In recent years, the discrete model

has been partially substituted by a dimensional method for emotions. The dimensional model [16] of emotions supposes that the emotional state can be represented by a continuum of three dimensions, namely, pleasure, arousal, and dominance; thus, this model is also called the PAD model, which is discussed in Section 3. This model has been supported by several studies and statistical analyses [17, 18]. The dimensional model asserts that emotions are not discrete and specific states, but they are fuzzy and interacting with one another.

In this study, a novel learning method based on fuzzy cognitive maps (FCMs) for speech emotion recognition is proposed. Different acoustic features are extracted and fused to overcome the deficiency of one another in classifying certain emotions. Two corpora with five emotions (i.e., sadness, anger, happiness, surprise, and boredom) are selected to test the performance of the network and features. This study is divided into two stages. The first stage is the analysis of speech features to extract various acoustic features for fusion. Fundamental speech features, namely, prosodic features (i.e., speed, energy, zero-crossing rate (ZCR), and pitch) and quality features (i.e., formant), are extracted to recognize emotions in speech. A new spectral feature [19], which is the combination of the methods of the Teager energy operator (TEO) and the Hilbert–Huang transform (HHT), is also extracted. The second stage is the design of the FCM-based recognition system. This stage includes the network architecture design and weight determination. The weights are divided into weight 1 (between emotions and emotions) and weight 2 (between emotions and classes). Weight 2 can be obtained on the basis of the PAD model. After the network input and weight 2 are determined, the structure of the network is derived using certain mathematical derivations. The performance of the proposed combination features and network is evaluated on TYUT2.0 and EMO-DB databases. They both show good performances.

The rest of this paper is organized as follows: Section 2 presents a review of related work. Section 3 describes the three-dimensional PAD model. Section 4 discusses the speech emotion databases and acoustic features. Section 5 provides an overview of FCMs. Section 6 demonstrates the detailed procedure of constructing an FCM-based speech emotion recognition system. Section 7 presents the experiments to determine the recognition rates of the proposed system and combination features. The results are also compared with other conventional classifiers in Section 7. Section 8 concludes the study and presents our future work.

2. Related Work

This section provides a brief review of several important speech features, speech emotion recognition techniques, and FCM learning algorithms.

2.1. Speech Features. Different speech features represent different speech information, such as emotion and speaker, in a highly overlapped manner. Several studies have used various features and their combinations. For instance, Ooi et al. [20] used the combination of pitch, log energy, ZCR, TEO, and mel-frequency cepstral coefficient (MFCC) features to recognize six universal emotions (i.e., happiness, anger, disgust, sadness, surprise, and fear). The recognition rates of the radial basis function NN using the feature combination on the eNTERFACE'05 and RML databases were 75.89% and 65.87%, respectively. MFCC, pitch, and wavelet features were used in the study of Lanjewar et al. [13] to classify seven emotions, namely, happiness, anger, disgust, sadness, surprise, fear, and neutral. For the EMO-DB database, the KNN and Gaussian mixture model networks achieved 51.67% and 66% average classification rates, respectively. Huang et al. [21] used traditional emotional features, such as time, amplitude, and fundamental frequency, to recognize four basic emotions (i.e., sadness, anger, surprise, and happiness); the highest recognition rate on a self-constructed Mandarin database was 86.5%. Xu et al. [22] adopted the statistical information of pitch, ZCR, energy, formant, duration, and MFCC features to classify emotions. On the EMO-DB database (eNTERFACE'05 database), the recognition rates of the KNN, SVM, and naive Bayesian classifiers were 71.25% (54.33%), 72.42% (61.44%), and 72.44% (53.67%), respectively.

2.2. Classification Methods. Different classification methods have been utilized for speech emotion recognition. Most of these methods mainly use probability algorithms or statistical methods for modeling. One of the most popular classification methods for speech emotion recognition is KNN [13, 22]. NNs have also been widely applied in speech emotion recognition [14, 20]. Furthermore, SVM classifiers have been extensively utilized in many studies related to speech emotion recognition [12, 22]. SVMs are based on statistical learning algorithms to model sequential data.

2.3. Fuzzy Cognitive Map Approach. Fuzzy cognitive map approach, which is a soft computing method that provides a powerful and flexible framework for knowledge representation and reasoning, is a convenient tool for dynamic system modeling [23]. Several researchers have recently made progress in the areas of classification [24] and emotional prediction [25] with FCM-based models. These models involve updating the node state values and the causal relationships between concepts to simulate dynamic behavior. A number of weight-learning methods, such as Hebbian learning [26, 27], genetic algorithm (GA) [28], and swarm intelligence optimization algorithm [29], have been applied to learning weights of an FCM. However, most of these methods require domain experts who can specify in advance the initial weight matrix of an FCM. Moreover, GA and swarm intelligence optimization algorithms present low running speeds and cause network instability.

3. PAD Model

Mehrabian and Russell [16] proposed PAD model, in which emotions are presented in a three- dimensional mood space.

Pleasure (P) reflects human responses to environments. High pleasure describes joy or satisfaction, whereas low pleasure indicates boredom and anger. Arousal (A) means

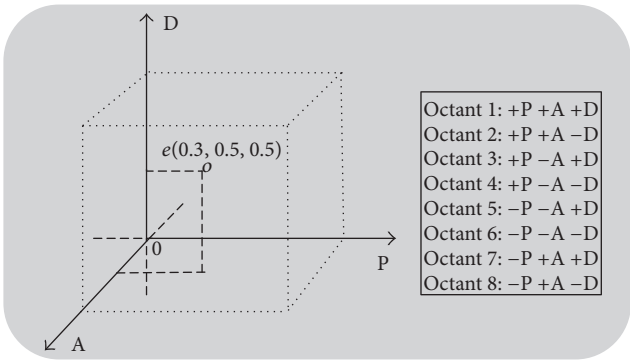

Figure 1: PAD model.

mental and/or physical activation, positive and negative alike. High arousal includes anger and surprise, whereas low arousal includes sadness. Dominance (D) is a feeling of control and influence over relationships, situations, and others, as opposed to submissiveness.

On the basis of these definitions, emotion is described in each of the three mood space axes as follows: +P and −P represent pleasant and unpleasant, respectively; +A and −A represent arousable and unarousable, respectively; and +D and −D represent dominant and submissive, respectively. Accordingly, eight octants exist in the PAD space (Figure 1). For instance, the PAD value for an emotion (0.3, 0.5, 0.5) belongs to octant 1, where all PAD components are positive (see Figure 1).

Research shows that the use of P, A, and D three dimensions can effectively explain the human emotions. For example, Mehrabian [30] used 42 mood scales developed by other investigators to test the PAD model. Findings showed that almost all the reliable variance in the 42 scales was explained in terms of the three PAD scales. The latter results were interpreted as indicating that the PAD model provided a reasonably general characterization and measurement of emotional states. Moreover, these three dimensions are not limited to the subjective experience of describing emotions, which has a good mapping relationship with the emotional physiological arousal and external behavior. Previous studies paid attention to emotional valence and activation degree. However, these two dimensions cannot effectively distinguish certain emotions, such as anger and fear. By contrast, the PAD emotion scale can distinguish between anger and fear. Although both are high-arousal and low-pleasure emotions, these emotions are opposite in terms of dominance; that is, anger is a high-dominance state, whereas fear is a low-dominance state.

The PAD emotion scale is established on the basis of the PAD model. This scale is an elaborate tool developed by Mehrabian to measure emotion, initially consisting of 34 items. Later, the researcher further proposed a simplified version of the PAD emotion scale, where each dimension was described using four items (12 items in total). Li et al. [31] revised the simplified version of the PAD emotion scale to form a Chinese version of PAD scale. The Chinese PAD scale is an abbreviated nine-point semantic differential scale, where each item is composed of a pair of different emotional state adjectives, with space between each pair divided into nine segments. Each pair represents an emotional value that differs in one dimension but remains the same in the other two dimensions. For example, a project that measures pleasure is made up of "excitement" and "anger." The emotions they represent are the opposite of pleasure and are roughly the same in terms of arousal and dominance.

Table 1: PAD values of five basic emotions.

Number	Emotion	Mean		
		P	A	D
1	Happiness	2.77	1.21	1.42
2	Boredom	−0.53	−1.25	−0.84
3	Sadness	−0.89	0.17	−0.7
4	Anger	−1.98	1.10	0.6
5	Surprise	1.72	1.71	0.22

Subjects need to assess the target emotion based on emotional intensity. According to the Chinese version of the PAD scale, from left to right, the score on the item is recorded as "−4" to "4"; in the middle, it is recorded as "0." The final dimension value is the average of the scores of the four items that evaluated the dimension.

Li [32] evaluated the PAD values of 14 types of specific emotions based on the PAD three-dimensional emotion model and the Chinese PAD emotion scale. The PAD values of some emotions are shown in Table 1.

4. Speech Emotion Databases and Acoustic Features

4.1. Speech Emotion Databases. The emotion recognition task is performed on the Taiyuan University of Technology (TYUT2.0) database and Berlin speech emotion database (EMO-DB) [33].

4.1.1. TYUT2.0 Database. The TYUT2.0 database includes 678 utterances extracted from different Chinese radio dramas. Each sentence in the corpus is marked with one of the four emotion classes (happiness, sadness, anger, and surprise). The distributions of the utterances across these four emotion categories are shown in Figure 2.

Unlike most previous databases that utilized speech data from studio-recorded emotional utterances, the TYUT2.0 database was constructed by selecting Chinese radio dramas from which to collect emotional speech data because the emotional utterances intercepted from radio dramas are more realistic than studio-recorded emotional predesigned texts read by talents. The database was established in two steps. First, the speech utterances were intercepted from radio dramas and saved as 11.025 kHz, 16-bit, mono, wave files. The interception criterion is that the emotion conveyed in a sentence is obvious; the text content is not limited. Second, the speech data effectiveness was evaluated. The comprehensive evaluation index established in [34] was used to evaluate 837 sentences. As a result, 678 utterances from more than 200

Table 2: Prosodic features and their statistical parameters.

Prosodic features	Statistical parameters
Energy	Maximum; minimum; average; and rates of maximum, minimum, and average
Pitch frequency	Maximum; minimum; average; and rates of maximum, minimum, and average
ZCR	Average zero-crossing rate
Speed	Average speed

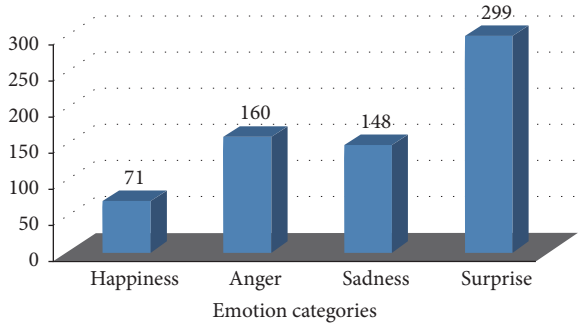

Figure 2: Utterance distributions across the emotion categories in TYUT2.0 database.

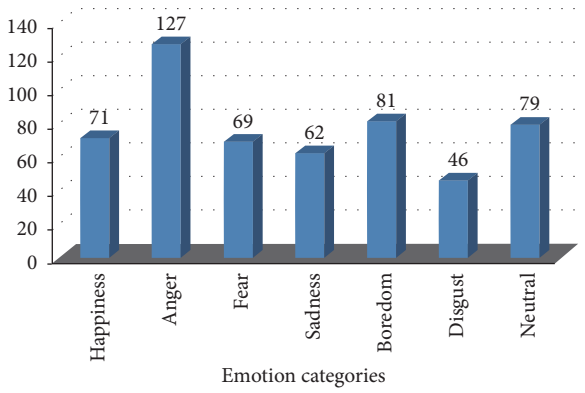

Figure 3: Utterance distributions across the emotion categories in EMO-DB database.

speakers were obtained to achieve the diversity of the corpus and pose certain difficulties to emotion recognition.

4.1.2. EMO-DB Database. EMO-DB is a database of studio-recorded predesigned texts read by talents. This database includes 535 sentences (as shown in Figure 3) of emotional speech in 7 emotional states (i.e., anger, happiness, sadness, fear, disgust, boredom, and neutral). The database involves 10 actors (5 males and 5 females) reading 10 sentences (5 short sentences and 5 long sentences) in German language.

4.2. Acoustic Features. This section introduces the features that we use in this study. The Hilbert marginal Teager energy spectrum coefficient (HTSC) and prosodic features are described in the following subsections.

4.2.1. HTSC Features. This subsection introduces HTSC, which consists of two major parts, namely, HHT and TEO. The detailed extraction procedure for these features is described in [19].

HHT is a method developed by Huang and Liu [35]. It has been proved to be a powerful tool for analyzing nonstationary and nonlinear signals [36]. The HHT has two stages. The first stage is the empirical mode decomposition (EMD), in which signals are decomposed into a series of intrinsic mode functions (IMFs). The second stage is the Hilbert spectral analysis, in which the Hilbert transform is applied to each IMF. The HHT is not limited by time-frequency uncertainty, and, thus, it provides a high-resolution time-frequency analysis.

The speech emotion signal $x(t)$ input is assumed to be given. The HTSC extraction flowchart is shown in Figure 4.

Figure 5 presents the emotional speech Hilbert marginal Teager energy spectra for the two databases. As shown, the four emotions in each speech database present significant differences.

4.2.2. Prosodic Features. In addition to HTSC, which focuses on spectral features, this study also uses the voice toolbox written by Professor Zhang Zhixing of Taiwan to extract prosodic features. Prosodic features, which are also known as suprasegmental features, describe voice changes in pitch, tone, speed, and other aspects. The categories of prosodic features adopted in this study are energy, pitch frequency, ZCR, and speed (speed refers to the pronunciation speed, measured with the number of syllables per second.). The statistical parameters of these features, which are extracted for emotion recognition, are shown in Table 2.

4.2.3. Quality Features. Sound quality is a subjective parameter of speech for assessing whether the speech is pure, clear, and easily identifiable. Acoustic performances that affect sound quality are wheezing, vibrato, and asphyxiation. They often occur when the speaker is emotionally agitated, and they are difficult to suppress. In listening experiments of speech emotion, the change in sound quality is closely related to the expression of speech emotion. The commonly used quality features in speech emotion recognition are formant, bandwidth, and frequency perturbation. In this study, the quality features extracted are formants 1–3 and their statistical parameters, as shown in Table 3.

5. FCMs

The objective of this study is to simulate the human brain perception of the emotional information in speech. Human

Table 3: Quality features and their statistical parameters.

Quality features	Statistical parameters
Formant	
F1	Maximum; minimum; mean; variance; median; and rates of maximum, minimum, and average
F2	Maximum; minimum; mean; variance; median; and rates of maximum, minimum, and average
F3	Maximum; minimum; mean; variance; median; and rates of maximum, minimum, and average

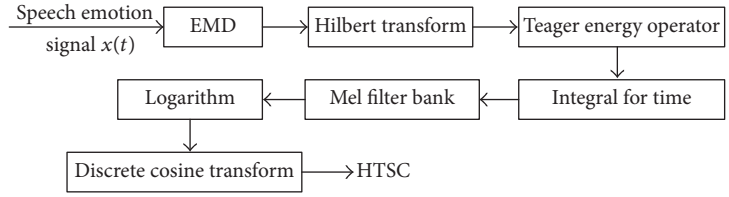

Figure 4: Flowchart of HTSC extraction.

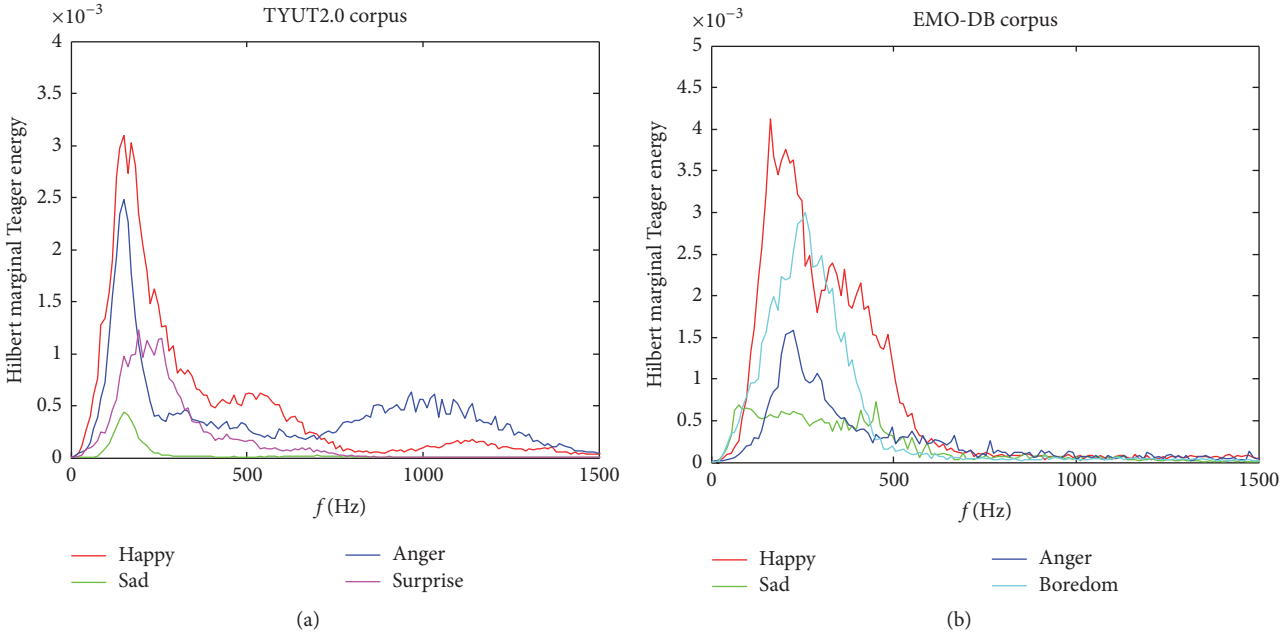

Figure 5: Hilbert marginal Teager energy spectra for the two databases.

beings can understand the emotional state of one another through speech because the human brain can perceive and reason about the emotional information of a speaker via speech signals. Different classification methods have been utilized for the speech emotion recognition. The existing approaches to speech emotion recognition are mainly based on probability algorithms or statistical methods for modeling. In this study, cognitive psychology is used to find a recognition method that is approximate to human cognition and reasoning. Therefore, an FCM model is selected for speech emotion recognition. In an FCM, the connections between concepts are the basic units of knowledge storage; that is, knowledge is stored in the form of concept nodes and the relationship between these concept nodes. Relating connected expressions is a type of logical knowledge, and the use of logical knowledge is a process of reasoning; thus, the process of establishing relationships between concepts can be regarded as a process of reasoning on the basis of the connection. The connections express the relationships and facilitate the implementation of reasoning. Recognition is a prediction task in data mining. Its goal is to predict the class to which an attribute set belongs on the basis of the values of its attributes. The recognition model is established based on the known training samples, which can predict the target class to which the unknown attribute set belongs.

5.1. FCM Fundamentals. In 1986, Kosko [37] proposed FCMs, which are fuzzy-weighted digraphs used for prediction. They are tools for representing and investigating systems and human behavior with the use of nodes and edges. A simple FCM model is shown in Figure 6.

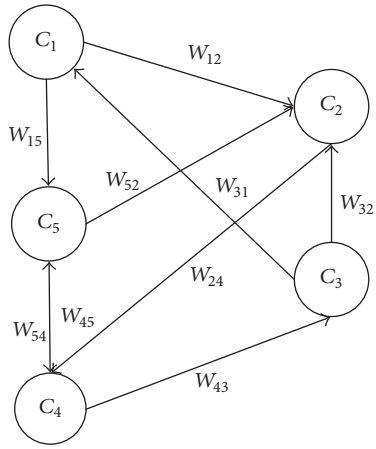

Figure 6: FCM example.

Figure 6 depicts a simple FCM, which includes five nodes and nine edges. Nodes C can be system events, goals, feelings, and trends that reflect system attributes, features, qualities, and conditions. The edges represent the causal relationships between the nodes. The causal relationships between the nodes of an FCM (also called weights W_{ij}) have different intensities represented by fuzzy numbers $[0, 1]$ or $[-1, 1]$.

The causal relationships are categorized into three types:

$$\begin{aligned} &W_{ij} > 0, \quad C_i \text{ to } C_j \text{ as positive causal relationship} \\ &W_{ij} < 0, \quad C_i \text{ to } C_j \text{ as nagative causal relationship} \\ &W_{ij} = 0, \quad C_i \text{ to } C_j \text{ no causal relationship.} \end{aligned} \tag{1}$$

The W_{ij} values are represented in an $n \times n$ (n is the number of nodes) matrix called an adjacency matrix.

The adjacency matrix corresponding to the FCM model in Figure 6 is as follows:

$$\mathbf{W} = \begin{bmatrix} 0 & W_{12} & 0 & 0 & W_{15} \\ 0 & 0 & 0 & W_{24} & 0 \\ W_{31} & W_{32} & 0 & W_{34} & 0 \\ 0 & 0 & 0 & 0 & W_{45} \\ 0 & W_{52} & 0 & W_{54} & 0 \end{bmatrix}. \tag{2}$$

5.2. Dynamic FCM. An FCM can simulate the dynamic behavior of a system by updating the state values of the nodes. The node state of the FCM at time t can be easily expressed by a vector function:

$$\vec{C}(t) = \left(C_1(t), C_2(t), \ldots, C_n(t)\right), \tag{3}$$

where $C_1(t), C_2(t), \ldots, C_n(t)$ are the state values of the n nodes in the FCM network at time t. Before the analysis, the initial state values of the FCM nodes should be given. When $t = 0$, initial state vector is denoted by $\vec{C}(0) = (C_1(0), C_2(0), \ldots, C_n(0))$.

Subsequently, the node state value is updated by

$$C_j(t+1) = f\left(\sum_{i \neq j, i=1}^{n} C_i(t) W_{ij}\right), \tag{4}$$

where $C_i(t)$ is the node value of C_i at time t. The node value of C_j at instant $t + 1$ is $C_j(t + 1)$. Furthermore, W_{ij} indicates the intensity of the relationships between nodes C_i and C_j; $f(x)$ is the activation function, which may be a sigmoid, hyperbolic tangent, or threshold linear function.

When $f(x)$ is selected as the linear function, that is, $f(x) = x$, $x \in$ arbitrary value, then the state values of the n nodes in the FCM at time $t + 1$ are calculated by

$$\begin{aligned} C_j(t+1) &= \left(C_1(t+1), \ldots, C_n(t+1)\right) \\ &= \left(C_1(t), \ldots, C_n(t)\right) \times W \\ &\qquad \left(j = 1, 2, \ldots, n\right). \end{aligned} \tag{5}$$

After finite iterations, the FCM reaches two states: (a) it stabilizes to a fixed pattern of node values, so-called hidden patterns or fixed-point attractors; (b) it remains in circulation among several fixed states, known as the limit cycle. When the FCM reaches the (a)/(b) state, the system reaches the steady/equilibrium state.

6. FCM-Based Speech Emotion Recognition System

6.1. Architecture of FCM. In this study, the use of FCMs is proposed as a technique for speech emotion recognition. The architecture of an FCM-based speech emotion recognition system or network (e-FCM) is composed of two layers (Figure 7), namely, input and output layers. The input layer collects data from the speech features. It includes all the features (linear or nonlinear) that can reflect the emotional state. The output layer, which is directly connected to the input layer, is composed of emotion classes. The PAD model suggests that emotions are continuous and linked with one another. The nodes in the output layer are interrelated, and they are represented by directed arcs (Figure 7). The connections between the input layer and the output layer are one-way arcs, indicating the connections between the speech features and the emotion classes.

The entire emotion recognition system is an FCM. The speech features and emotional classes constitute the input of the FCM model, which completely considers the relationships between classes and between classes and features. As a result, a weight matrix is formed to simulate the dynamic behavior of the FCM model.

The features are represented by F_i ($i = 1, 2, \ldots, n$), and the classes are represented by C_j ($j = 1, 2, \ldots, m$). The weight matrix formed by the relationship between the features and classes is denoted by W_i and also called the input weight matrix. The weight matrix formed by the relationship between classes is denoted by W_o and also called the output

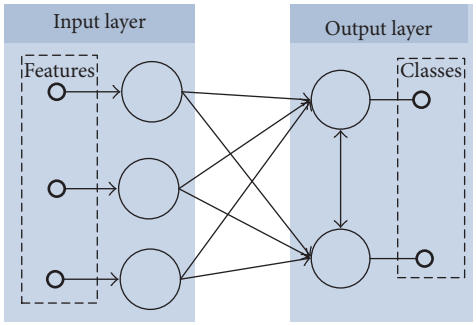

Figure 7: FCM-based speech emotion recognition system architecture.

weight matrix. The weight matrix of the system can be simplified as an $(n+m)\times m$-order matrix:

$$W = \begin{bmatrix} W_i \\ W_o \end{bmatrix}. \tag{6}$$

The output of e-FCM at time $t+1$ according to (5) is

$$\begin{aligned} C_j(t) &= \left(C_{(n+1)}(t+1), \ldots, C_{(n+m)}(t+1)\right) \\ &= \left(F_1(t), \ldots, F_n(t), C_{(n+1)}(t), \ldots, C_{(n+m)}(t)\right) \\ &\quad \times W \quad (j = 1, 2, \ldots, m). \end{aligned} \tag{7}$$

Therefore, the feature values are constant in e-FCM, and only the values of the classes are updated.

6.2. Proposed Learning Algorithm for FCMs. Most learning algorithms [26, 27] for FCMs require domain experts who can predetermine the initial weight matrix. Moreover, GA [28] and swarm intelligence optimization algorithms [29] present low running speeds and induce network instability. These drawbacks result in the failure to match the natural human-machine interactive speech emotional information in real time. This paper proposes a new learning method, which includes the use of the three-dimensional PAD model and certain mathematical derivations, to overcome the aforementioned shortcomings.

As mentioned, in the PAD model, emotions are continuous, and they have certain relationships between them. Therefore, the relationships between the emotional classes (also called output weights W_o) are determined by the emotional PAD value. We construct a three-dimensional emotional space, with P, A, and D, as the axes of the emotional space. The emotional PAD values shown in Table 1 are mapped in this space (Figure 8). The space distance is used to map the relationships between classes and ultimately determine the weights between emotions.

The distance between any two emotional classes is calculated by the Euclidean distance as follows:

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \tag{8}$$

where d_{12} represents the space distance between points 1 and 2 and (x_1, y_1, z_1) and (x_2, y_2, z_2), respectively, represent

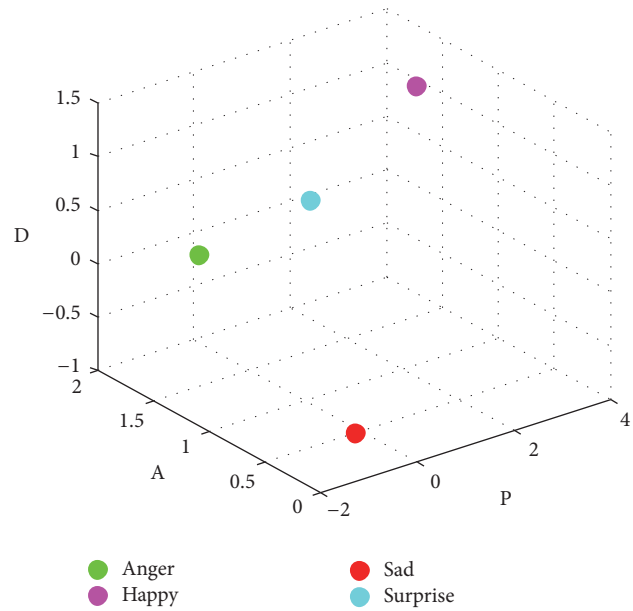

Figure 8: Four basic emotions in the TYUT2.0 database in a three-dimensional space.

the coordinates of points 1 and 2, or the PAD emotions, in the three-dimensional space. The relationships between classes are obtained by calculating the reciprocal of the space distance between any two emotions.

After the output weights W_o (the relationships between the emotional classes) are obtained, an extremely simple and efficient method to train the FCM (i.e., training W_i) is utilized.

The FCM model in Figure 9 represents an example of an FCM-based speech emotion recognition system. Equation (9) shows the adjacency matrix.

$$\mathbf{W} = \begin{bmatrix} 0 \cdots 0 & W_{1(n+1)} & W_{1(n+2)} & W_{1(n+3)} & W_{1(n+4)} \\ \vdots \ \ldots \ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 \cdots 0 & W_{n(n+1)} & W_{n(n+2)} & W_{n(n+3)} & W_{n(n+4)} \\ 0 \cdots 0 & 0 & W_{(n+1)(n+2)} & W_{(n+1)(n+3)} & W_{(n+1)(n+4)} \\ 0 \cdots 0 & W_{(n+2)(n+1)} & 0 & W_{(n+2)(n+3)} & W_{(n+2)(n+4)} \\ 0 \cdots 0 & W_{(n+3)(n+1)} & W_{(n+3)(n+2)} & 0 & W_{(n+3)(n+4)} \\ 0 \cdots 0 & W_{(n+4)(n+1)} & W_{(n+4)(n+2)} & W_{(n+4)(n+3)} & 0 \end{bmatrix}. \tag{9}$$

In the process of emotion recognition, the system is divided into two stages, namely, training and testing. The network is trained to obtain the input weights (W_i) by designing the initial state vector and using the proposed learning algorithm. The initial state vector with n nodes in input layer and four nodes in output layer is as follows:

$$\begin{aligned} &\vec{C}(0) \\ &= \left(\overbrace{F_1, F_2, \ldots, F_n}^{\text{Input}}, \overbrace{C_1(0), C_2(0), C_3(0), C_4(0)}^{\text{output}}\right). \end{aligned} \tag{10}$$

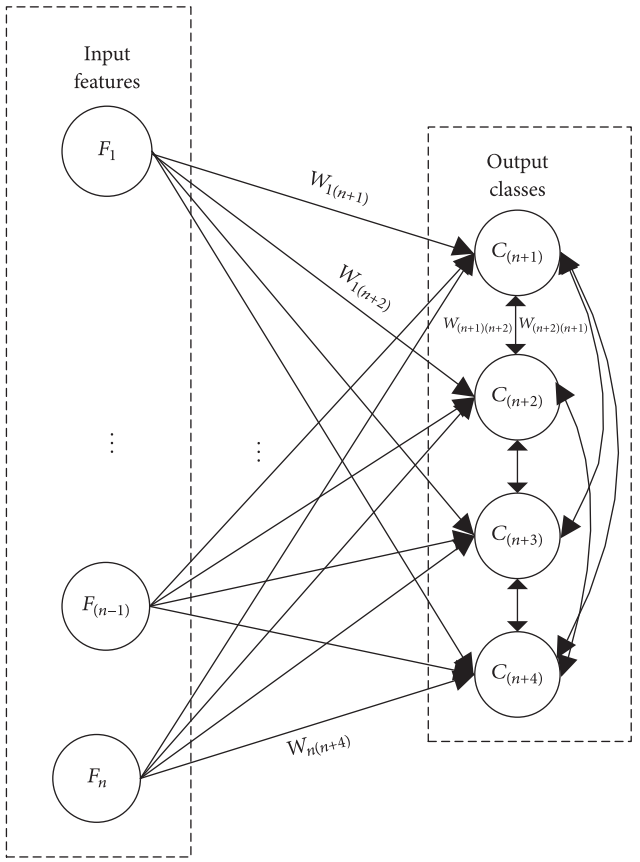

Figure 9: FCM-based speech emotion recognition system/network (e-FCM).

As indicated in (7), the node state value changes as follows during the training process of e-FCM:

$$C(t) = [F, C(t-1)] \begin{bmatrix} W_i \\ W_o \end{bmatrix}, \tag{11}$$

where t is the current state of the network, $t-1$ is the previous state of the network, and F represents the features of the training samples of emotional speech.

Then, when $t = 1$,

$$C(1) = [F, C(0)] \begin{bmatrix} W_i \\ W_o \end{bmatrix} = FW_i + C(0) W_o. \tag{12}$$

For $t = 2$,

$$\begin{aligned} C(2) &= [F, C(1)] \begin{bmatrix} W_i \\ W_o \end{bmatrix} = FW_i + C(1) W_o \\ &= FW_i + \left(FW_i + C(0) W_o\right) W_o \\ &= FW_i + FW_i W_o + C(0) W_o^2. \end{aligned} \tag{13}$$

For $t = 3$,

$$\begin{aligned} C(3) &= [F, C(2)] \begin{bmatrix} W_i \\ W_o \end{bmatrix} = FW_i + C(2) W_0 \\ &= FW_i + \left(FW_i + \left(FW_i + C(0) W_o\right) W_o\right) W_o \\ &= FW_i + FW_i W_o + FW_i W_o^2 + C(0) W_o^3. \end{aligned} \tag{14}$$

For time t,

$$\begin{aligned} C(t) &= FW_i + C(t-1) W_o \\ &= FW_i + FW_i W_o + FW_i W_o^2 + \cdots + FW_i W_o^{(t-1)} \\ &\quad + C(0) W_i^t \\ &= FW_i \left(I + \sum_{i=1}^{t-1} W_o^{t-1} \right) + C(0) W_o^t. \end{aligned} \tag{15}$$

The minimum training error for the network can be achieved by finding $\widehat{C}(t)$ that satisfies the condition

$$\left\| \widehat{C}(t) - T \right\| = \min \left\| C(t) - T \right\|, \tag{16}$$

where T is the objective function that refers to the initial state values of the output nodes. Equation (16) can be rewritten as

$$\left\| \widehat{C}(t) - C(0) \right\| = \min \left\| C(t) - C(0) \right\|. \tag{17}$$

According to (17), training FCM equates to solving the minimum value of cost function as follows:

$$E = \left(FW_i \left(I + \sum_{t=1}^{t-1} W_o^{t-1} \right) + C(0) W_o^t - C(0) \right)^2. \tag{18}$$

For $P = I + \sum_{t=1}^{t-1} W_o^{t-1}$, $E = (FW_i P + C(0)(W_o^t - I))^2$.

Equations (17) and (18) imply that after the feature values and the input weights are determined, training the network is equivalent to finding a least-squares solution $\widehat{W_i}$.

$$FW_i P = C(0) \left(I - W_o^t \right), \tag{19}$$

where I is a unit matrix, W_o is a square matrix, and F is an emotional speech feature sequence, which is often a column vector and is thus irreversible. According to Moore–Penrose generalized inverse matrix theory [38, 39], solving the MP generalized inverse matrix of F is equivalent to finding the minimum-norm least-squares solution of (19).

$$\widehat{W_i} = F^+ P^{-1} C(0) \left(I - W_o^t \right), \tag{20}$$

where F^+ is the MP generalized inverse matrix of F.

The following are the steps of e-FCM.

Given a training set of emotional speech features F, we have the following steps.

Step 1. The output weights (W_o) between emotional classes are calculated using the PAD emotion scale.

Step 2. The input weights (W_i) are calculated by $W_i = F^+ P^{-1} C(0)(I - W_o^t)$.

Step 3. The testing set is used as the input to test network performance.

Table 4: Emotion categories and utterance distributions selected in the experiment.

Database	Anger	Happiness	Sadness	Surprise	Boredom	All
TYUT2.0	62	59	62	62	—	245
EMO-DB	81	71	62	—	80	294

Table 5: Recognition results on TYUT2.0 (%).

Features	Emotion				Average accuracy
	Anger	Happiness	Sadness	Surprise	
HMTC	63.64	25.00	33.33	83.33	51.06
Prosodic	50.00	66.67	50.00	33.33	50.00
Formant	66.67	41.67	25.00	58.33	47.92
Prosodic + formant	83.33	75.00	58.33	66.67	70.83
HMTC + prosodic + formant	66.67	75.00	66.60	83.33	72.92

7. Experiment and Analysis

TYUT2.0 and EMO-DB databases are selected for the validation experiments. A total of 245 samples from the original database of TYUT2.0 and 294 samples from the EMO-DB database (as shown in Table 4) are included in our experiments. The samples are randomly selected from the database, and 75% of them are used for the training set, and the remaining 25% are used for the testing set. The TYUT2.0 database includes four categories of emotions. For comparison, four emotions (anger, happiness, sadness, and boredom) are selected from the EMO-DB database. The features described in Section 4.2 are used in our experiments. The experiments with single feature and multifeature combination are described in Section 7.1. The performance evaluation of the proposed learning algorithm for e-FCM is presented in Section 7.2. Comparisons with existing systems on these databases are conducted in our experiments.

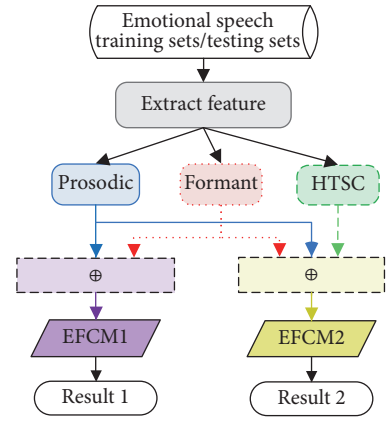

Figure 10: Diagram of the multifeature combination based on emotional speech signals.

7.1. Multifeature Combination Experiment. The statistical parameters of the three acoustic features described in Section 4.2 are evaluated on the EMO-DB and TYUT2.0 databases. Comparisons with multifeature combination on the two databases are also conducted, as shown in Figure 10.

In the first experiment, the TYUT2.0 database is selected to evaluate the features. The experimental results are given in Table 5.

Table 5 shows that prosodic and formant features achieve recognition rates of 50% and 47.92%, respectively. The HMTC feature exhibits a recognition rate of 51.06%, which is 1.06% and 3.14% higher than those of prosodic and formant features, respectively. The prosodic and formant feature combination achieves a recognition rate of 70.83%, and the three-feature combination presents a recognition rate of 72.92%.

In the second experiment, EMO-DB database is selected with four emotions (anger, happiness, sadness, and boredom) to evaluate the features. The experimental results are shown in Table 6. The prosodic and formant features achieve the same recognition rate of 62.89%. The proposed spectral feature (i.e., HMTC) has a recognition rate of 72.16%, which is 10% higher than the rates obtained by prosodic and formant features. The prosodic and formant feature combination yields a recognition rate of 74.23%, and the three-feature combination obtains a recognition rate of 85.37%, which is 11.14% higher than the rates obtained by the two-feature combination.

7.2. e-FCM Experimental Results and Comparison. The performance of the proposed e-FCM is evaluated on the TYUT2.0 and EMO-DB databases. Comparisons with existing networks, such as back propagation (BP) and SVM, on the two databases are conducted. The average recognition rates are shown in Table 7.

As shown in Table 7, the traditional networks BP and SVM achieve recognition rates of 43.75% and 64.58% on the TYUT2.0 databases, respectively. The recognition rate of the proposed e-FCM is 72.92%, which is 29.17% and 8.34% higher than the rates obtained by traditional networks.

The recognition rates on the EMO-DB databases in Table 7 indicate that e-FCM can achieve a better performance than the classical recognition networks based on probability

Table 6: Recognition results on EMO-DB (%).

Feature	Emotion				Average accuracy
	Anger	Happiness	Sadness	Boredom	
HMTC	88.46	41.67	85.00	74.07	72.16
Prosodic	80.77	37.50	65.00	66.67	62.89
Formant	65.38	50.00	85.00	55.56	62.89
Prosodic + formant	73.08	79.17	80.00	66.67	74.23
HMTC + prosodic + formant	84.62	54.17	90.00	98.11	85.37

Table 7: Average recognition results on TYUT2.0 and EMO-DB (%).

Network	TYUT2.0	EMO-DB
BP	43.75	65.04
SVM	64.58	81.22
e-FCM	72.92	85.37

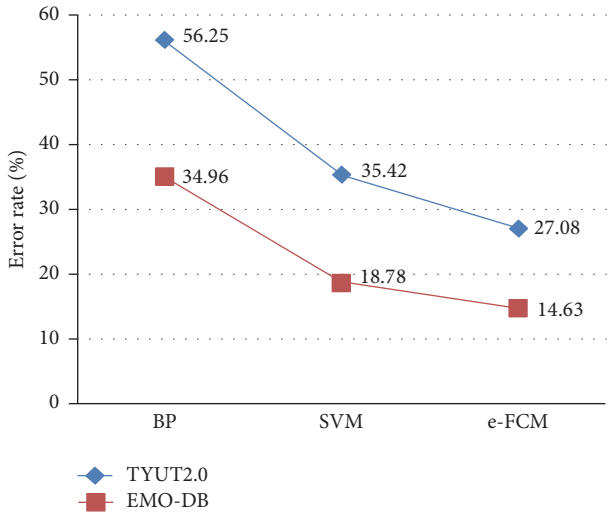

Figure 11: Error rate on TYUT2.0 and EMO-DB.

and statistical methods under most conditions. The recognition rate of e-FCM is 85.37%, which is 20.33% higher than that of BP, 7.32% higher than that of KNN, and 4.15% higher than that of SVM.

Regarding the comparison of the recognition rates of the two databases (Table 7), it is not difficult to find the following: the recognition rate on the EMO-DB database is generally higher than that on the TYUT2.0 database. This deviation is ascribed to the different methods of constructing the two databases. EMO-DB is a database of studio-recorded predesigned texts read by actors, whereas the TYUT2.0 database includes extracted utterances from radio dramas. In addition, fixed speakers were used in the former but not in the latter. As a result, the utterances from the TYUT2.0 database are more realistic than those from the EMO-DB database.

The recognition error rates on the two databases are shown in Figure 11. With the TYUT2.0 database, SVM and e-FCM gave an error reduction from 37.03% to 51.86%, relative to BP. With the EMO-DB database, SVM and e-FCM gave an error reduction from 46.28% to 58.15%, relative to BP. This shows that the e-FCM has more advantages on the TYUT2.0 database than on the EMO-DB database. This result indicates that e-FCM can better handle natural emotional speech.

With the TYUT2.0 database, the lowest error rate is e-FCM (27.08%). This conclusion is true for EMO-DB database. It shows that the e-FCM is more robust than BP and SVM, and it could provide better matches between the testing and training conditions. Therefore, e-FCM is preferable for high-performance speech emotion recognition systems.

8. Conclusions

A novel speech emotion recognition system or network (e-FCM) is proposed in this paper. The architecture of e-FCM has two layers, namely, input layer and output layer. The input layer collects data from the speech features. The spectral feature HTSC is compared and combined with basic speech features, such as energy, pitch, ZCR, speed, and formant. The output layer is composed of emotion classes. The new learning algorithm for e-FCM is established for speech emotion recognition. The e-FCM system consists of two weights, namely, the input weight and the output weight. The PAD emotion scale is used to calculate the output weights, and certain mathematical derivations are performed to determine the input weights. The e-FCM system is tested on the databases of TYUT2.0 and EMO-DB to validate its effectiveness. The results of e-FCM are compared with those of traditional networks. The experimental and comparison results indicate that e-FCM exhibits a good performance. The new spectral feature also obtains better results than other basic speech features do. The combination features exhibit good results. For future studies, other emotional speech features can be considered in establishing a speech emotion recognition network to increase the recognition rate in identifying emotional speech in real life. This paper focuses on five emotions, and we intend to identify more emotion classes, such as hate and fear, in our future work.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This project was sponsored by the National Natural Science Foundation of China (Grant no. 61371193) and the

Youths Foundation of Shanxi Province, China (Grant no. 201601D202045).

References

[1] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: prosody and everything else," *The Workshop on Child*, 2012.

[2] H. Bořil, S. O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association 2010 (Interspeech 2010)*, vol. 1-2, pp. 502–505, Japan, 2010.

[3] W. Wang and J. Wu, "Emotion recognition based on CSO&SVM in e-learning," in *Proceedings of International Conference on Natural Computation (Icnc 2011)*, pp. 566–570, Shanghai, China, 2011.

[4] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Proceedings of the INTERSPEECH 2005 - Eurospeech, European Conference on Speech Communication and Technology*, pp. 493–496, Lisbon, Portugal, September 2005.

[5] A. Origlia, V. Galatà, and B. Ludusan, "Automatic classification of emotions via global and local prosodic features on a multilingual emotional database," *Atmospheric Environment*, vol. 36, no. 30, pp. 4823–4837, 2010.

[6] T. Iliou and C.-N. Anagnostopoulos, "Statistical evaluation of speech features for emotion recognition," in *Proceedings of the 2009 4th International Conference on Digital Telecommunications, ICDT 2009*, pp. 121–126, fra, July 2009.

[7] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[8] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[9] R. Chauhan, J. Yadav, S. G. Koolagudi, and K. S. Rao, "Text Independent Emotion Recognition Using Spectral Features," *Communications in Computer and Information Science*, vol. 168, pp. 359–370, 2011.

[10] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '07*, pp. IV17–IV20, usa, April 2007.

[11] M. Lugger, M.-E. Janoir, and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition," in *Proceedings of the Signal Processing Conference, 2009 European*, pp. 1225–1229, gbr, August 2009.

[12] Z. Wanli, L. Guoxin, and W. Lirong, "Speech Emotion Recognition Using Fourier Parameters," in *Proceedings of the Ieee Transactions on Affective Computing*, vol. 6, pp. 69–75, Dalian, China, 2015.

[13] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques," in *Proceedings of International Conference on Advances in Computing, Communication and Control, ICAC3 2015*, pp. 50–57, ind, April 2015.

[14] D. Philippou-Hübner, B. Vlasenko, R. Böck, and A. Wendemuth, "The performance of the speaking rate parameter in emotion recognition from speech," in *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2012*, pp. 296–301, aus, July 2012.

[15] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[16] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*, MIT Press, 1974.

[17] B. Kühnlenz, S. Sosnowski, M. Buß, D. Wollherr, K. Kühnlenz, and M. Buss, "Increasing Helpfulness towards a Robot by Emotional Adaption to the User," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 457–476, 2013.

[18] N. Herbeth and D. Blumenthal, "Product appraisal dimensions impact emotional responses and visual acceptability of instrument panels," *Food Quality and Preference*, vol. 29, no. 1, pp. 53–64, 2013.

[19] W. Zhang, X. Zhang, and Y. Sun, "Speech emotion recognition based on the HHT marginal Teager energy spectrum," in *Proceedings of the National Conference on Man-Machine Speech Communication 2013 (NCMMSC 2013)*, pp. 91–94, Taiyuan, China, 2013.

[20] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[21] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and SVM," *Mathematical Problems in Engineering*, vol. 2014, Article ID 749604, 2014.

[22] X. Xu, C. Huang, C. Wu, Q. Wang, and L. Zhao, "Graph learning based speaker independent speech emotion recognition," *Advances in Electrical and Computer Engineering*, vol. 14, no. 2, pp. 17–22, 2014.

[23] A. Press, in *International journal of man-machine studies*, pp. 940-941, Academic Press, 1969.

[24] G. A. Papakostas, D. E. Koulouriotis, A. S. Polydoros, and V. D. Tourassis, "Towards Hebbian learning of Fuzzy Cognitive Maps in pattern classification problems," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10620–10629, 2012.

[25] J. L. Salmeron, "Fuzzy cognitive maps for artificial emotions forecasting," *Applied Soft Computing Journal*, vol. 12, no. 12, pp. 3704–3710, 2012.

[26] J. L. Salmeron and E. I. Papageorgiou, "Fuzzy grey cognitive maps and nonlinear Hebbian learning in process control," *Applied Intelligence*, vol. 41, no. 1, pp. 223–234, 2014.

[27] A. P. Anninou and P. P. Groumpos, "Modeling of Parkinson's disease using fuzzy cognitive maps and non-linear hebbian learning," *International Journal on Artificial Intelligence Tools*, vol. 23, no. 5, Article ID 14500109, 2014.

[28] W. Stach, L. Kurgan, and W. Pedrycz, "A divide and conquer method for learning large fuzzy cognitive maps," *Fuzzy Sets and Systems. An International Journal in Information Science and Engineering*, vol. 161, no. 19, pp. 2515–2532, 2010.

[29] E. Yesil, C. Ozturk, M. Furkan Dodurka, and A. Sakalli, "Fuzzy cognitive maps learning using Artificial Bee Colony optimization," in *Proceedings of the 2013 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2013*, ind, July 2013.

[30] A. Mehrabian, "Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[31] X. Li, H. Zhou, S. Song, T. Ran, and X. Fu, "The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales," in *Affective Computing and Intelligent Interaction*, vol. 3784 of *Lecture Notes in Computer Science*, pp. 513–518, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[32] X. M. Li, "PAD three-dimensional emotion model," *Computer world*, vol. 29, no. B14, 2007.

[33] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the INTERSPEECH 2005 - Eurospeech, European Conference on Speech Communication and Technology*, pp. 1517–1520, Lisbon, Portugal, 2005.

[34] J. Song, X. Zhang, Y. Sun, and C. Jiang, "Establishment of emotional speech database based on fuzzy comprehensive evaluation method," *Modern Electronics Technique*, vol. 39, no. 13, pp. 51–54, 2016.

[35] N. E. Huang and H. H. Liu, "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis," in *Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences*, vol. 454, 1971, pp. 903–995, 1998.

[36] D. G. Aggelis, A. C. Mpalaskas, D. Ntalakas, and T. E. Matikas, "Effect of wave distortion on acoustic emission characterization of cementitious materials," *Construction and Building Materials*, vol. 35, pp. 183–190, 2012.

[37] B. Kosko, "Fuzzy cognitive maps," *International Journal of Man-Machine Studies*, vol. 24, no. 1, pp. 65–75, 1986.

[38] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley & Sons, New York, NY, USA, 1971.

[39] D. Serre, "Matrix Factorizations and Their Applications," in *Matrices*, vol. 216 of *Graduate Texts in Mathematics*, pp. 207–223, Springer New York, New York, NY, 2010.